Hamed Rezazadegan Tavakoli

VISUAL SALIENCY AND EYE MOVEMENT: MODELING AND APPLICATIONS
HAMED REZAZADEGAN TAVAKOLI

VISUAL SALIENCY AND EYE MOVEMENT: MODELING AND APPLICATIONS

Academic dissertation to be presented with the assent of the Doctoral Training Committee of Technology and Natural Sciences of the University of Oulu for public defence in Auditorium TS101, Linnanmaa, on 14 November 2014, at 12 noon

UNIVERSITY OF OULU, OULU 2014
Rezazadegan Tavakoli, Hamed, Visual saliency and eye movement: modeling and applications.
University of Oulu Graduate School; University of Oulu, Faculty of Information Technology and Electrical Engineering, Department of Computer Science and Engineering; Infotech Oulu
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract
Humans are capable of narrowing their focus on the highlights of visual information in a fraction of time in order to handle enormous mass of data. Akin to human, computers should deal with a tremendous amount of visual information. To replicate such a focusing mechanism, computer vision relies on techniques that filter out redundant information. Consequently, saliency has recently been a popular subject of discussion in the computer vision community, though it is an old subject matter in the disciplines of cognitive sciences rather than computer science. The reputation of saliency techniques – particularly in the computer vision domain – is greatly due to their inexpensive and fast computation which facilitates their use in many computer vision applications, e.g., image/video compression, object recognition, tracking, etc.

This study investigates visual saliency modeling, which is the transformation of an image into a salience map such that the identified conspicuousness agrees with the statistics of human eye movements. It explores the extent of image and video processing to develop saliency techniques suitable for computer vision, e.g., it adopts sparse sampling scheme and kernel density estimation to introduce a saliency measure for images.

Also, it studies the role of eye movement in salience modeling. To this end, it introduces a particle filter based framework of saccade generation incorporated into a salience model. Moreover, eye movements and salience are exploited in several applications.

The contributions of this study lie on the proposal of a number of salience models for image and video stimuli, a framework to incorporate a model of eye movement generation in salience modeling, and the investigation of the application of salience models and eye movements in tracking, background subtraction, scene recognition, and valence recognition.

Keywords: computer vision, pattern recognition, saliency map, vision system, visual attention
Rezazadegan Tavakoli, Hamed, Visuaalinen salienssi ja silmänliike: mallintaminen ja sovellukset.
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Tieto- ja sähköteknikan tiedekunta, Tietoteknian osasto; Infotech Oulu
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Ihmiset kykenevät kohdistamaan katseensa hetkessä näkymän keskeisiin asioihin, mikä vaatii näköjärjestelmältä valtavan suuren tietomäärien käsittelyä. Kuten ihmisen myös tietokoneen pitäisi pystyä käsittelemään vastaavasti suurta määrää visuaalista informaatiota. Tällaisen mekanismin toteuttaminen tietokonenäöllä edellyttää menetelmiä, joilla redundanttista tietoa voidaan suodattaa. Tämän vuoksi salienssista eli silmiinpistävyydestä on muodostunut viime aikoina suosittu tutkimusaihe tietotekniikassa ja erityisesti tietokonenäöön tutkimusyhteisössä, vaikka sitä sinänsä on jo pitkään tutkittu kognitiivisissa tieteissä. Salienssimenetelmiä tunnetaan erityisesti tietokonenäössä johtuen pääasiassa niiden laskennallisesta tehokkuudesta, mikä taas mahdollistaa menetelmien käytön monissa tietokonenäöön sovelluksissa kuten kuvan ja videon pakkaamisessa, objektin tunnistuksessa, seurannassa, etc.

Tässä väitöskirjassa tutkitaan visuaalisen salienssin mallintamista, millä tarkoitetaan muun muassa kuvasta salienssikartaksi siten, että laskennallinen silmiinpistävyys vastaa ihmisen silmänliikkeistä muodostettavaa statistiikkaa. Työssä tarkistetaan tällä työmenetelmällä, miten kuvan- ja videonkäsittely voidaan käyttää kehitettämään salienssimenetelmiä tietokonenäön tarpeisiin. Työssä esitellään esimerkiksi harvaa näytteistystä ja ydinestimointia hyödyntävä kuvien salienssimittäminen.

Työssä tutkitaan myös silmänliikkeiden merkitystä salienssin mallintamisen kannalta. Tästä varten esitellään partikkelisuodatusta hyödyntävä lähestymistapa sakakidien geneerointiin, joka voidaan liittää salienssimallit. Lisäksi silmänliikkeitä ja salienssia hyödynnetään useissa sovelluksissa.

Suoritetun tutkimuksen tieteellisiin kontribuoituihin sisältyvät useat esitetyt salienssimallit kuvasta ja videoista saatavalle herätteelle, lähestymistapa silmänliikkeiden laskennalliseen mallintamiseen ja geneerointiin osana salienssimallia sekä salienssimallien ja silmänliikkeiden sovellettavuuden tutkiminen visuaalisessa seurannassa, taustanvähennyksessä, näkymänanalyysissä ja valenssin tunnistuksessa.

Asiasonat: hahmontunnistus, näköjärjestelmä, salienssikartta, tietokonenäkö, visuaalinen tarkkaavaisuus
To my parents
Preface

This is a compilation thesis (compendium), i.e., it consists of several academic papers with a clarification about their interrelation. The dissertation is conducted under the supervision of Prof. Janne Heikkilä and Dr. Esa Rahtu on the theme of visual saliency.

The first exposure of me to the topic of saliency was in 2009, where Assoc. Prof. Hamid-Reza Pourreza introduced me the VOCUS framework of Frintrop (2006). Later in collaboration with Assoc. Prof. Mohammad-Shahram Moin, I worked on tracking that emerged into sort of saliency-based target representation (Tavakoli & Moin 2010). In August, 2010, I joined Center for Machine Vision research (CMV) to conduct research on visual saliency. The result of the research is the present compendium which consists of eight papers.

To keep consistent with the tradition, the story behind each paper is summarized. The very early paper in this collection is influenced by the former research of Dr. Esa Rahtu (Rahtu et al. 2010), my idea on adopting center-surround comparisons rooted in (Tavakoli & Moin 2010), and Prof. Heikkilä’s emphasize on efficiency. Consequently, we formulated a Bayesian center-surround model published in Paper I. I authored the paper, implemented the method, and conducted the experiments under their supervision.

The idea behind Paper II came to me by reading (Martinez-Conde et al. 2004). Intrigued to investigate the role of eye movements, I formulated, implemented, and drafted the first version of the paper during July 2011 (summer holidays) in the solace of university silence. The first draft utilized a grid-based implementation. However, valuable comments of reviewers provided the motivation for particle filter formulation which was carried out under the supervision of Prof. Heikkilä.

Paper III is the fruitful result of collaboration with Dr. Ali Borji from the University of Southern California. A discussion on the evaluation metrics, meanwhile, I was inquiring content analysis using eye movements and salience, led us into adopting the idea of task decoding for the purpose of model evaluation and investigate the effect of emotional stimulus on the performance of models. In this publication, I developed the details of the eye movement based features, implemented it, and conducted the related experiments in regard with task decoding. Also, I developed the idea of a scanpath-based evaluation metric and implemented it.
Paper IV is the outcome of collaboration with Dr. Victoria Yanulevskaya. In essence, it is the application of eye movement based features in emotional content analysis. I authored most of the paper, and did the implementation of the proposed framework, and conducted all the experiments. Dr. Yanulevskaya provided her baseline framework (Yanulevskaya et al. 2012) and contributed in writing of the related work section. The features were refined in discussion with Prof. Heikkilä. Dr. Rahtu helped polishing the manuscript and Prof. Sebe provided valuable comments.

I authored, implemented, and carried out all the experiments of Paper V, which was approved by Prof. Heikkilä and Dr. Rahtu for submission. I should note Dr. Rahtu’s comment on the possibility of exploring the video domain using technique of Paper I as an incentive.

Prof. Heikkilä initially shared his idea and implementation for exploring saliency for video with me. I adopted the idea and modified it as presented in Paper VI. Eventually, I developed and implemented a new model and conducted the experiments. Dr. Rahtu helped me improving the content and enhancing the mathematical representation of the paper.

Paper VII is in practice an extension of Paper VI targeted specifically for the background subtraction purpose. I drafted the paper, implemented the method, and conducted the experiments. Prof. Heikkilä and Dr. Rahtu approved the paper for submission.

Paper VIII is the consequence of collaboration with Assoc. Prof. Moin and a continuation of (Tavakoli & Moin 2010). I authored the paper, implemented the method and carried out the experiments. Helpful comments of Prof. Heikkilä and Assoc. Prof. Moin assisted me improving the quality and content of the paper.

Oulu, September 2014
Hamed Rezazadegan Tavakoli
Acknowledgements

First of all, I would like to thank all the people who contributed to the formation of this thesis. My profound gratitude goes to Prof. Janne Heikkilä and Dr. Esa Rahtu, my mentors and supervisors, and Dr. Ali Borji, Dr. Victoria Yanulevskaya, Assoc. Prof. Mohammad-Shahram Moin and Prof. Nicu Sebe. Without their guidance and support, I would certainly not have been able to complete this work. I put gratitude towards Prof. Heikkilä’s mathematical skills. I thank him for the opportunity to conduct the research, valuable advice, and support. I thank Dr. Rahtu for his helpful comments, guidance and support. I salute Dr. Borji for his scientific approach, ideas, and talks which, despite the great distance, resulted in probably the most influential contribution of this thesis. I would like to thank Dr. Yanulevskaya and Prof. Sebe for their friendship as well as the opportunity to visit the University of Trento, and participate in their productive research. I thank Assoc. Prof. Moin for his support in continuing a career abroad, and adore his scientific reasoning and comments.

I would like to thank my pre examiners, Assoc. Prof. Olievier Le Meur and Prof. John K. Tsotsot for their valuable comments which enriched the current thesis. I would like to thank Assoc. Prof. Le Meur for his comments that provided some ideas for future work. I put gratitude towards Prof. Tsotsos’s vast knowledge on the field and thank him for his valuable comments and advice which lighted a torch for being a successful researcher.

I also wish to thank all my colleagues and friends for the unforgettable recollections which they gifted me during my PhD. In particular, I thank Miss Yaghoobi for all the enjoyable discussions and memorable moments that helped me to keep my spirits up.

My very special thanks go to my family for their unflagging support. I thank my parents for motivating me to be persistent in my path. Definitely, without their love, it was impossible to complete this thesis work.

This work was supported by the InfoTech Oulu doctoral program. Also, Nokia foundation support is gratefully acknowledged.

Oulu, September 2014
Hamed Rezazadegan Tavakoli
## Abbreviations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>Image pixel</td>
</tr>
<tr>
<td>f</td>
<td>Fixation</td>
</tr>
<tr>
<td>s</td>
<td>Saccade</td>
</tr>
<tr>
<td>I((x))</td>
<td>Image Patch Centered at (x)</td>
</tr>
<tr>
<td>Sal(_x)</td>
<td>Saliency at (x)</td>
</tr>
<tr>
<td>C</td>
<td>Central Region of an Image Patch</td>
</tr>
<tr>
<td>S</td>
<td>Surrounding Region of an Image Patch</td>
</tr>
<tr>
<td>(\mathcal{N}(\cdot))</td>
<td>Normalizing Function</td>
</tr>
<tr>
<td>(\mathcal{G}(\cdot))</td>
<td>Gaussian Kernel</td>
</tr>
<tr>
<td>(\mathcal{K}(\cdot))</td>
<td>Isotropic Kernel</td>
</tr>
<tr>
<td>(\delta(\cdot))</td>
<td>Delta Function</td>
</tr>
<tr>
<td>D((\cdot,\cdot))</td>
<td>Distance Function</td>
</tr>
<tr>
<td>(|\cdot|)</td>
<td>(l_2)-norm</td>
</tr>
<tr>
<td>(|\cdot|_F)</td>
<td>Frobenius Norm</td>
</tr>
<tr>
<td>(\text{LSN}_r^n)</td>
<td>Local Similarity Number Operator with Radius (r) and (n) Neighbors</td>
</tr>
<tr>
<td>disk(_i)</td>
<td>Disk Structure of Element Size (i)</td>
</tr>
<tr>
<td>(e_k)</td>
<td>Discrete eye state at time step (k)</td>
</tr>
<tr>
<td>*</td>
<td>Convolution Operator</td>
</tr>
<tr>
<td>(\oplus)</td>
<td>Dilation Operator</td>
</tr>
<tr>
<td>(W^\dagger)</td>
<td>Pseudoinverse of a Matrix (W)</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>Attenuation Parameter</td>
</tr>
<tr>
<td>(\eta)</td>
<td>Normalizing Constant</td>
</tr>
<tr>
<td>(\tau)</td>
<td>Threshold Value</td>
</tr>
<tr>
<td>(\text{AUC})</td>
<td>Area Under the ROC Curve</td>
</tr>
<tr>
<td>(\text{BoW})</td>
<td>Bag-of-Visual-Words</td>
</tr>
<tr>
<td>(\text{CLE})</td>
<td>Constrained Levy Exploration</td>
</tr>
<tr>
<td>(\text{DBN})</td>
<td>Dynamic Bayesian Networks</td>
</tr>
<tr>
<td>(\text{DCT})</td>
<td>Discrete Cosine Transform</td>
</tr>
<tr>
<td>(\text{EMD})</td>
<td>Earth Mover’s Distance</td>
</tr>
<tr>
<td>(\text{FFT})</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>(\text{FIT})</td>
<td>Feature-Integration Theory</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>HCI</td>
<td>Human-Computer Interaction</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>HVS</td>
<td>Human Visual System</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Component Analysis</td>
</tr>
<tr>
<td>ICL</td>
<td>Incremental Coding Length</td>
</tr>
<tr>
<td>IOR</td>
<td>Inhibition of Return</td>
</tr>
<tr>
<td>KL</td>
<td>Kullback-Leibler</td>
</tr>
<tr>
<td>LHMC</td>
<td>Levy Hybrid Monte Carlo</td>
</tr>
<tr>
<td>LSN</td>
<td>Local Similarity Number</td>
</tr>
<tr>
<td>LSP</td>
<td>Local Similarity Patterns</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov Chain Monte Carlo</td>
</tr>
<tr>
<td>MEP</td>
<td>Mean Eye Position</td>
</tr>
<tr>
<td>NMF</td>
<td>Non-negative Matrix Factorization</td>
</tr>
<tr>
<td>NSS</td>
<td>Normalized Scanpath Saliency</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial Basis Function</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>ROI</td>
<td>Region of Interest</td>
</tr>
<tr>
<td>SA</td>
<td>Simulated Annealing (SA)</td>
</tr>
<tr>
<td>SBSG</td>
<td>Stochastic Bottom-up Saccade Generation</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>SW</td>
<td>Spectral Whitening</td>
</tr>
<tr>
<td>WTA</td>
<td>Winner-Take-All</td>
</tr>
</tbody>
</table>
List of original publications

This thesis is based on the following articles, which are referred to in the text by their Roman numerals (I–VIII):


The present author contributed to various perspectives of each paper such as idea, implementation, conducting experiments, and writing. Depending on the significance of his role, his presence varies from a co-author to a corresponding/first author.
Contents

Abstract
Tiivistelmä
Preface
Acknowledgements
Abbreviations
List of original publications
Contents
1 Introduction
  1.1 Visual saliency and visual attention
  1.2 Objectives
  1.3 Contributions
  1.4 Outline of the thesis
2 Image saliency
  2.1 The history
  2.2 Saliency modeling
    2.2.1 Center-surround models
    2.2.2 Information theoretic models
    2.2.3 Frequency domain models
    2.2.4 Top-down manipulated models
    2.2.5 Connection based models
    2.2.6 Learning based models
    2.2.7 Miscellaneous models
  2.3 The center-surround hypothesis for saliency
  2.4 Bayesian center-surround salience model
  2.5 Local similarity number model
  2.6 Discussion
3 Eye movements
  3.1 The rudiment
  3.2 Scanpath generation
  3.2.1 Scanpath and visual attention
  3.2.2 Scanpath and oculomotor

9 11 13 15 17 21 27 29 30 32 33 35 37 38 39 39 41 43 43 45 46 47 48 50
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Stochastic bottom-up saccade generation</td>
<td>51</td>
</tr>
<tr>
<td>3.4</td>
<td>Discussion</td>
<td>54</td>
</tr>
<tr>
<td>4</td>
<td>Video saliency</td>
<td>57</td>
</tr>
<tr>
<td>4.1</td>
<td>Spherical center-surround</td>
<td>58</td>
</tr>
<tr>
<td>4.2</td>
<td>Temporal motion model</td>
<td>59</td>
</tr>
<tr>
<td>4.3</td>
<td>Spatio-temporal motion model</td>
<td>61</td>
</tr>
<tr>
<td>4.4</td>
<td>Discussion</td>
<td>63</td>
</tr>
<tr>
<td>5</td>
<td>Applications</td>
<td>65</td>
</tr>
<tr>
<td>5.1</td>
<td>Salience modeling in action</td>
<td>65</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Recognition</td>
<td>65</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Object detection</td>
<td>66</td>
</tr>
<tr>
<td>5.1.3</td>
<td>Compression</td>
<td>68</td>
</tr>
<tr>
<td>5.1.4</td>
<td>Video summarization</td>
<td>68</td>
</tr>
<tr>
<td>5.1.5</td>
<td>Object tracking</td>
<td>68</td>
</tr>
<tr>
<td>5.1.6</td>
<td>Thumbnailing and retargeting</td>
<td>69</td>
</tr>
<tr>
<td>5.1.7</td>
<td>Segmentation</td>
<td>69</td>
</tr>
<tr>
<td>5.1.8</td>
<td>Medical purposes</td>
<td>70</td>
</tr>
<tr>
<td>5.1.9</td>
<td>Graphics and art</td>
<td>71</td>
</tr>
<tr>
<td>5.1.10</td>
<td>Robotics</td>
<td>71</td>
</tr>
<tr>
<td>5.1.11</td>
<td>Others</td>
<td>72</td>
</tr>
<tr>
<td>5.2</td>
<td>Exploiting proposed models</td>
<td>72</td>
</tr>
<tr>
<td>5.2.1</td>
<td>LSN model and tracking</td>
<td>72</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Motion detection &amp; background subtraction</td>
<td>73</td>
</tr>
<tr>
<td>5.3</td>
<td>Eye movement use</td>
<td>74</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Salience modeling</td>
<td>75</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Scene recognition</td>
<td>75</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Valence recognition</td>
<td>76</td>
</tr>
<tr>
<td>5.4</td>
<td>Discussion</td>
<td>77</td>
</tr>
<tr>
<td>6</td>
<td>Evaluating saliency models</td>
<td>79</td>
</tr>
<tr>
<td>6.1</td>
<td>Bias and evaluation</td>
<td>79</td>
</tr>
<tr>
<td>6.2</td>
<td>Salience evaluation metrics</td>
<td>81</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Similarity-based evaluation</td>
<td>81</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Fixation-based evaluation</td>
<td>82</td>
</tr>
<tr>
<td>6.3</td>
<td>Scanpath analysis and salience models</td>
<td>83</td>
</tr>
<tr>
<td>6.4</td>
<td>Benchmarks</td>
<td>84</td>
</tr>
</tbody>
</table>
1 Introduction

The “saliency” (or “salience”) is generally used to refer to the state by which any peculiarity stands out relative to its surrounding. For many years, it has been a subject matter in disciplines of cognitive science such as psychology, artificial intelligence, and neuroscience. Later, it drew attention of computer vision scientists. Particularly, the computer vision community is interested in distinct perceptual quality that makes items stand out from their neighborhood (i.e. “visual salience”).

1.1 Visual saliency and visual attention

The visual saliency a.k.a. saliency was popularized in the computer vision community in the late 90s by Itti et al. (1998). They implemented a biologically-plausible model based on the architecture proposed by Koch & Ullman (1985). Subsequently, more models have been introduced because of salience success in curtailing the processed data and improving efficiency. Eventually, research on saliency became a subject matter in computer vision.

Saliency is a subset of a broader topic named computational model of visual attention (CMVA). It includes a formal description of attention computation. It also covers the span of theoretical studies of brain and behavioral investigations. For instance, a researcher presents a stimulus (e.g., image) to subjects (e.g. human) and observes their behavior to provide convincing proof of a theory, a model or a hypothesis.

In the computer vision community, there is often a confusion between attention and saliency. The two terms are used interchangeably, though there is a profound difference. Attention is a general concept that depends on many cognitive factors. It is easily influenced by the assigned task (e.g., free-viewing and interactive tasks, similar to game playing (Borji et al. 2011b, 2012d)) and the strategy of solving the task. Moreover, subjective factors such as age and experience regulate it, e.g., the encounter of an experienced detective with a crime scene is different from a young policeman or a civilian eyewitness. The mechanism of attention is either expectation-driven top-down (TD) and/or scene-driven bottom-up (BU). It is often speculated that the first is deliberative and task dependent, while the latter is reflexive, fast, likely feed-forward.

1In this thesis, the term saliency/salience refers to visual saliency.
Saliency relies on the perceived visual stimulus (i.e., it is a bottom-up process) or the extracted features from the stimulus which can be manipulated by top-down (TD) cues.

Computer vision perspective promotes categorizing saliency study into two main areas of debate: saliency modeling and saliency segmentation. Although the two are closely related, their objectives are distinct. The goal of saliency modeling is predicting locations that grab attention; saliency segmentation, however, tries to segment the most eminent item in a scene. While the segmentation usually applies to images containing one well-identifiable object (e.g., a bucket against a white background), the salience modeling is challenged by real world images with complicated scenery. Consequently, the evaluation criteria and ground truth are different. Assessment of saliency segmentation is often performed by measuring the precision-recall of the output of methods against the ground truth data. The ground truth data is obtained from explicit judgments of observers who annotate the salient area. On the other hand, evaluation of saliency modeling is sophisticated and relies on observers’ eye movements.

The success of saliency techniques – particularly in the computer vision domain – is greatly due to their inexpensive and fast computation. This facilitates their use as a preprocessing step in many applications. For instance, image/video compression can benefit from variable compression rate which is achieved by higher compression rate on non-salient regions and careful handling of salient area (e.g., Guo & Zhang 2010). Saliency can provide an efficient framework for object recognition in conjunction with selective attention-based methods by underlining the informative regions (e.g., Walther & Koch 2006, Kanan & Cottrell 2010). Similar idea exists in many applications that use saliency such as tracking (e.g., Borji et al. 2012a), content-aware image
re-targeting (e.g., Jacobson et al. 2010), image thumbnailing (e.g., Marchesotti et al. 2009), video summarization (e.g., Ma et al. 2005), etc.

1.2 Objectives

Computer and human shall deal with a tremendous amount of visual information. To efficiently process the visual information, it is fascinating that a human being is capable of narrowing his/her focus on the highlights in a fraction of time. There may exist a variety of theories in cognitive sciences that explain such a mechanism. Among them, however, saliency modeling has recently obtained high reputation in computer vision. Because of the growing and overwhelming influence of salience and eye movements on computer vision, the aim of this thesis is to advance the frontiers by proposing new techniques and exploring their possible applications. Although it may be inspired by the advancements in biological and behavioral studies, this thesis tries to provide solutions enhanced for computer vision.

This thesis is mostly focused on the area of salience modeling. It explores the extent of image and video processing to develop techniques suitable for computer vision. The role of eye movement is investigated in the hope of improving salience modeling. Moreover, eye movements and salience are exploited in applications such as tracking, background subtraction, image category decoding, and valence recognition.

1.3 Contributions

Here, the major contributions of this thesis are summarized as follows:

– A simple and efficient saliency model which exploits Bayesian formulation and relies on a small number of samples taken sparsely around a central pixel. It is locally estimable, fast to compute, and biologically inspired. The description of the model is provided in Paper I.
– A framework to simulate eye movements and generate salience based on the particle filters and utilizes the salience model proposed in Paper I. The results demonstrate that incorporating eye movements will improve the performance of salience modeling. The detail of the model is published in Paper II.
– Assessment of salience and the statistics of eye movements in recognition of image categories and emotional valence. To this end, several features were extracted from
eye movements and utilized in a classification task. Similarly, salience models applied and compared to human performance. Related publications are Papers III & IV.

- A spatio-temporal salience model. A spherical representation utilizes the principles of Paper I and extend it to the video domain. Paper V summarizes this model.
- Salience modeling and background subtraction. The purpose of background subtraction is localizing objects of interest. It usually requires some background model. Contrarily, a general model-free approach is made by proposing application of saliency modeling techniques. Two saliency models are proposed. The relevant Papers are VI and VII.
- Salience modeling and tracking. Application of salience models in tracking existed beforehand. Here, a simple salience model is proposed based on the concept of rarity. Afterwards, a tracking algorithm uses the proposed model to define the tracking target of interest (see Paper VIII).
- A metric of evaluation. It introduces a scanpath based metric of evaluation, which takes regional interaction of image elements into account. The relevant publication is Paper III.

1.4 Outline of the thesis

This chapter summarized a short introduction to the topic. It provided the objective and contributions of the thesis. Chapter 2 familiarizes the reader with the history and background of the saliency modeling. It elaborates the center-surround hypothesis and introduces two salience models. The relevant publications are Papers I & VIII.

Chapter 3 initially discusses eye movements and its statistical characteristics. It is followed by a discussion with an argument on scanpath generation schemes. Eventually, it introduces a stochastic bottom-up saccade generation model. The relevant publications are Papers II, III, & IV.

Chapter 4 discourses the domain of video stimuli. It elaborates three techniques for video which are published in Papers V, VI, & VII, respectively. The first method extends the model of Paper I to handle video stimuli. The two other models, explore the applicability of salience modeling in background subtraction.

Chapter 5 examines the application of salience modeling and eye movements. It initially introduces the span of applications. Afterwards, it explains how the proposed models contribute to a specific application area. The relevant Papers are III, IV, VI, VII, & VIII.
Chapter 6 explains the evaluation criteria and its challenges. It initially talks about the bias in datasets and the metrics of evaluation. Afterwards, it introduces a metric of scanpath evaluation. The relevant Paper is III. Eventually, it is followed by Chapter 7 which provides a brief summary.
2 Image saliency

This chapter provides a broad overview of saliency modeling. Initially, it discusses the origins; and the link between salience modeling and visual attention is elaborated. The second section mentions the computer vision perspective. It provides a taxonomy of salience models. Afterwards, the center-surround hypothesis is explained. Eventually, the models of Papers I & VIII are briefly introduced.

2.1 The history

To dig into the history of salience modeling, one probably has to go back to the 19th century, the time of initial psychological studies on attention. Even though psychologists credit philosophers as the pioneers, the memory of the early era of philosophy evanesces because of the significant contributions of psychologists. Meanwhile, due to its old roots, the term “attention” is so believed to be obvious that James (1890, pages 403-404) writes:

“Every one knows what attention is. It is the taking possession of 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…”

Aside from his vivid language description of attention, his theory laid a sound foundation for modern attention research. Also, differentiation between the sensorial (objects of sense) and intellectual (ideal objects) attention was a remarkable insight facilitating the advancement of the field.

During several years, similar to the term, the field has emerged. It has developed to cover different perspectives and led to different schools of thoughts. Tsotsos et al. (2005) identify several turning points in the evolution of attention models. Among them, some – particularly, more important to the computer community – are signified here. Until the mid-20th century, different perspectives (including psychological, behavioral, neuroanatomical and neurophysiological) are influenced by the discoveries and findings of the 19th century. In the 50s, the sparks of the influence of information processing appears. Broadbent (1958) tries to justify several experimental results on hearing using a filter theory of attention. Although his theory proceeded by several variants
and alternatives, they are mostly founded on the same grounds of treating human as an information processing unit. The late 70s is somehow the beginning of a new era in attention research as Treisman et al. (1977) laid the foundation of “Feature-Integration Theory”.

In the domain of attention, Feature-Integration Theory (FIT) (Treisman & Gelade 1980) is probably one of the most notable theories of the last decades. Furthermore, it can be somehow considered as the origin of many current visual salience models. It states that the visual scene is coded into several dimensions (e.g., color, orientation, spatial frequency, . . . ). While the features (i.e., dimensions’ values) are registered in a parallel manner in an early stage, objects are focally identified later. Hence, in order to form a single object, stimulus locations are processed serially – through a map of locations – to combine the present features.

Koch & Ullman (1985) proposed the most influential model of visual attention, which partially benefits from FIT. The main advantage of this model is that it has a detailed mathematical description, which makes its implementation possible. This theory has three main components as follows: parallelly represented features, a selective mapping, and a shifting mechanism of the focus of attention. In a feed-forward fashion, features are initially extracted and registered in several parallel topographical maps. These maps are fused into a central map to determine the amount of conspicuity of different locations in a visual scene. The central map is called saliency map and resembles the map of locations in FIT. Afterwards, a Winner-Take-All (WTA) neural network selectively transforms the most salient features into a central non-topographical representation and identifies the most salient region. Eventually, a decaying property helps to simulate shifting mechanism of attention. The amount of maximal active unit in the saliency map decays and WTA’s response to the new configuration identifies the new most salient area. This model is notable to us as it significantly helped to the popularization of the term “saliency map”. Later, Itti et al. (1998) provided the first complete implementation of this model which is thus far the most famous and well-cited work in the computer vision community.

It is impossible to neglect the influence of biology of vision on computer vision scientists since the emergence of the field. In computer vision, some of the attempts to model visual attention date back to the early 90s. For instance, Swain & Stricker (1993) argued on the use of visual attention in active vision systems and robotics. Milanese et al. (1994) proposed an attention model to identify regions of interest. Jagersand (1995) utilized the Kullback-Leibler divergence to define an information theoretic
contrast measure and estimate salience to model attention. In computer vision, this research direction flourishes by the publication of Itti et al. (1998). The main reason is that 1) it is the first successful implementation of a testified pure psychological model and 2) it somehow ignited the use of eye movement statistics as an evaluation mean in the community. Accordingly, it grounds the basis of what is referred to as salience modeling.

2.2 Saliency modeling

The idea of salience has somehow existed since the early days of computer vision (e.g., Moravec (1981) used the concept in vision based robot navigation). But, there did not exist a unified agreement on its definition until introduction of computational models of attention to the computer vision community. These models are the strong link between biological vision and computer vision, particularly considering attention modeling (Tsotsos 2011). Taking attention into consideration, visual saliency is the result of a perceptual quality that makes an item stands out from its neighborhood. It is the consequences of interaction between several stimuli. In other words, a feature (e.g., edge, corner, . . . ) is not salient by itself. As an example, an edge is salient if its surrounding is proportionally plain; otherwise, it may not look that salient.

Saliency map is the scalar quantity of saliency at every location in the visual field. It defines spatial distribution of saliency that facilitates the deployment of attention. Sha’asua & Ullman (1988), probably as one of the pioneers in computer vision, used saliency map terminology and concept in order to represent globally salient contours. They applied an iterative scheme using a network of locally connected elements to derive the saliency map. Although it seems that their work is biologically motivated and promising, it is adversely affected by the experiments that do not exceed the domain of intuition. Almost ten years later, the first successful biologically plausible model appears. It is accompanied with a sound quantitative basis for saliency evaluation which tries to find an agreement between human eye movement statistics and saliency map. This analogy is fairly supported by psychophysical and behavioral studies (Itti et al. 1998).

Visual saliency modeling a.k.a. saliency modeling is the transformation of an image into a salience map such that the identified conspicuousness agrees with the statistics of human eye movements. While the process of saliency map computation mostly relies on the stimulus (i.e., it is bottom-up), sometimes, the extracted features can
be partially manipulated by a given task (i.e., top-down cues). Thus, the type of stimulus and assigned task matter as they influence the eye movements. An exemplar of stimulus influence is the emotional stimuli which influence eye movement patterns (Niu et al. 2012). Considering the task influence, the seminal work of Yarbus (1967) is the most recognized study to date (see, Borji & Itti 2014). All the mentioned sophistication makes salience modeling a challenge, but it is, nonetheless, interesting and crucial. It helps broadening the understanding of human cognitive behavior and benefits engineering applications.

There exists numerous models and techniques, more than 30 (see, paper III), in the literature. Toet (2011) divides the models into computational and psychophysical. The first defines any model or techniques that transforms an image into a saliency map and the latter refers to the estimates directly derived from human eye statistics and usable in the assessment of a computational model. Judd et al. (2012) implicitly identifies several categories, including models motivated by computational models of visual attention such as (Koch & Ullman 1985), models with top-down components, Fourier based (i.e., process the image in the Fourier domain), region based, models that learn parameters (e.g., using some machine learning tool to learn the parameters), and models with center-bias. Models with top-down components incorporate sort of higher information which can be contextual, task regulated or object detector based. A model is region based if it considers some local grouping of pixels. Models with center-bias are getting advantage of the phenomena of eye fixation tendency towards the center of an image.

Another categorization is provided by Borji & Itti (2012b) while they reviewed visual attention models. It includes a taxonomy that covers saliency modeling techniques. To them, a saliency modeling technique, depending on its computation methodology and theoretical grounds, belongs to one of the proposed groups of cognitive, information theoretic, graphical models, spectral analysis, pattern classification, Bayesian, decision theoretic, and others. Adopting a similar approach, this study devises the models into several categories discussed below. It provides some of the exemplars in each category. However, it does not have any restriction and a method can belong to more than one category. It is worth to mention that an earlier revision of this taxonomy exists in Paper I.

### 2.2.1 Center-surround models

Many of these models rely on FIT and attention model of Koch & Ullman (1985). They often consist of the three steps of feature extraction, center-surround feature comparison
and conspicuity map fusion. For instance, Itti et al. (1998) subsamples a given image into a Gaussian pyramid. At each pyramid level (i.e., scale), extracted feature channels consist of red, green, blue, yellow, intensity, and local orientations. They apply these features in the computation of conspicuity maps for all the scales in each feature channel. The conspicuity maps are computed by comparing the value of features to the value of the features of their surroundings. Afterwards, the conspicuity maps of each feature channel are combined across all scales to produce another level of conspicuity maps. Eventually, the saliency map is obtained from the linear combination of these conspicuity maps. Walther & Koch (2006) adopt the same approach and extended it to a framework of attention-based object recognition. Frintrop (2006) provided an implementation based on the aforementioned structure for laser sensor data along the images.

Choi et al. (2006) proposed a similar center-surround model. It extends the feature maps by including symmetry features. Furthermore, it added an Independent Component Analysis (ICA) layer to model the visual cortex redundancy reduction. Initially, the features are extracted similarly to the model of Itti et al. (1998). The center-surround difference is computed over the features to produce several conspicuity maps. These maps are convolved with ICA-driven filters and linearly combined to produce the final saliency map.

Le Meur et al. (2006) extended the feature extraction phase to coincide better with Human Visual System (HVS) and psychovisual space. Initially, the image is transformed to Krauskopf’s color space which is validated by psychophysical experiments on color masking. Before extracting the visual features, Contrast Sensitivity Functions (CSF) is applied to boost the sensitivity of chromatic components as in HVS. The extracted features receive a differential visibility threshold modification based on the influence of context (i.e., go through a masking effect). These features are combined through series of center-surround interactions and filtered to produce the final salience map.

Murray et al. (2011) also focused on an appropriate color appearance to efficiently mimic human vision. Their model utilizes Wavelet decomposition to generate the features for center-surround comparison. Eventually, their framework integrates scale through a simple inverse wavelet transform over the set of weighted center-surround outputs to estimate the salience.

Gao et al. (2007) introduced a center-surround process in combination with a decision-theoretic hypothesis. It equates the saliency of an image location with a classification problem that opposes stimuli at center and surround. Eventually, the saliency is locally – in a window with respect to its neighborhood – quantified by the
mutual information between features and class labels. A similar idea is also presented in (Gao & Vasconcelos 2007) which elaborated the consistency of the model with psychophysics.

Seo & Milanfar (2009a) presented another center-surround technique which utilizes the decision-theoretic hypothesis of Gao et al. (2007). Their main contribution is the application of local regression kernels to extract features. Their model measures the similarity of a pixel to its surroundings using a self-resemblance metric. It represents each pixel by a matrix of local regression kernels. Then, it computes the resemblance of each pixel to its surrounding by cosine similarity. The model was later extended to the video domain (Seo & Milanfar 2009b).

2.2.2 Information theoretic models

These models are rooted on the grounds of information theory. They usually treat the salience as the amount of information. Thus, the notion of most salient region refers to the most informative region. These techniques often include one of the quantities of information (e.g., entropy, mutual information, . . . ) as a key ingredient.

Bruce & Tsotsos (2006) proposed a technique based on self-information maximization to compute salience likelihood. Their model is motivated by the premise that “localized saliency computation serves to maximize information sampled from one’s environment” (Bruce & Tsotsos 2009). Hence, salience of an image patch is formulated as its likelihood estimate on the basis of its surround.

Mancas (2007) applied a rarity based scheme in which rarity is interpreted as a close relevant of self-information. The basis of rarity computation is counting the number of similar regions. To measure similarity, it compares histograms of image regions by taking their spatial relation into account. Two types of rarity are defined: Global and Local. Global rarity considers the uniqueness of a region in the whole image. In contrast, local rarity examines uniqueness in a smaller neighborhood.

Hou & Zhang (2008) introduced a model based on Incremental Coding Length (ICL). It defines ICL as a mechanism of energy distribution in the attention system and apply it to estimate entropy gain of features. According to the ICL definition, salient features happen to elicit entropy gain and are therefore assigned high energy. The objective is the maximization of the entropy of the sampled visual features. They demonstrated that their model selectively attends salient locations by identifying features with large coding length increments. A similar idea for information measurement is also promoted by Li
et al. (2009). In the latter approach, the ICL is approximated by finding the sparsest linear representation. The saliency map is similarly dependent on the coding length, though.

Li et al. (2010b) applied conditional entropy to compute saliency and infer proto-objects in video. To them, salience follows a local paradigm, it is the minimum uncertainty (i.e., minimum conditional entropy) of a local region with respect to its surround. Later, they approximate this conditional entropy using the lossy coding length scheme under multivariate Gaussian data assumption. In other words, they assume that the local region and its surround are both multivariate Gaussian data.

The application of information theory is not limited to the computation of salience map as described above. There exist techniques that utilize measures of information as a means of conspicuousness aggregation, scale selection, or distance metric (e.g., Kullback-Leibler divergence). It is sometimes difficult to include them in this class category, depending on the relevance of contribution, here are some of them. Ban et al. (2008) & Wei et al. (2010) proposed a model based on maximum entropy to compute dynamic saliency (i.e., saliency for video). They treat entropy as a utility of aggregation to obtain the final salience map from several conspicuousness maps. Kadir & Brady (2001) proposed a multiscale algorithm that uses maximum entropy as a scale selection mechanism.

### 2.2.3 Frequency domain models

In this family of models, an image is processed in the frequency domain in order to infer salience. Hou & Zhang (2007) proposed spectral residual saliency model. They analyzed the log-spectrum of an image to compute saliency. They denoted that the statistical singularities in the spectrum are probably responsible for anomalous regions in the image where proto-objects become conspicuous (i.e., saliency).

Guo et al. (2008) demonstrated that the phase spectrum outperforms the amplitude transform in salience prediction. The model is further extended to a quaternion representation of an image combining intensity, color and motion features. Also, it is applied to image and video compression tasks by taking advantage of the multiscale representation of the wavelet (Guo & Zhang 2010).

Bian & Zhang (2009) introduced Spectral Whitening (SW) model. The idea behind the model is that visual system bypasses the redundant features while responding to rarity. An image is transformed to the Fourier domain and is normalized with respect to
its amplitude. Afterwards, the salience is obtained from the smoothing of the inverse Fourier transform of the whitened signal. Later, the same idea is applied to a more sophisticated framework (Bian & Zhang 2010) in which salience relies on the divisive normalization in the frequency domain.

Li et al. (2011) proposed a model that combines the information of frequency domain analysis with spatial domain local information. In the frequency domain, the focus is on the estimation of repeating patterns which are non-salient. This model argues that the spikes in the amplitude spectrum corresponds to repeating patterns. Thus, spectrum smoothing suppresses these indistinct patterns and produces frequency based saliency. Also, the spatial domain analysis enhances those regions that are more informative in a center-surround manner and results in salience. The output of the two mentioned channels (frequency and spatial) are combined to produce the final saliency map. Later, Li et al. (2013) proposed a hypercomplex representation of the frequency channel of the mentioned model. Moreover, they introduced a scale-space analysis to utilize the best scale of spectrum smoothing.

Hou et al. (2012) introduced the image signature as a foreground quantifier. They experimentally demonstrated that image signature overlaps with visual conspicuousness by developing a salience algorithm based on it. Although the image signature is grounded on the advanced theoretical framework of sparse signal mixing, it is as simple as the signum of the Discrete Cosine Transform (DCT) of a given image. To produce a saliency map, a foreground image is initially obtained from the reconstruction (i.e., inverse DCT) of its image signature. Subsequently, it smooths the self-Hadamard (entrywise) product of the foreground image to produce the saliency map.

One can apply different transform functions such as Fast Fourier Transform (FFT), DCT, or their quaternion representation to analyze an image in the frequency domain. Schauerte & Stiefelhagen (2012b) argued that the quaternion representations perform relatively better in comparison to FFT and DCT. Later, Schauerte & Stiefelhagen (2012a) proposed the quaternion DCT representation of the model of Hou et al. (2012).

Although frequency domain techniques are successful, their biological plausibility is not well investigated. To date, it seems no one has proposed a concrete biological foundation for them.
2.2.4 Top-down manipulated models

Here, the discussed saliency maps are influenced by a kind of higher information. This information can be contextual and/or task regulated. Also, there are some salience modeling techniques that get advantage of human attention to some particular items (e.g., faces). They usually incorporate the output of an object classifier to boost the performance of their model.

Here, an example is provided to familiarize the reader with the concept of task regulation. Borji et al. (2010b) proposed an online learning scheme based on salience concept in interactive environments. Their model has three layers. In the first layer, it computes a saliency map and select the most salient point. The next layer performs a task specific operation (e.g., object recognition) on the selected salient point. Based on the decision outcome of the operation, it builds an object-based binary tree in the last stage. The internal nodes of this tree decide about existence of objects and regulate the salience map while the leaves correspond to a world state. The tree is updated interactively – consequently, the salience map – because it suggests an agent about the state of the world in which a feedback signal is provided upon agent action. Although this kind of interactive scenario probably helps understanding the concept of the top-down regulation of salience maps, it is very difficult to classify it – also any of its variants (Borji et al. 2011b, 2012c,d, 2013b) – under the umbrella of saliency modeling as they almost build a full attention model. The next examples are those that are closer to the topic of salience modeling and a salience map is part of the system output.

Oliva et al. (2003) demonstrated the use of saliency maps and contextual priors in object detection. Having a specific assigned task (e.g., pedestrian detection), a specific region (e.g., horizon) of the image is more likely to contain the object of interest (e.g., pedestrian). Hence, the salience map is modulated by contextual knowledge about the task. In this case, the final salience map is the multiplication of contextual prior and stimuli-driven salience map. While salience is treated locally, contextual priors are inferred from global features. This idea was later developed to a full contextual guidance model of attention (Torralba et al. 2006).

Ma et al. (2005) proposed a video content summarization based on saliency. The model applies top-down factors by extracting contextual priors from semantic cues (e.g., face, speech, and camera motion). In their model, the contextual priors are also treated as conspicuity maps. To generate salience map, the model initially produces several conspicuity maps (including priors). Afterwards, it produces a comprehensive
salience map by nonlinear fusion of these maps. Eventually, the salience map may be used as importance ranking or the index of video content.

Navalpakkam & Itti (2006) proposed a model that integrates top-down influence and stimuli-driven salience. The stimuli-driven component computes the visual salience of scene locations in different feature maps extracted at multiple spatial scales. Further, the salience map is tuned by the accumulated statistical knowledge of visual features of the desired search target, to maximize target detection speed. This is achieved by maximizing the signal-to-noise ratio (SNR) of the target versus distractors which results in optimum linear weights for feature combination.

Peters & Itti (2007) trained a classifier to capture the task-dependent association between a given scene and human eye movement data. For a given image or video frame, the classifier determines the top-down prior. The final saliency map is modulated using the inferred top-down prior. Similarly, Li et al. (2010a) introduced a probabilistic multi-task learning approach in which task-related “stimulus-to-salience” mapping functions are learned. Their model also learns different strategies for fusing top-down information and stimulus-driven salience.

Zhang et al. (2008) formulated salience in terms of stimuli-driven salience and top-down information. The top-down information consists of a location prior and a likelihood to favor features which are consistent with our knowledge of the target. Kanan & Cottrell (2010) introduced a classification framework which extended this model to simulate a focus of attention mechanism and demonstrated its success in a variety of classification tasks.

Gao et al. (2009) provided an extension to (Gao et al. 2007) for a classification task. It mostly relies on top-down information. The saliency is defined as the confidence with which locations in the visual field can be classified as containing stimuli drawn from the target class. The model is successfully applied in object localization and image classification.

As mentioned earlier, there are models that try to boost the performance by incorporating some object classifier output. These models are motivated by the findings that suggest a human pays attention to some specific items such as faces and text independent of the task (Cerf et al. 2009). Judd et al. (2009) proposed a model that follows the same track. It consists of three feature channel of low (e.g., intensity, color, etc), mid and high-level. The mid-level feature is a horizon line which is assumed to be the resting bed of objects. The high-level features are faces and people. These features are combined by training a linear Support Vector Machine (SVM) classifier to relate them to human eye
movement. Borji (2012) extended this idea by considering more object classes and explored performance of different classifiers.

### 2.2.5 Connection based models

Connection based models consider a structural relation between image elements. They model salience as the emergence of interactions of interconnected image elements. These connections can be represented by neural networks or graphical models in which rudimentary elements correspond to properties of pixels or regions.

Harel et al. (2007) introduced a model that utilizes fully-connected graphs. Initially, it extracts several feature maps. Afterwards, it adopts a Markovian approach towards salience computation. Building a fully-connected graph over image locations, weights between two nodes are assigned proportional to the similarity of feature values and their spatial distance. The most dissimilar node of the graph will gain the most salient value by defining a Markov chain on the graph. It is done by normalizing the weights of the outbound edges of each node to 1; drawing an equivalence between nodes and states and between edges weights and transition probabilities.

Gopalakrishnan et al. (2009) provided a model that resembles (Harel et al. 2007), the two models are different in details, however. It considers both global and local properties of the image. The first is obtained from a random walk on a complete graph and the second is extracted from a random walk on a k-regular graph. Eventually, the salience map is determined in terms of most globally salient node (i.e. most globally isolated) that falls on a compact object (i.e., easily reachable in a local neighborhood).

Wang et al. (2010) employed the information maximization principle to measure visual saliency on a network of random walks. The model extracts several sub-band feature maps using learned sparse codes. On a complete graph, it runs a random walk to simulate the information transmission between interconnected elements of these feature maps. They proposed a salience measure called “site entropy rate” which computes the average information transmission between a node and all the other nodes.

Pang et al. (2008) presented a stochastic model that utilizes dynamic Bayesian networks to predict where humans typically look in a video. The model consists of deterministic saliency maps and stochastic saliency maps. While the deterministic saliency maps present the current video frame salience, the stochastic saliency map integrates the information from the past into them. In consequence, a more realistic saliency is computed which is later used to predict likely locations one person may
look. Miyazato *et al.* (2009) introduced a variation of this model that improved the performance and reduced the execution time.

Avraham & Lindenbaum (2010) proposed a model called “Extended Saliency” which utilizes graphical models. It starts with a rough preattentive segmentation and uses a tree-based Bayesian network to approximate the likelihood of segments’ attractiveness. To compute the saliency, they introduced a six-step algorithm which consists of 1) candidate selection by segmentation, 2) candidate prior assignment, 3) similarity measurement between candidates, 4) label dependency representation using DBN, 5) selection of N most likely joint assignments, 6) salience deduction by marginalization.

These models are very successful as reported. However, their major drawback is that it is difficult to replicate them. They are often complicated and implementing them requires attention to a lot of minute details. Hence, except those that made their code available, these models are less renowned.

### 2.2.6 Learning based models

These models try to establish a relation between either low-level features and human eye movement statistics or the features themselves. It can happen by learning a classifier, some parameters, and/or some priors. They differ from the top-down manipulated models with respect to the training data. In other words, there is no higher information present.

Itti & Baldi (2005a,b, 2009) introduced a model based on surprise. They defined surprise as a measure to assess how data affects an observer in terms of difference between posterior and prior beliefs about the world. The premise is that given a prior distribution of belief, the fundamental effect of a new data observation on the observer is to change the prior into a posterior distribution. In fact, if the posterior is identical to prior, there exists no surprise. The model learns a prior using features extracted in a center-surround fashion; this prior acts as a building block of the posterior distribution. This concept applied to both space and time.

Kienzle *et al.* (2009) fitted a nonlinear model to associate image patches with human eye movement statistics. They extracted a $13 \times 13$ patch which is stored in a 169-dimensional feature vector. Each patch is accompanied with a binary label denoting relation to eye fixation. A nonlinear model is fitted to the training data using SVM. Afterwards, a perceptive field is extracted from the learned model. The perceptive field
identifies the most relevant image patterns that guide fixation selection. The saliency map is estimated by combining four primitive perceptive fields.

Zhao & Koch (2011) learned the optimal weights for fusing conspicuity maps from observers’ eye movement data. Also, they learned a bias toward fixating at the center of the image. They also extended this model to investigate a nonlinear learning framework (Zhao & Koch 2012).

The bias toward fixating at the center of the image is a very general prior that one can learn from eye movement statistics. There is evidence that shows such a bias boosts the performance of salience modeling algorithms (Zhang et al. 2008, Judd et al. 2009, Borji et al. 2013a). While some people may assume a simple Gaussian prior, others may learn it from the statistics of human eye movement. For instance, Yang et al. (2010) learned the center prior along with a prior to choose the optimum size of the surrounding region from human fixation information.

### 2.2.7 Miscellaneous models

Salience modeling has a broad span. Some of the proposed models are difficult to assign to the aforementioned groups. Also, there are few models similar to them which makes categorization difficult. As an example, Garcia-Diaz et al. (2009a,b, 2012) introduced a model that mostly relies on feature whitening and normalization. First, the color space is whitened. Then, orientation features are extracted from these whitened color channels. The extracted feature maps are decorrelated, whitened and normalized to produce oriented conspicuity maps. The fusion of these maps also goes through several normalization steps.

### 2.3 The center-surround hypothesis for saliency

Taking the visual salience definition into account and consistent with Itti et al. (1998) and Paper I, for an item that is located in a region called center, the center-surround hypothesis states that an item (i.e., center) is salient with respect to its surround if and only if the features of the item contrast with the features of its surround. Consequently, the salience of an item denoted as Sal\_i corresponding to the center C\_i (i.e., the salience of the center) with respect to its surrounding S\_i is defined as

\[
Sal_i = D(C_i, S_i),
\]

(1)
Surround
Center
Fig 1. An example to represent center-surround concept. (a) an image patch is divided into center and surround, (b) the realization of the center-surround concept on a real image. (Reprinted from Paper I, Copyright (2011), with kind permission from Springer Science and Business Media).

where $D(\cdot, \cdot)$ is a distance function that measures the amount of dissimilarity between center and surround. Fig 1 depicts the concept of center-surround in which a rectangular image patch is divided into a central region and a surrounding area.

Different methods apply to measure the aforementioned contrast. Itti et al. (1998) defined the center-surround salience as the across-scale point-by-point subtraction of features between fine and coarse scales in order to compute conspicuity maps. Erdem & Erdem (2013) investigated two covariance distance metrics to measure the region covariance dissimilarity of image patches. Gao et al. (2008) elaborated the biological plausibility of formulating center-surround salience in terms of mutual information. In their formulation, Kullback-Leibler divergence defines the contrast between center and surround. Huang & Ahuja (2012) tried to unify the center-surround framework by defining contrast function as divergence between probability distributions of center and surround.

Scale is a challenge to many algorithms in computer vision. To deal with this problem, some get advantage of a priori information about the target of interest. But, there are cases that no a priori data exists. Salience modeling falls into this latter category, since there exists no assumption about the salient item. In the center-surround hypothesis, the scale is defined as the size of the local neighborhood that the salience is computed with respect to it. In other words, it is defined in terms of the size of center and the size of surround. Depending on the model, the center’s size can vary from a pixel (e.g., paper I) to an image patch of arbitrary size (e.g., Erdem & Erdem 2013). Similarly, the size of surround can vary from a pixel to the whole image.
In the arena of salience modeling, there exist three approaches towards the scale challenge. Multiscale salience modeling is one popular method. Several conspicuity maps are usually computed from several scales. These maps are fused to produce a master saliency map (e.g., Itti et al. 1998, Murray et al. 2011, Paper I). Proper scale selection is another strategy in which an algorithm deduces the best scale by analyzing the conspicuity maps of each scale (e.g., Kadir & Brady 2001, Li et al. 2013). The last scheme is the use of one arbitrary scale that is usually provided by the author (e.g., Seo & Milanfar 2009a, Hou & Zhang 2007). The main motivation for such an approach is that multiscale analysis is sometimes computationally expensive or does not add to the performance of the model.

**Global and local salience** (Mancas 2007) is a concept that is partly related to scale. In general, an item is globally salient if it is eminent with respect to all the elements of the image. On the other hand, it is locally salient if it stands out with respect to a local neighborhood. In the center-surround domain, this is in direct relation to the size of surround. The global saliency is computed using a whole-image-sized surround (e.g., Borji & Itti 2012a). Otherwise, it is local saliency.

### 2.4 Bayesian center-surround salience model

This section discusses the principles of Paper I. The proposed model provides a Bayesian formulation for saliency modeling under the umbrella of center-surround hypothesis. Defining each pixel $x$ as a pair of $(x, y)$ coordinates, it adopts a circular center-surround representation (see, Fig 2(a)) in which the center is denoted by C and S identifies the surround.
Let’s assume $I(x)$ is an image patch centered at $x \in C$. $H_x$ is a binary random variable defining saliency at coordinate $x$ as follows:

$$H_x = \begin{cases} 
1, & \text{if } x \text{ is salient} \\
0, & \text{otherwise.}
\end{cases}$$ (2)

The amount of conspicuity at $x$ relies on $p(H_x = 1 | I(x), x) \equiv p(1 | I(x), x)$. Knowing that the salience is related to the feature and the position (Tatler 2007), Bayesian rule applies to write

$$p(1 | I(x), x) = p(1 | x)p(I(x) | 1, x) / p(I(x) | x).$$ (3)

By expanding the denominator to $p(I(x) | x) = p(I(x) | 1, x)p(1 | x) + p(I(x) | 0, x)p(0 | x)$, estimation of (3) requires $p(I(x) | H_x, x)$ (i.e., $p(I(x) | 1, x)$ and $p(I(x) | 0, x)$). To obtain these quantities, it adopts the discriminant hypothesis of Gao et al. (2007) and utilizes a non-parametric kernel density estimation to obtain $p(I(x) | H_x, x)$. Moreover, estimation of $p(1 | I(x), x)$ is performed under two assumptions of 1) singleton center (i.e., $C$ consists of one pixel), and 2) the surround consists of a sufficiently small number of pixels which are scattered at distance $r$ from the center (see, Fig 2(b)). Thus, given $n \in S$ samples scattered at distance $r$ from $x$, denoted $x_{i,r}$, it approximates the amount of conspicuity at $x$ as

$$p_n^r (1 | I(x), x) = \left( 1 + \frac{\sqrt{2\pi}\sigma_C}{np(1 | x)} \sum_{i=1}^n \mathcal{G}(I(x) - I(x_{i,r})) \right)^{-1},$$ (4)

where $\sigma_C$ is the standard deviation corresponding to estimation of $p(I(x) | 1, x)$ and $\mathcal{G}(.) = (\sqrt{2\pi}\sigma_S)^{-1} \exp(- ||.||^2 / 2\sigma_S^2)$ is a Gaussian kernel. Eventually, the final saliency is defined as

$$Sal_x = \mathcal{N} \left( \mathcal{G}_\sigma * \sum_{n,r} p_n^r (1 | I(x), x) \right)^\alpha,$$ (5)

where $\mathcal{N}(.)$ is a normalizing function that normalizes values to $[0,1]$, $\mathcal{G}_\sigma$ is a Gaussian smoothing kernel of standard deviation $\sigma$, * is the convolution operator and $\alpha$ is an attenuation control parameter.

This model is capable of incorporating any kind of prior information. In essence, $p(1 | x)$ can represent any a priori knowledge about the location of a salient object. Nonetheless, the proposed model uses only a center-bias prior. It learns a normal
bivariate Gaussian distribution, which constitutes the center-bias prior, using eye fixation statistics. It is worth mentioning that once \( p(1|x) \) is obtained, one can compute \( p(0|x) = 1 - p(1|x) \).

### 2.5 Local similarity number model

This section introduces the model of Paper VIII. The proposed model is particularly developed for tracking. It relies on the operator, named Local Similarity Number (LSN), which gets advantage of the center-surround differences.

Given an image patch \( I(x) \) centered at \( x \in C \), the LSN operator counts the number of similar pixels around \( x \). It is defined as follows:

\[
\text{LSN}_n^r(x, d) = \sum_{i=1}^{n} \mathcal{T}(D(I(x), I(x_i)), d),
\]

where \( x_i, r \in S \) is the \( i \)th pixel at distance \( r \), \( D(.,.) \) is a distance function, \( d \) is a tolerance value, and \( \mathcal{T}(a, d) = 1 \) if \( |a| \leq d \); otherwise, it is zero.

Considering the tracking application, it will be shown that (6) is adequate to define the target of interest. Nevertheless, one can extend it and define the salience as

\[
\text{Sal}_x = \mathcal{N}\left(\left(\mathcal{G}_\sigma \ast \sum_{n,r,d} e^{-\text{LSN}_n^r(x, d)}\right)^\alpha\right),
\]

where \( \mathcal{N}(.) \) is a normalizing function that normalizes values to \([0,1]\), \( \mathcal{G}_\sigma \) is a Gaussian smoothing kernel of standard deviation \( \sigma \), \( \ast \) is the convolution operator and \( \alpha \) is an attenuation control parameter.

### 2.6 Discussion

This chapter provided a concise history of salience modeling and discussed its origins and organization. While preserving the constitution of attention modeling, it tried to provide a direct insight into salience modeling challenge. To this end, it categorized the models into seven categories. In each class, it introduced and discussed the related techniques. Furthermore, it discussed the grounds of center-surround models and expatiated on the related hypothesis. It provided the details of Bayesian center-surround salience model and LSN salience model, which were introduced in Paper I & VIII, respectively.
The Bayesian center-surround model provides a Bayesian formulation to compute salience. In order to produce the result, it adopts a discriminant approach which has a biologically plausible foundation (see, Gao et al. 2008, Gao & Vasconcelos 2009). Thus, while the proposed model is not a direct derivation from biology, it provides an efficient approximation to a well-established biologically linked technique.

The Bayesian center-surround is formulated so flexible that makes it possible to implement variant derivatives and extensions. The two most notable variants are 1) variable surround and 2) top-down manipulation. Although the Bayesian center-surround model uses a fixed-size surround, which means the samples \( x_{i,r} \in S \) are at a fixed distance \( r \) from the center, this model can utilize a variable surround. In other words, each sample \( x_i \in S \) can be located at an arbitrary distance of \( r_i \). It is easy to manipulate the proposed Bayesian model by any top-down task because of \( p(1|x) \). For instance, in an interactive scenario similar to (Borji et al. 2010b), the amount of \( p(1|x) \) can be adjusted to reflect the new circumstance. Furthermore, while the original implementation uses CIE-Lab color space as feature, one can use any feature vector of preference.

The local similarity number model is particularly developed for tracking application. It is formulated in terms of a center-surround difference operator, called local similarity number. This operator simply counts the number of similar pixels scattered around a central pixel. It uses a representation similar to the representation of the Bayesian center-surround model. Nonetheless, its origin lies on texture analysis because the LSN operator is in essence a simplified texture operator with links to local binary patterns (Ojala et al. 2002) (see, Paper VIII for details).
3 Eye movements

The research on eye movement is proceeding at a brisk pace and has influenced computer vision field. Nonetheless, the origin of the research on eye movement is somehow opaque and less known. In a recent effort, Wade & Tatler (2005) tried to trace the history of eye movement research and demonstrated various aspects of it. Similar to the studies on attention, the origin of eye movement research is older than the emergence of computer sciences. In this study, the focus is on the role of eye movement in computer vision with respect to salience modeling.

In the 90s, Reinagel & Zador (1997, 1999) studied the relation between eye movements and natural image statistics in a free-viewing task. They reported that, typically, spatial frequency content is significantly higher at a fixated location than a random location. Later, Itti et al. (1998) utilized this finding to analyze salience performance in terms of a measure of spatial frequency content, which is the average of the numbers of non-negligible FFT coefficients in $16 \times 16$ patches extracted from a given location at each feature channel. They inferred results similar to Reinagel & Zador (1997) which indicated their model, like human, is attracted to informative locations. This is probably the first try to assess a model by establishing a relation, though implicit, between eye movement statistics and salience models.

Motivated by the necessity of direct mapping of human perceptual behavior to the computer vision domain, Privitera & Stark (2000) recorded eye movements as a ground truth to assess a region of interest detector. By the advent of technology and popularity of eye tracking devices, many datasets have become available to address the relation of salience and eye movements (e.g., Bruce & Tsotsos 2006, Le Meur et al. 2006, Judd et al. 2009, Ramanathan et al. 2010). Each dataset is unique because of its comprising attributes, e.g., stimulus type and quantity, the duration of stimulus presentation, number of subjects (i.e., observers), observer distance to the display, demographic information of observers (e.g., age, gender, religious views, . . . ), and experimental setup. Despite all the nuances of the experiment apparatus, these datasets share a common bed which is the eye movement data. Hence, this chapter provides a closer look into the eye movement data.
The following section explains the ingredients of eye movement data. Afterwards, the chapter continues with an introduction to the techniques for scanpath generation. This is followed by a brief explanation of the model introduced in Paper II.

3.1 The rudiment

Eye-tracking based datasets usually provide similar data in slightly different formats. The eye movement data consists of an ordered set of fixations and saccades. Fixations are image locations where the human observer maintains his/her gaze for a reasonable time period. Each fixation \( f \), denoted as \( f = (x, y, ts, te) \), is associated with a location, fixation_location\( f = (x, y) \), and has a duration characteristic defined as fixation_duration\( f = te - ts \) where \( ts \) and \( te \) are the start time and the end time of the fixation occurrence, respectively.

Saccades are fast movements of the eye. A saccade is defined as \( s = (x_s, y_s, x_e, y_e, ts, te) \) where \( x_s \) and \( y_s \) denote the start location, \( x_e \) and \( y_e \) represent the end location, \( ts \) is the start time and \( te \) is the end time of its occurrence. As depicted in Fig 3, each saccade \( s_i \) is associated with several properties. Some of them are as follows:

- Duration; it is defined as saccade_duration\( s_i = te - ts \).
- Slope; it is the direction and steepness of the straight line that fits to the saccade path. It can be estimated by the line that connects the start of the saccade and its end, written as saccade_slope\( s_i = (y_e - y_s)/(x_e - x_s) \), (see, tan(\( \theta \)) in Fig 3(a)).
- Length; it identifies the dimension of the strait line that fits to the saccade path. It can be approximated as saccade_length\( s_i = \| \vec{s}_i \| \) where \( \vec{s}_i = [x_e - x_s, y_e - y_s] \) and \( \| . \| \) is the l2-norm.
- Velocity; it measures the rate of fixation location change which is estimated by saccade_velocity\( s_i = \| \vec{s}_i \|/(te - ts) \).
- Orientation; it defines the angle between two succeeding saccades computed as saccade_orientation\( s_i, s_{i-1} = \cos^{-1} \left( \frac{(\vec{s}_i \cdot \vec{s}_{i-1})}{\| \vec{s}_i \| \cdot \| \vec{s}_{i-1} \|} \right) \), (see, \( \phi \) in Fig 3(a)).

Probably, fixation location is the most well recognized property of the eye movements data in the salience modeling domain. It has a crucial role in most of the learning based models and evaluation of salience modeling techniques. Fixation location is often represented as a fixation map or a fixation density map. A Fixation map is in essence a binary image indicating fixation locations of all subjects. Fig 3(d) visualizes a fixation map overlaid on its corresponding image. The Gaussian filtered version of fixation
map is called fixation density map (see, Fig 3(e)). It is likely to be the most similar to the salience map feature which can be extracted from the eye movements. Therefore, it is sometimes the basis of similarity based evaluation criteria (Le Meur & Baccino 2013). Furthermore, it can be used as a feature in some recognition tasks (e.g., Borji & Itti 2014). While fixation map and fixation density map provide a realization of the most attended location, saccade information provides a good insight into each subject’s viewing behavior. Fig 3(c) depicts the scanpath — a sequence of eye movement data containing trajectories of the eyes — of an observer which is a rough visualization of saccade data and envisions shifts in the observer’s gaze.

### 3.2 Scanpath generation

Scanpath generation can be studied from two perspectives, visual attention and oculomotor concept. The first perspective is closely related to the concept of the focus of attention shift, either covertly or overtly. While covert attention considers the shifts of the focus of attention in the absence of eye movements, the overt attention involves eye movements (Itti & Koch 2001). Although it seems to be a link between overt and covert attention (e.g., Peterson et al. 2004), the true interaction mechanism is vague and unknown.

The oculomotor concept covers the models that generate a scanpath which has the statistical characteristics of human scanpath. Indeed, the emphasize is mostly on the statistics of the saccadic eye movements (e.g., Brockmann & Geisel 1999). The research
in this area provides one of the vital elements of the overt attention (i.e., eye movements). Notwithstanding the intimate link between the two perspectives, it seems necessary to keep them distinct because it makes the study of a complex biological mechanism simpler. This study categorizes models which has no effect on the salience space as oculomotor. To be accurate, oculomotor models often perform an exploration on a saliency map. On the other hand, in an overt attention model, it expects that sensor shift affects the formation of salience. For instance, Dominey & Arbib (1992) introduced a biological plausible model of overt attention that has a compensatory salience shifting property.

### 3.2.1 Scanpath and visual attention

The eye movement based datasets are dependent on the human eye gaze. An eye-tracker follows the position of the eye and measures the eye movement in order to obtain the point of gaze. It eventually produces the fixation and saccade data from the point of gaze information. While fixations somehow represent the focus of attention, saccades indicate gaze-shift towards the next interesting point. The combination of fixations and saccades partially describes the behavior of the observer’s visual attention. This behavior is the intention of the models of visual attention. Thus, the models of visual attention try to replicate the focus of attention shift Clark & Ferrier (1988). To this end, they usually produce an artificial scanpath, which is a crude approximation of eye movement trajectories, by inferring an ordered set of most salient locations.

In attention models, scanpath generation usually consists of 1) a focus of attention selection mechanism, 2) attention shifting mechanism. For instance, Itti et al. (1998) implemented a salience-based bottom-up visual attention model in which initially a master saliency map is computed. Afterwards, a winner-take-all neural network forms a distributed maximum detector (Koch & Ullman 1985) which picks the most salient location. The attention is deployed to this maximum salient point. An inhibitory tagging mechanism, which is called Inhibition Of Return (IOR), suppresses the currently attended location and facilitates selection of the next most salient location. Repeating this process generates a scanpath and provides a mechanism that probably describes attention deployment within the first fractions of seconds after a scene is presented. Hence, it is expected that a complete attention mechanism includes top-down and
volitional biasing. Nonetheless, here, the focus is on the bottom-up attention models in the domain of computer vision.

Wang et al. (2011) introduced a technique which consists of three key components: visual working memory, reference sensory responses, and foveal imaging. The visual working memory accumulates the fixated locations, meanwhile, forgetting them at a set rate. This is somehow equivalent to a computational inhibition of return, which prevents visiting the same location for some time. The reference sensory responses simulate neural responses of the primary visual cortex to an image at a uniform high resolution. On the other hand, foveal imaging provides a detailed response around a fixation location and a coarse response in the periphery of fovea. This response is useful in updating the visual working memory. Eventually, the information discrepancy between reference sensory responses and visual memory facilitates the selection of the next fixation location.

Liu et al. (2013) estimated scanpath by learning semantic content with Hidden Markov Models (HMM)s. In an HMM structure, they proposed that 1) the hidden states can represent latent semantic concept, 2) the prior distribution of the hidden states describes visual attraction to the semantic concept, 3) the transition probabilities represent gaze shift patterns. Hence, it learns the semantic content in three steps: 1) decomposing the image into several region segments, 2) representing each region by a Bag-of-Visual-Words (BoW) descriptor, and 3) learning an HMM over the regions. In order to generate a scanpath, it fuses the semantic content with a spatial position prior, which is a 2D Cauchy distribution, and a salience-based transition probability prior.

Hou & Zhang (2008) introduced a dynamical mechanism which is argued to mimic the inhibition of return phenomenon. It utilizes a saccade generation scheme in which the premise is the correlation of the location of succeeding fixation with a Laplacian distribution over time. In this mechanism, the two influential factors are non-homogeneous composition of features and foveal structure. The first one fosters fixation transition via feature preference and the latter provides a sampling bias. In consequence, their interplay results in an active search behavior and saccadic eye movement functioning.

---

2These models may have been inspired by some biological motivations, but none has a claim to be biologically plausible.
3.2.2 Scanpath and oculomotor

The oculomotor concept investigates mechanisms that can replicate the movements of eyes. The objective is generating a scanpath that often induces the statistics of eye movements. In this section, most of the discussed models consider the eye movements as a sampling phenomenon which often occurs involuntarily.

Brockmann & Geisel (1999) introduced a model to generate scanpath in a natural environment. The scanpath was interpreted as the realization of a stochastic process with non-local transition probabilities. Therefore, the proposed model utilized a random walk on a salience field to derive the probability of fixating a location at discrete time steps. It studied different saccade magnitude probability functions. Eventually, it demonstrated that a most efficient way of generating a scanpath somehow follows a power law distribution in the frequency of occurrence of shift magnitudes.

Boccignone & Ferraro (2004) adopted the findings of Brockmann & Geisel (1999) and introduced Constrained Levy Exploration (CLE) algorithm. It uses a Markov Chain Monte Carlo (MCMC) approach, particularly a Metropolis-Hastings algorithm, to search the salience space produced by the model of Itti et al. (1998). To propose the location of a succeeding fixation, it produces a jump with a random length which is obtained from a weighted Cauchy-Levy distribution. This jump is accepted if it lands on a more salient location than the origin. Otherwise, the acceptance depends on a randomly generated number. Meanwhile, their model tries to explore the whole salience space by generating random saccades towards less salient regions, it often prefers jumps towards more salient locations. Consequently, it is prone to neglecting some far away candidate locations and sticking to a region. To compensate, Levy Hybrid Monte Carlo (LHMC) algorithm was introduced (Boccignone & Ferraro 2010) in which an independent extra set of momentum variables makes the exploration of the salience space easier. Later, Boccignone & Ferraro (2011b, 2012) integrated a set of deterministic rules into CLE and improved the robustness. Eventually, the model was extended to cover scene information, fixation duration, and inhibition of return (Boccignone & Ferraro 2013).

For time-varying scenes, Boccignone & Ferraro (2011a) proposed an active sampling strategy in which gaze relocation is under the influence of a stochastic differential equation whose noise source is a mixture of \( \alpha \)-stable distributions. Nonetheless, they use only a binary switch to alternate between Gaussian and Cauchy distributions. At each time step, it models the oculomotor state in a gaze fixation with respect to the state
of the world. The model demands an observation likelihood to judge the world. To obtain the observation likelihood, the image is initially subdivided into several windows. Afterwards, the likelihood is obtained based on the number of proposed candidate locations in each window and their corresponding salience value. The dynamical behavior of the system is determined by the jump type (i.e., Gaussian vs. Cauchy) and its occurrence probability at the current state. Later, Boccignone & Ferraro (2014) generalized this framework to include a variety of \(\alpha\)-stable distributions and treat different eye movements.

### 3.3 Stochastic bottom-up saccade generation

This section briefly introduces a Stochastic Bottom-up Saccade Generation (SBSG) and fixation prediction scheme. The model is originally first published in Paper II. Fig. 4 depicts a conceptual description of the proposed model. At each time step, a mechanism generates a saccadic behavior that shifts the sensor (i.e., eye) resulting in a new sensor response. This response will be interpreted and a fixation status is decided which is followed by an update to the visual memory. The visual memory keeps track of the latest status of the eyes, including motor and response information.

Despite the proposed model is mostly motivated by the findings on the contributions of saccadic movements to human perception, which promote the incorporation of eye movements in salience modeling (e.g., Tatler & Vincent 2009), it has links to both aforementioned perspectives of eye movements. On the one hand, it shares ideas with oculomotor concept on the treatment of eye movements by considering it a sampling mechanism. This is somehow supported by the temporal blindness investigations, in which it is demonstrated that the absence of eye movements can result in vision loss (Land et al. 2002).

On the other hand, under the umbrella of overt attention, the proposed model is related to the area of visual search in which, in its most common form, the task is the localization of a target as rapidly as possible (e.g., Geisler & Cormack 2011). For instance, Najemnik & Geisler (2005) introduced a Bayesian ideal observer for optimal simulation of eye movement behavior in a target search task. In this perspective, contrarily to the oculomotor models which explore salience space for no reason, the proposed model can be interpreted as a model that searches stimulus space to find the maximum saliency meanwhile it dynamically computes the saliency.
Motivated by the influence of immediate preceding saccade and fixation properties on a saccade (e.g., Hooge et al. 2005, Unema et al. 2005, Tatler & Vincent 2009), the premise is that eye movement can be modeled as a Markovian process of order one. Hence, the proposed model utilizes a stochastic Bayesian filtering approach, in particular a particle filter (Arulampalam et al. 2002, Chen 2003), to model saccadic eye movements. It defines eye movement as the evolution of a discrete state \( \{e_k, k \in \mathbb{N}\} \) given by

\[
e_k = f_k(e_{k-1}),
\]

where \( f_k \) is a stochastic function that maps the eye state \( e_{k-1} \) to \( e_k \) and \( \mathbb{N} \) is the set of natural numbers. An eye state, denoted as \( e_k = [u_k, b_k, v_k, c_k] \), is determined in terms of eye coordinates in image plane and size of focus (i.e., \( u_k \)), a vector of binary variables \( b_k \) such that \( b_k(u) = 1 \) if \( u \) is not visited, a Bernoulli variable \( v_k \) that represents fixation status, and a binary vector \( c_k \) such that \( c_k(u) = 1 \) if \( u \) is not visited as a fixation.

Incorporating the concept of inhibition of return in visual working memory, the dynamics of the model accumulates visited locations and inhibits immediate return to attended locations. To implement this phenomena, the proposed model applies a stochastic function \( f_u \), which is defined as \( u_k = f_u(u_{k-1}, b_{k-1}) \). It is determined by

\[
p(u_k|u_{k-1}, b_{k-1}) = \eta p(d(u_k, u_{k-1}))(b_{k-1} - \Gamma(u_{k-1}))u_k,
\]

where \( d \) is a distance function, \( p(\cdot) \) is a pdf determining probability of a jump with length \( d \), \( \Gamma(u) = [\gamma_1(u) \ldots \gamma_n(u)]^T \) such that \( \gamma_i(u) = \delta(u - i) \) and \( \eta \) is a normalizing constant.

The state of the eyes (i.e., fixation/saccade occurrence) is defined by a Bernoulli process \( v_k \) which is determined by
\[
p(v_k) = \begin{cases} \frac{P_0}{1 - P_0} v_k = 1 \\ \frac{1 - P_0}{1 - P_0} \end{cases}, \tag{10}
\]

where \( P_0 \) is the probability of fixation occurrence. Eventually, the proposed model utilizes the aforementioned inhibitory like mechanism and the state of eyes to define the dynamics of the system, which is expressed in terms of

\[
p(e_k | e_{k-1}) = p(u_k | u_{k-1}, b_{k-1})p(b_k | u_{k-1}, b_{k-1})p(c_k | c_{k-1}, v_{k-1}, u_{k-1})p(v_k), \tag{11}
\]

where

\[
p(c_k | c_{k-1}, v_{k-1}, u_{k-1}) = \delta(\sum_v c_k - c_{k-1} + \Gamma(u_{k-1}) \delta(v_{k-1}))
\quad \text{and} \quad
p(b_k | u_{k-1}, b_{k-1}) = \delta(\sum_b b_k - b_{k-1} + \Gamma(u_{k-1}))
\quad \text{while} \quad
b_k \text{ and } c_k \text{ are updated by } b_k = b_{k-1} - \Gamma(u_{k-1}) \text{ and } c_k = c_{k-1} - \delta(v_{k-1}) \Gamma(u_{k-1}).
\]

The proposed model utilizes a stochastic measurement function \( h_k \), defined as

\[
z_k = \begin{cases} 1 & Sal_{u_k} \geq \tau \\ 0 & \text{otherwise} \end{cases},
\]

where \( \tau \) is a threshold value and \( Sal_{u_k} \) is the salience value at \( u_k \). In other words, it relies on the salience value of current eye position and a specific size of focus. Eventually, the observation model follows:

\[
p(z_k | u_k, v_k, c_k) = \begin{cases} P_1 & v_k = 1, z_k = 1 \\ 1 - P_1 & v_k = 1, z_k = 0 \\ P_1 e^{-\min(\phi(u_k, c_k)) / 2 \sigma^2} & v_k = 0, z_k = 1 \\ (1 - P_1) e^{-\min(\phi(u_k, c_k)) / 2 \sigma^2} & v_k = 0, z_k = 0 \end{cases}, \tag{13}
\]

where \( \phi(.) \) is a function that computes the distance of the current eye position to fixation locations, \( \sigma \) is the standard deviation of the smoothing kernel and \( P_1 \) is the probability value of having a fixation on an item.

The proposed model assesses salience using the Bayesian center-surround salience model explained in Chapter 2 (published in Paper I). Nonetheless, it is slightly modified to evaluate salience on a given image coordinate at a single scale in which the scale represents the size of focus. At the end of the simulation period, the proposed model applies an algorithm to generate a scanpath, as depicted in Fig 5, from the visited locations of a particle with maximum observation value. It also produces a fixation map and a salience map from fixation locations.
Fig 5. Scanpath simulation, the result of the proposed model is on the left and four different human scanpaths are on the right. The beginning of the path is identified by a green mark and the end is marked blue. The proposed model is not supposed to replicate exactly an observer, though it shall behave similar to them. In other words, it should fixate almost on the same locations that observers commonly fixate. (Reprinted from Paper II, Copyright (2013), with permission from Elsevier).

3.4 Discussion

This chapter is devoted to eye movements because of its vital role in salience modeling. In a relatively brief introduction, it tried to identify the origins of the use of eye movements in salience modeling within the scope of computer vision. Moreover, it provides a simple engineering insight into the properties of eye movements. It investigated the techniques for scanpath generation by categorizing them into visual attention and oculomotor. The first category signifies the links between the attention modeling and saliency modeling, while the oculomotor concept highlights the sampling premise. The chapter ended with the details of a saccade generation scheme which was introduced in Paper II.

The stochastic bottom-up saccade generation model simulates saccadic eye movements by using a particle filter. It applies a system model that prevents immediate return to a visited location and proposes candidate locations at a distance from current location. A candidate location is tagged as fixation if its salience value exceeds a threshold. A complication of the method is its reliance on several parameters such as the appropriate length of a saccade, probability of fixation occurrence, etc. To avoid the argument on the optimal selection of these parameters, the SBSG model learns them from eye movements data in a free-viewing task.

The most indispensable parameter is $p(d(u_k, u_{k-1}))$ which, in practice, determines the length of a random jump. The model learns it by fitting a mixture of Gaussian to the
saccade length information extracted from observers’ eye movements data. An analysis shows that the mixture of Gaussian fits better to the underlying saccade data than the other distributions such as Cauchy or Levy (see, Paper II). The parameter $P_0$ determines the chance of a fixation occurrence. It is set equal to the proportion of the number of fixations to the total number of saccades and fixations. Similarly, to learn $P_1$, which assess the probability of fixating on an item, the model computes the ratio between the average number of fixations that fall within the 99% most salient area of fixation density map and the total number of occurred fixations. It also sets $\tau$ to the average of the salience value of the fixation points belonging to the 99% most salient area of fixation density map.

In the proposed model, Bayesian filtering provides the solution. Although Bayesian optimization explains human active search strategies (Borji & Itti 2013), the optimality of the proposed solution, which relies on a suboptimal technique (i.e., particle filters) (Arulampalam et al. 2002), requires further investigation. Indeed, it demands several controlled experiments to validate the visual search behavior of the proposed model against human. That goes beyond the scope of the current study (i.e., salience modeling). Moreover, it is still an open challenge in active search area and visual attention modeling.

Finally, the proposed model somehow resembles a bottom-up overt attention model. Nevertheless, it seems easily extendable to top-down or a full attention model (i.e., having both top-down and bottom-up components). In theory, the proposed model is capable of handling top-down information since it uses salience measure of Paper I. In order to implement such an extension, one should also modify fixation selection and visual memory to reflect the influence of top-down cue. This can be a topic of further research.
4 Video saliency

This chapter investigates the research on salience in presence of time-varying stimulus or videos. The arguments on time-varying stimulus are almost identical to the statements about still images. Nonetheless, it is often expected that a model that handles such a stimulus considers temporal information (e.g., motion) because of their influential role (Wolfe 2004). The same model categorization can apply to them, and many models, which were studied earlier, can be easily extended to handle videos by incorporating some temporal information, e.g., Itti & Dhanale (2003) extended the model of Itti et al. (1998) to include motion and flicker cues. In essence, temporal information integration somehow motivates fusion of temporal and spatial information. Having spatio-temporal purposes in mind, some models consider temporal pathway and static pathway, e.g., Marat et al. (2009) considered two pathways of cortical like filters to process dominant motion and spatial information separately. In this approach, each pathway can have a unique treatment for the stimuli. Similar architecture is adopted by Zhai & Shah (2006), Mancas et al. (2011) and Paper VII.

On the other hand, the formulation of some models is such that one can apply them to the time-varying stimulus, as well as a still image. A good exemplar is the model of Itti & Baldi (2005a,b, 2009) which exploits the deviation in the posterior and prior beliefs to identify salience. Although this methodology is applicable to both spatial and temporal stimulus, its fame is because of its success on videos. The model of Bruce & Tsotsos (2006, 2009) responds to time-varying stimulus by providing it some spatio-temporal basis functions. Similarly, Zhang et al. (2009) explained how to utilize the formulation of Zhang et al. (2008) in order to process videos. Seo & Milanfar (2009b) demonstrates the use of a spatio-temporal volume of features in the self-resemblance salience framework of Seo & Milanfar (2009a). Later, this study presents how to generalize the Bayesian center-surround framework to video by using a spherical center-surround representation, which is introduced in Paper V.

The aforementioned models, almost all, rely on the spatio-temporal information. Contrarily, one can particularly target the temporal cue, e.g., Wixson (2000) proposed an algorithm to detect salient motion by intermediate-stage vision integration of the optical flow; it demonstrates the power of motion in detecting salience in the presence of time-varying stimulus. Relying, only, on temporal information, Paper VI develops a
salience modeling scheme for fast motion detection. The rest of this chapter introduces three models published in Papers V, VI & VII, respectively.

### 4.1 Spherical center-surround

This section describes the model of Paper V. The proposed model shares a common ground with the Bayesian center-surround (see, Chapter 2 and Paper I). In other words, it follows the center-surround hypothesis while utilizing the same Bayesian formulation. Nevertheless, it uses a spatio-temporal volume of information. This section exclusively discusses the information representation.

In the case of a still image, the easiest representation of a center-surround patch is a rectangular image patch. To process videos, most of the models (e.g., Seo & Milanfar 2009b, Mahadevan & Vasconcelos 2010) extend such an image patch to a spatio-temporal rectangular volume, which is depicted in Fig 6, because of easier implementation and slightly efficiency advantage. Similarly, one may suggest a cylindrical volume for circular patches of Paper I arguing better rotational symmetry advantages (Dakin & Herbert 1998). Although cylindrical representation is enough to formulate the Bayesian center-surround, there is a better choice in order to implement the model under the two assumptions of 1) singleton center, and 2) sparse surround. Paper V adopts a spherical representation.

Fig. 7 illustrates the proposed spherical representation. Taking only a small number of pixels uniformly scattered at a distance from a central pixel, this representation facilitates the implicit incorporation of the sense of flicker and motion. The implementation of the proposed model ensures an estimate of the sense of flicker by implicit comparison of a
pixel to its history. For a video volume $V$, the model estimates flicker by comparing $V(x, t) \in C$ with $V(x, t + \Delta t) \in S$ while performing kernel density estimation. Similarly, it incorporates a rough sense of motion by collating the center, which is denoted as $V(x, t)$ and consists of one pixel, with $V(x + \Delta x, t + \Delta t) \in S$. The spatial information is preserved by considering samples $V(x + \Delta x, t) \in S$.

In real time applications, the spherical representation can be easily replaced by a semi-spherical representation. Paper V finds no significant overall performance difference between the sphere and semi-sphere representation for a small video buffer in which small is 39 frames for the semi-spherical representation. Fig 8 visualizes several saliency maps obtained by the proposed model.

Fig 8. Spherical center-surround and visualization of saliency maps. The first row depicts video frames and the second row portrays their corresponding saliency map.

4.2 Temporal motion model

Paper VI introduced the temporal motion model for the purpose of fast motion detection. It is mostly grounded on the significance of motion cue as an undoubted attention
Fig 9. Temporal motion model, a step by step description diagram. It accepts a small video buffer and produces the salience of the last frame using its history and flicker concept.

guiding attribute (Wolfe 2004). It utilizes flicker which is defined as the divergence of the value of a pixel over time.

Flicker detects temporal change, e.g., Itti & Dhavale (2003) computed flicker as the absolute difference of two consequent frames in a sequence. The proposed model adopts a different approach. It considers flicker as the deviation of the value of a pixel from its mean value in a small temporal window. Fig 9 provides an insight on how the method works.

It treats a video volume $V$ as a sequence of images $V = \{v_t, v_{t+1}, ..., v_{t+n}\}$, in which $v_i$ is the column representation of the luminance of video frame at time $i$ in LMS color space. It subtracts the mean of each row of the image sequence, denoted $\mu V$, from each frame (i.e., $\tilde{V} = V - \mu V$). The premise is that $\mu V$ identifies the average static information of a sequence (i.e., background). Consequently, $\tilde{V}$ provides an approximation of flicker (i.e., change over time) in the scene. Afterwards, the proposed model applies the procedure of Principal Component Analysis (PCA) with whitening to obtain a dimension reduced back projected sequence, denoted $\tilde{V}_w^p$. Eventually, it defines salience as

$$ Sal = \mathcal{N} \left( \sum_i |\tilde{V}_w^p, i| \right), \quad (14) $$

where $\mathcal{N}(\cdot)$ is a normalizing function, and $\tilde{V}_w^p, i$ is the $i_{th}$ frame of $\tilde{V}_w^p$. Fig. 10 depicts the salience measure obtained by the temporal motion model.

Fig 10. Salience using temporal motion model, original frame and its corresponding salience measure on its right side.
4.3 Spatio-temporal motion model

Paper VII introduced the spatio-temporal motion model. This model considers two pathways, namely temporal and spatial, for processing an image sequence. However, it utilizes them slightly differently compared to similar architectures, which are introduced in (Zhai & Shah 2006, Marat et al. 2009). As depicted in Fig 11, it cascades the two pathways.

In order to judge salience in the spatial domain, it relies on Non-negative Matrix Factorization (NMF), which substitutes ICA to approximate neural responses. It motivates use of NMF because of its sparse representation (Hoyer 2003), data encoding ability with few active elements (Hoyer 2004), and part-based representation (Rajapakse & Wyse 2003). Particularly, it expects that an intermediate representation performs better than a holistic one because it encodes both local and global information. Comparing NMF with ICA and PCA, it underlines the superiority of NMF.

In order to learn the NMF basis, it samples one million patches of size $24 \times 24$ from McGill color image dataset (Olmos & Kingdom 2004). It converts each patch to gray-scale and stacks them in a column order to make a $n \times m$ feature matrix, $F = [f_1, f_2, \ldots, f_m]$ with the approximative factorization of

$$F \approx (WH)_{i,j} = \sum_{k=1}^{r} w_{ik} h_{kj}, \quad (15)$$
where each column of $W$ represents the so-called basis vector, and each column of $H$ consists of the encoding coefficients, which determine the strength of each basis vector. The basis vectors are obtained by solving

$$\minimize_{W, H} f(W, H) = \frac{1}{2} \| F - WH \|_F^2,$$

subject to $W, H > 0$ (16)

where $\| . \|_F$ is the Frobenius norm.

Eventually, to derive the spatial conspicuousness of the luminance channel of an image/video frame $I$, it normalizes the image as $\tilde{I} = (I - \mu_I)/\sigma_I$ and computes $C_i(I) = |S_i \ast \tilde{I}|$, in which $S_i = w_i, i = 1, \ldots, r$, $w_i$ is the $i$th row of $W^\top$, and $\ast$ is the convolution operator. For each video frame, the proposed model computes $r$ spatial conspicuousness maps. Afterwards, it makes $r$ video buffers, so each buffer, denoted $V_i$ $i = 1, \ldots, r$, contains the spatial conspicuousness obtained from a particular $S_i$. These video buffers and the original video are fed to the temporal pathway described in section 4.2. Consequently, the model applies (14) to obtain $Sal_i, i = 1 \ldots r + 1$. The final salience is defined as

$$Sal = \mathcal{N} \left( \sum_{i} b_i Sal_i \right),$$

where $\sum_{i=1}^{r+1} b_i = 1$, and $\mathcal{N}(.)$ is a normalizing function. The salience measure is further post-processed to produce a smooth saliency map. Fig 12 shows some examples of such saliency maps.

Fig 12. Spatio-temporal motion model, original frame and its corresponding smooth saliency map on its right side.
4.4 Discussion

This chapter was mostly a brief overview of the video saliency contributions of this study. Essentially, the video salience models are a generalization of the models applied to an image. While one may motivate a frame by frame use of a model developed for image saliency, the temporal information shall be taken into account in dealing with video. Consequently, it categorized the models based on the basis of handling such information.

This chapter introduced three models. The first model, spherical center-surround, shows how changing the data representation can transform an image saliency model into video saliency model. In essence, this model somehow seems to consider information of the past and future. Nonetheless, a semi-sphere is practically implemented for real-time processing of the video frames.

The temporal motion model only relies on the temporal cue. It is particularly developed for fast motion detection. This chapter introduced the salience measure only. However, the salience measure is further processed to infer the foreground in a background models challenge Vacavant et al. (2013). The results on the proposed challenge signifies the role of temporal cue and flicker.

The last model, spatio-temporal model, develops two pathways for processing the information. While some models often utilize two parallel pathways, it treats them sequentially. In fact, the stimuli initially pass via the spatial tract before entering the temporal path. This model somehow extends the temporal motion model. It targets the area of background subtraction where there is inadequate initial background information.
5 Applications

This chapter is devoted to the application of salience modeling in the domain of computer vision. While there are several surveys which studied different aspects of eye movements, saliency and visual attention (e.g., Frintrop et al. (2010), Toet (2011), Borji & Itti (2012b), Judd et al. (2012), Borji et al. (2013a), Filipe & Alexandre (2013)), they often pay less attention to the applications of salience modeling. This chapter initially tries to portray the span of applications by exploring several examples. Afterwards, it describes how the present study utilized salience and eye movements in some of these applications.

5.1 Salience modeling in action

The low cost localization of the region of interest often helps to the efficiency of a wide range of computer vision related applications. Such a preprocessing step is sometimes crucial depending on the amount of information and application constraints. Although the agreement on a general solution seems difficult, the study of human attention mechanism helps proposing a less application specific solution. One can exploit the premise that a human often focuses on the regions with eminent information and utilizes salience modeling for such a purpose. In this respect, the concept of saliency modeling use is redundant information curtailment, efficient resource allocation, wise problem formulation, and/or less problem specific expressions.

The integration of salience modeling techniques into computer vision applications is either as a simple preprocessing step or as a more sophisticated interactive asset in the resolution. As an example in the latter category, Borji et al. (2011a) proposed a saliency-based method for quick localization of object candidates in which they bias the saliency by putting costs on the computation of the features of the model for object detection. Indeed, the saliency map is part of the final solution in their framework.

5.1.1 Recognition

In computer vision, recognition is the task of identifying a specific target, e.g., object, scene, face, in image or video. It shares similar concepts with cognitive studies such as neuroscience and artificial intelligence in formalizing target perception. Thus, it
provides an excellent grounds for the application of visual attention models and salience modeling.

In order to recognize objects, one can apply salience in a two stage approach of attentional preprocessing and classifying recognizer. For instance, “Visual Object detection with a CompUtational attention System” (VOCUS) (Frintrop 2006) adopts the concept of salience to detect regions of interest followed by anomaly detection or specific target recognition. It further investigates sensor fusion schemes in order to improve saliency and object recognition in robot vision.

A pattern recognition model proposed by Salah et al. (2002) exploits the concepts of the primate selective attention mechanism for efficient recognition tasks, e.g., handwritten digit recognition and face recognition. It uses a saliency scheme to simulate primitive attentive levels of human visual system which is followed by Markov models to control the processed regions of the image. To perform recognition, it considers an intermediate level of neural networks which processes the observed parts of the image and generates posterior probabilities as observations to the Markov model. Eventually, the model recognizes objects of interest aptly.

Siagian & Itti (2007) utilized the salience models and early visual features to differentiate outdoor scenes from other sites. In essence, it utilizes salience to capture the “gist” of the scene into a low-dimensional feature vector. It merges the gist of image and a local object recognition technique, which is somehow based on salience modeling, in order to recognize the observed scene.

Gao et al. (2009) formulated a visual saliency model that exploits top-down information. The proposed model efficiently recognizes objects embedded in significantly cluttered scenes. Similarly, Kanan & Cottrell (2010) utilized a salience model in a recognition framework. Nonetheless, these models can be studied under the umbrella of visual attention models because, in essence, they utilize top-down modulated FOA shift mechanisms either implicitly or explicitly. In this context, an example is (Walther 2006) which studies the application of visual attention models in object recognition, detection and tracking.

5.1.2 Object detection

Object detection is the process of finding instances of real-world objects in images or videos. Traditionally, an object detection model uses a sliding window to assess the presence or absence of an object (e.g., Papageorgiou & Poggio 2000). Since the sliding
window architecture is computationally expensive, the use of attention mechanism and salience-based models is motivated. In the context of object detection, visual search and attention models are broader concepts that are sometimes difficult to distinguish from salience concept. Some of the methods discussed here are more likely to be attention models that utilize salience.

The concept of salience can be used in making a bottom-up voting scheme which one can use to obtain candidate object locations. For instance, Bouchard & Triggs (2005) applied a similar idea to suppress the redundant local features in order to introduce a hierarchical part-based object detector. In the context of priors, Oliva & Torralba (2001), Torralba (2003) fused salience and contextual priors to obtain the likelihood of object existence at a location. Similarly, Ehinger et al. (2009) applied salience, scene context, and target features to search people in images. They demonstrated that the fusion of several information cues outperform each individual cue.

Fritz et al. (2005) proposed an attentive object detection scheme in which the premise is that the focus of attention on discriminating patterns enables efficient detection. They interpreted the attention as the selective response of tuned early features to generic task related visual features. In principle, they obtained a locally relevant to object information appearance and mapped it to a discriminating representation using decision trees. They demonstrated the effectiveness of attentive object detector in rapid object detection.

Butko & Movellan (2009) relied on a visual search model which schedules eye fixations to maximize the long-term information accrued about the location of the object. They elaborated the eye movement model of Najemnik & Geisler (2005) and incorporated a digital fovea, which produces image patches of different size centered on a given fixation point such that the resolution falls off in the periphery, for faster object detection.

Navalpakkam & Itti (2006) investigated the role of top-down and bottom-up information in an attention model applied to quick object detection. The model relies on the premise that the detection speed depends on the ratio between the strength of the signal detecting the target and the signal detecting non-target area. It defines SNR as the ratio of expected salience of target over distractors to choose optimal top-down weights that maximize the target’s salience. Eventually, it is shown that the combination of saliency and such top-down information speeds up the object detection.
5.1.3 Compression

A challenge in image/video compression is the prevention of the corruption of visually important elements meanwhile redundant information curtailment. To preserve vital information, one may suggest careful handling of the regions of interests (Kunt et al. 1985). In this context, visual saliency replicates the region of interest detector, e.g., Ouerhani et al. (2001) applied saliency maps as region of interests detectors to formulate an adaptive JPEG-based image compression algorithm in which the number of encoding bits of a pixel is determined according to their salience.

Dhavale & Itti (2003) adopted the same concept to perform foveated compression. The usefulness of this concept was later validated by assessing algorithmic foveation using recorded eye movements (Itti 2004). Likewise, Guo & Zhang (2010) introduced a hierarchical selectivity framework based on saliency to construct an image representation utilizing trees. The hierarchical selectivity leads to the development of a multiresolution wavelet domain foveation which improves the coding efficiency in image and video compression.

5.1.4 Video summarization

Accessibility and management of large video libraries requires identification and extraction of essential video content. Nonetheless, manual extraction of such information is a laborious task. This motivates an automated framework of information prioritization and filtering. In such a framework, saliency provides a feasible solution without fully semantic understanding of video content. A good exemplar is (Ma et al. 2005) which extracts key-frames and video skims using an attention curve obtained from saliency.

5.1.5 Object tracking

Object tracking is the process of associating a target object in consecutive video frames. It is a vital component of applications such as surveillance, human-computer interaction, traffic monitoring, and activity recognition. A crucial part of a tracking framework is object representation which is doable using appearance models (Yilmaz et al. 2006). In the domain of appearance based tracking, one can utilize saliency to discriminate between background and the object of interest meanwhile learning the object appearance.
Mahadevan & Vasconcelos (2009) applied the discriminant saliency model of Gao & Vasconcelos (2009) to discriminate between target (i.e., salient stimuli) and background (i.e., non-salient stimuli). Their proposed model uses salient features to model the target. It learns the target representation at a time and apply it as a matched filter later.

Frintrop (2010) proposed a component-based descriptor to represent the target. It is integrated into a CONDENSATION-based object tracker (Isard & Blake 1998). To compute the target descriptor, it initially computes six salience like feature maps, which represent different receptor responses in the human visual system, using the model of Frintrop (2006). In each feature map, it finds local intensity maxima regions which are segmented to form a component in the target descriptor. Borji et al. (2012a) extended this framework to adapt itself to background changes.

5.1.6 **Thumbniling and retargeting**

A straightforward image thumbnailing is to shrink the original image. Nonetheless, such a solution is not as effective as thumbnail cropping and retargeting in which the image is manipulated while its visual eminent content is preserved. In fact, a challenge of thumbnailing and retargeting is the recognition of the region of interest. Since salience modeling provides a viable solution to detection of region of interest in the input image, one can adopt it in the mentioned tasks. For instance, Marchesotti et al. (2009) introduced a framework to generate intelligent thumbnails. It derives a saliency map for a given image and refines it to extract the thumbnails. Other analogous examples are (Le Meur et al. 2006, Goferman et al. 2010, Lu et al. 2010, Kim et al. 2011, Erdem & Erdem 2013).

5.1.7 **Segmentation**

Segmentation is the process of partitioning an image into segments. To obtain a fully automatic framework, Mishra et al. (2009) suggested the use of fixation points as initial seeds for finding bounding contours of objects. The premise is that every fixation lies inside a particular region of arbitrary shape and size in the scene. While their algorithm has no mechanism for fixation generation, they motivated use of saliency-based attention models as a replicate.

Saliency-based segmentation often focuses on segmenting the most salient object. For instance, Fu et al. (2008) introduced “Salience cuts”. It combines the salience model
of Hou & Zhang (2007) and graph cuts based object segmentation (Boykov et al. 2001) to automatically segment objects. In this framework, salience map provides the labeled seeds for the graph cuts.

Yanulevskaya et al. (2013) adopted proto-objects (Walther & Koch 2006) concept to segment the most salient object. In essence, their proposed model initially generates several proto-object candidates using the method of Van de Sande et al. (2011). For each candidate, it computes center-surround saliency and integrated saliency (i.e., locally rare features) measures to determine the most salient proto-object which is eventually segmented.

The segmentation task associates the two areas of salience modeling and salience segmentation. Hence, the salience modeling algorithms can apply to segmentation task; however, they often require some post-processing step and tuning. On the other hand, it is less successful to assess salience segmentation algorithms against observers’ eye movement data. In fact, direct assessment of one class of techniques against the other class is not recommended (Borji et al. 2013a).

In the context of segmentation, salience segmentation techniques are more successful than salience modeling in dealing with computer vision segmentation data bases. Some examples of such techniques can be found in (Achanta et al. 2008, 2009, Achanta & Susstrunk 2010, Cheng et al. 2011, Lu et al. 2011).

5.1.8 Medical purposes

Medical purposes can include a variety of subject matters such as biomedical engineering and medical imaging. Biomedical engineering tries to close the gap between engineering and medicine by combining the design and problem solving principles of engineering with medical and biological sciences to advance the healthcare treatment. It covers development of devices that help treatment of diseases. To develop low vision assistive devices, one may utilize saliency models and visual attention models. For instance, Parikh et al. (2010) exploited a saliency model in development of a electronic retinal prosthesis to treat blinding diseases like retinitis pigmentosa and age-related macular degeneration.

Medical imaging carries the flagship of applied imaging techniques facilitating clinical analysis and medical intervention. It seeks to reveal internal structures of body as well as diagnose and treat diseases. To diagnose, it sometimes necessitates recognition of abnormal regions of tissue. Interpreting abnormality as saliency, one can
effectively employ saliency modeling to delineate unnatural regions, e.g., tumors in mammogram (Hong & Brady 2003).

5.1.9 Graphics and art

The recent enhancements in computerized graphics and art have facilitated the design of high fidelity virtual environments and artistic images. Achieving such an accuracy requires intelligent treatment of the scene, e.g., in order to generate high quality graphics efficiently, one can adopt saliency to render graphics in an adaptive manner. In such a process, saliency detects visually attractive regions of meshes which are rendered with higher accuracy (Lee et al. 2005, Bulbul et al. 2010, Kim et al. 2010, Song et al. 2014). In the domain of computer graphics, this technique is usually renowned as “Mesh Saliency”. Other examples are non-photorealistic photo rendering, picture collage, and dynamic lightening. DeCarlo & Santella (2002) applied a model of human perception to identify highlights of visual elements to generate automatic non-photorealistic photo renders. In a picture collage application, which represents a visual image summary by arranging all input images on one canvas, Wang et al. (2006) applied saliency to maximize visible visual information that fits into the canvas. El-Nasr et al. (2009) introduced an adaptive lighting algorithm. It exploits saliency to dynamically adjust the lighting color and brightness of interesting objects in interactive 3-D scenes.

5.1.10 Robotics

Robotics probably adopted the attention concept since the very early days. In a complex real world, a robot has to decide and act under limited resources. To decide the actions, it has to filter the irrelevant information and concentrate on the priorities. Consequently, it often focuses on one task at a time. Indeed, it definitely benefits from visual attention and saliency in a broad range of applications such as localization (e.g., Frintrop et al. 2007), navigation (e.g., Chang et al. 2010, Roberts et al. 2012), human-robot interaction (e.g., Breazeal & Scassellati 1999, Muhl et al. 2007, Nagai 2009, Schillaci et al. 2013) and active vision (i.e., where to look next?) scenario (e.g., Breazeal et al. 2001, Vijayakumar et al. 2001, Dankers et al. 2007, Rasolzadeh et al. 2010, Borji et al. 2010a).
5.1.11 Others

Some of the other applications that utilized saliency are super-resolution (e.g., Sadaka & Karam 2009), advertisement (e.g., Liu et al. 2008), perceptual designing (e.g., Rosenholtz et al. 2011), image quality assessment (e.g., Ninassi et al. 2007, Ma & Zhang 2008), scene memorability (e.g., Mancas & Meur 2013), security and x-ray imaging (e.g., Schmidt-Hackenberg et al. 2012), gaze estimation (e.g., Valenti et al. 2012, Sugano et al. 2013), and motion detection and background subtraction (e.g., Mahadevan & Vasconcelos 2010). Having a broad horizon, saliency detection can be interpreted as a kind of interest point/region detector that somehow covers most of the examples. However, there are cases that a saliency model is deliberately applied to interest point/region detection (e.g., Kadir & Brady 2001). In conclusion, the application of saliency and visual attention is not limited to what was mentioned above and there still exist numerous applications to discuss.

5.2 Exploiting proposed models

This section briefly discusses how this study employed the proposed models in some applications. It discusses two applications of object tracking, and motion detection & background subtraction. The models and applications are originally presented in Papers VIII, VI, & VII.

5.2.1 LSN model and tracking

The proposed tracking method utilizes the Local Similarity Number (LSN) model to represent the target in a mean-shift tracking framework. The mean-shift framework achieves tracking by calculating the likelihood of a target model and candidate model. Computing the distance between two models, the tracking problem reduces to an optimization problem. This study adopts the iterative formulation of Comaniciu & Meer (2002) in which for a normalized target patch denoted as \( \{ t^*_i \}_{i=1..n} \)

\[
\hat{y} = \frac{\sum w_i t_i}{\sum w_i}, \tag{18}
\]

where \( w_i = \sum \sqrt{\frac{b}{P_n(y)}} \delta([b(t^*_i) - u_i]) \), \( \hat{y} \) is the new position of a target model formerly positioned at \( y \), \( t_i \) belongs to the candidate patch denoted as \( \{ t_i \}_{i=1..n} \). \( \delta \) is the Kronecker
The above framework requires a target model and a candidate model. In general, a target model is formulated as \( \hat{q}_u = \left\{ \hat{q}_u \right\}_{u=1...m} \), where \( \hat{q}_u = c \sum_{n=1}^{m} K\left( \|t_i\|^2 \right) \delta\left( b(t_i) - u \right) \), \( c = 1/\sum_{n=1}^{m} K\left( \|t_i\|^2 \right) \). \( K\left( . \right) \) is an isotropic kernel, and \( m \) is the feature vector dimension. Similarly, it defines the candidate model as \( \hat{p}(y) = \left\{ \hat{p}_u(y) \right\}_{u=1...m} \), where \( \hat{p}_u(y) = c \sum_{n=1}^{m} K\left( \|y-t_i\|^2 \right) \delta\left( b(t_i) - u \right) \) and \( c = 1/\sum_{n=1}^{m} K\left( \|y-t_i\|^2 \right) \).

In practice, mean-shift tracks any object whose appearance model is defined by a histogram. In order to represent the tracking target, Paper VIII adopts a joint saliency-color histogram. The histogram consists of 8-bin quantized color of RGB channels and a 5-bin saliency descriptor (i.e., a histogram of \( 8 \times 8 \times 8 \times 5 \)). It obtains the 5-bin saliency as \( mLSN(x,d) = \left\{ 1 + LSN_1(x,d) \mid LSN_1(x,d) \in \{0,1,2,3,4\} \right\} \), where \( LSN_1(x,d) \) is the local similarity number operator.

5.2.2 Motion detection & background subtraction

This section discusses the application of temporal motion model (Paper VI) and spatio-temporal motion model (Paper VII) to motion detection and background subtraction problem.

In general, motion detection is the process of detecting a change in position of an object; and background subtraction is the localization of objects of interest, e.g., detecting all moving objects in a traffic monitoring application. To achieve such objectives, traditional techniques build a background model and use it to discriminate objects against background (Bouwmans 2011). Nonetheless, the two objectives are intrinsically related by equating background subtraction with the detection of salient motion (Mahadevan & Vasconcelos 2010). Similar to the aforementioned applications, saliency modeling provides an effective solution to them. Major benefit of applying saliency modeling to such applications is that there is no need to train/maintain a background model.

To differ object from background, Papers VI & VII initially post-process the salience measure, denoted \( Sal \), to obtain a smoothed saliency map

\[
s_{map} = \mathcal{N}\left( (\mathcal{G}_{\sigma} * (Sal \oplus disk_3)^{\alpha}) \right),
\]

where \( \oplus \) is a dilation operator, \( disk_3 \) is a disk structure of element size 3, \( \mathcal{G}_{\sigma} \) is a Gaussian kernel of standard deviation \( \sigma \), \( * \) is the convolution operator, \( \alpha \) is an attenuation control.
parameter, and \( N(.) \) is a normalizing function. Afterwards, a simple linear classifier produces an object mask, denoted \( o_{\text{mask}} \), such that

\[
o_{\text{mask}} = \begin{cases} 
1 & s_{\text{map}} \geq \tau \\
0 & \text{otherwise}
\end{cases},
\]

(20)

where \( \tau \) is a threshold value.

5.3 Eye movement use

The role of eye movement in human vision has been a contentious subject matter. Although the presence of eye movement in visual perception is proven necessary (Land et al. 2002, Martinez-Conde et al. 2004), its true role and connection to the cognitive state of an observer still requires further research. For instance, the seminal work of Yarbus (1967) on the influence of task on eye movement patterns is still a topic of investigation with discouraging studies (Greene et al. 2012) and supporting findings (Henderson et al. 2013, Borji & Itti 2014).

Despite all existing arguments, the advent of eye-tracking technology is promoting the use of eye movement in vision related applications. For instance, egocentric vision, which offers a unique human-centric perspective of visual world, provides a new insight into eye movement based behavior analysis. It helps demonstrating that attention patterns can be used for recognizing social interactions (Fathi et al. 2012).

To date, eye movement modality has been used in several applications. Human-computer interaction (HCI) treats eye movement as an input medium in which the emphasis is on natural incorporation of eye movements into the user-computer dialogue (Jacob 1993). Decoding user activity benefits from eye movement patterns in which one infers individuals assigned task, e.g., text copying, reading, taking notes, watching video, and browsing the web, from his/her eye movements (Bulling et al. 2011). Similarly, eye movements help scene understanding task (Mathe & Sminchisescu 2012). In order to build assistive technology for the purpose of autism diagnosis and monitoring of social development of children, Ye et al. (2012) applied eye movement to detect eye contact. A recent study also demonstrated that eye movement is a means of depression detection (Alghowinem et al. 2013).

This study investigates the role of eye movements in conjunction with saliency modeling in several tasks. In the rest of this chapter, the focus is on the part of the application specific contributions of Papers II, III & IV. Paper II demonstrates the
5.3.1 Salience modeling

The SBSG model, which is originally introduced in Paper II, tries to incorporate eye movements in salience modeling in order to improve its performance. It grounds itself on the findings of Tatler & Vincent (2009) which demonstrates predictive power of oculomotor bias in attending informative locations and produces saliency maps from a set of generated fixations. It adopts a method which is similar to the production of fixation density maps from eye tracking data. Such a process consists of filtering out the saccadic eye movements, omission of fixation points without enough nearest neighbors, and smoothing. Similarly, SBSG builds a fixation map and smooth it to produce the saliency map from a set of artificially generated fixation points. Fig. 13 depicts an image and its corresponding fixation points and saliency map.

5.3.2 Scene recognition

Doing an intensive salience modeling evaluation, Paper III investigates the performance of salience models augmented with features from statistics of eye movements in the task of recognition of stimulus category (i.e., type of the scene). It deliberately avoids image features to provide an easy test bed.

It decodes the stimulus category of a subset of images from NUSEF dataset (Ramanathan et al. 2010) which contains five categories over a total of 409 images. The
categories include event, face, nude, portrait and others. For each image, it extracts a 872 dimensional feature vector. The feature vector includes fixation histogram (256D), fixation duration histogram (60 bins), saccade length (50 bins), saccade orientation (36 bins), saccade duration (60 bins), saccade velocity (50 bins), saccade slope (30 bins), saliency map (300D), saliency histogram at fixated locations (10 bins), and top ten salient locations. To summarize, the features fall into three category of fixation, saccade, and saliency.

To compute the saccade histograms for a given image, it initially computes corresponding features (e.g., saccade velocity, duration, orientation, etc.) for each observer and quantizes the values into several bins. Later, the histogram of saccade statistic for an image is computed from all the observers and is L1 normalized. While the same methodology applies to the calculation of the fixation duration histogram, the fixation histogram is made by dividing the image into a grid pattern (16 × 16) and counting the number of fixations in each grid. It obtains the coordinates of the ten most salient points by applying IOR to the saliency map (a vector of size 1 × 20) and vectorizes a saliency map of size 20 × 15 to form saliency map feature. The saliency histogram at fixated locations represents the distribution of saliency values at fixations.

It performs the classification using a multi-class SVM classifier by utilizing a regularized binary support vector classifier (i.e. C-SVC) with RBF kernel. It tunes the parameters of the SVM using twenty images of each category in a 5-fold cross validation scheme over 100 runs. Afterwards, it categorizes the images in a 5-fold cross validation process, which is repeated ten times.

5.3.3 Valence recognition

Paper IV investigates the application of eye movements in the task of valence recognition. The emotional valence of a scene refers to the extent to which an individual is attracted or repelled by the visual content. In this context, a scene is in one of the categories of pleasant, neutral, or unpleasant.

Traditionally, computer vision solutions to valence recognition rely on low-level visual descriptors such as color, texture, and shape (e.g., Machajdik & Hanbury 2010, Yanulevskaya et al. 2008, 2012); psychological models and art theories (e.g., Colombo et al. 1999, Solli & Lenz 2009, Wang & Yu 2005). Contrarily, Paper IV grounds itself on the findings which demonstrate influence of emotional content on the eye movement patterns of an observer (e.g., Nummenmaa et al. 2006, Niu et al. 2012, Humphrey et al.
Learning Phase

Affective Image Classification

Human

EM

Feature Extraction makes feature descriptors from eye movement (E.M) data

Label

Affective Image Corpus

Learn a Classifier

Test Phase

Affective Image Classification

Human

EM

Feature Extraction makes feature descriptors from eye movement (E.M) data

Classifier

Presenting images

unseen image

It is . . .

Fig 14. An overview of the valence recognition system. In the training, it learns a classifier using a corpus of affective images. In the test phase, it tries to predict the valence from eye movements. (©2014 IEEE. Reprinted from Paper IV, with permission from IEEE).

2012). It moves away from low-level visual features and considers eye movements, which are somehow a proxy to contextual information that is reported an essential missing component (Machajdik & Hanbury 2010). Furthermore, it studies the predictive power of different characteristics of eye movement in the task of valence recognition.

Fig. 14 depicts an overview of the proposed recognition system. In practice, it adopts the features introduced formerly for scene recognition. To perform classification, it applies SVM, which is formulated in terms of fast additive kernels (Maji et al. 2008), in 1-vs-rest scheme (i.e., there exist one classifier per valence category). Also, it trains separate classifiers for each feature type and combines them as needed by averaging the scores over individual classifiers.

Paper IV also conducts a series of experiments on the applicability of salience maps for valence recognition. The findings reveal that salience models, similar to the visual low-level descriptors, are not as successful as the statistics extracted from eye movement. This study suggests that eye movement is a strong modality for recognition of valence of a scene.

5.4 Discussion

There is a narrow margin between salience modeling and bottom-up visual attention models (see, Tsotsos 2011, for insight). Consequently, while most of the applications discussed above may have adopted the concept of saliency, in practice, they are occasionally elaborating an attention model. To avoid confusion and argumentation, this chapter only considered the application and averted model discrimination.
The second part of this chapter was about eye movements. It investigated its incorporation in salience modeling, scene recognition, and valence recognition. Considering the performance boost obtained by incorporating eye movements into salience models and behavioral studies, that signify the role of eye movements in perception, the future of saliency modeling is chained to attention models and capability of inducing oculomotor properties. Nonetheless, the implementation of such phenomena presents a bottleneck for introducing efficient models.

Eye movements provide not only the essential ingredient of saliency model assessment (i.e., fixations), but also useful information to decode the cognitive state of mind. Paper III augments salience models with these features to evaluate the performance of salience models in a scene recognition task. Later, Paper IV adopted the eye movements as a utility to recognize valence of a scene.

The valence recognition task investigated eye movements in a scenario that represents biosignal crowdsourcing. In other words, it fuses the information extracted from the eye movements of several observers in order to judge emotional content of a scene. A situation that may happen by popularity of personalized wearable biosignal recorders. While such a study is interesting, it is challenged with limitations and difficulties like: data gathering limitations, emotional level differences in people caused by gender (Lithari et al. 2010), education, age, etc. Nonetheless, it somehow lights a torch towards a future path of salience modeling research. It demonstrated that fixation duration, which is influenced by top-down information modulation (Irwin 2004), is the most influential feature that current salience models miss. Hence, the community should probably expect and investigate techniques for implementing more sophisticated models.
6 Evaluating saliency models

This chapter briefly talks about the evaluation techniques of salience models. The literature of salience modeling is replete with various evaluation criteria. Indeed, choosing the appropriate criteria is important in order to obtain a vivid understanding of a model and its performance. Nonetheless, the underlying bias in the eye movement data has imposed a metric challenge. To address such a challenge, several benchmarks are proposed, e.g., Judd et al. (2012), Borji et al. (2013a).

6.1 Bias and evaluation

To date, there seems to be a unanimous agreement on the presence of bias (i.e. skewness) in underlying structure of datasets. Consequently, there are studies to address the effect of bias in visual datasets. For instance, Torralba & Efros (2011) proposed a toy experiment to investigate bias in visual recognition datasets by trying to discriminate different datasets from given image samples of each. Not surprisingly, it concludes existence of strong build-in bias in datasets. They identify three biases in computer vision data, namely: selection bias, negative set bias, and capture bias.

Selection bias is caused by preference of a particular kind of image during data gathering. It results in identical and qualitatively similar images in a dataset. This is evident in the saliency segmentation datasets (e.g., MSRA benchmark dataset (Liu et al. 2007)) where the size of the object of interest is identical in several datasets (Borji et al. 2012b).

Negative set bias is the consequence of the lack of a rich and unbiased negative set. In other words, one should avoid being focused on a particular image of interest and datasets should model the whole world. Otherwise, a fair assessment of techniques would be difficult. Negative set bias is not only of special importance of machine learning cases, but also it may affect the eye movement recording experiments by familiarizing the observers to the stimuli in a free-viewing task over time. Thus, having a variety of images is motivated in such datasets.

Capture bias conveys the effect of image composition on the dataset. In other words, the location of elements and objects affects the dataset. The most popular kind of such a bias is the tendency of composing objects in the central region of the image. It, along with the viewing strategy, is a reason of the center bias phenomenon, in which
humans appear to preferentially look near the center of an image (Tatler 2007), in salience modeling datasets. Fig 15 depicts center bias in some of the popular datasets by visualizing the Mean Eye Position (MEP) of all observers.

The existence of bias in a dataset imposes limitations and concerns. While one can deliberately utilize bias as a prior in a specific task in order to improve the performance of an algorithm, it makes the generic quantitative evaluation of models difficult and sometimes even misleading. For instance, a toy saliency model, which consists of a Gaussian blob at the center of image, often scores higher than many saliency models (Judd et al. 2009). Consequently, explicit/implicit inclusion of a location prior somehow boosts the performance of models.

A similar to center bias phenomenon is “border effect”, in which an algorithm neglects the responses on the border of the image and equates them to zero (Van der Wal & Burt 1992, Tsotsos et al. 1995, Zhang et al. 2008). Although it may be in favor of a computer vision application, it definitely affects one’s understanding of an algorithm. In order to provide a bias free realization of the performance of models, three remedies are often suggested, that are: 1) adding central bias to all models, 2) collecting capture-bias-free datasets, 3) designing metrics to consider such phenomena.

Capture bias is probably the most prominent bias in saliency modeling datasets because it significantly affects the ground truth. However, it is not the sole bias. Selection bias affects the size of the elements of images in a dataset. Some methods start responding drastically to the change of smoothing factor value due to such variations, e.g., the model of Hou et al. (2012) is sensitive to the change of smoothing factor. In fact, to assess salience models, one should consider several influencing factors. To cover some of the shortcomings in salience model comparisons, Paper III argues about benchmarks and the challenges in this context. It extends the assessment of saliency
models to the stimuli/task decoding domain and compares 32 models in a unified quantitative framework over four widely-recognised datasets.

6.2 **Salience evaluation metrics**

This study devises the use of the measures of salience evaluation into two major groups of similarity-based and fixation-based. Although they may employ the same metric, they treat ground truth and saliency map differently. Similarity-based techniques establish a relation between saliency map and a reference map. To this end, usually a human fixation map is smoothed to obtain a continuous fixation density map as reference. On the other hand, the fixation-based methods analyze the saliency map on the location of fixations. The rest of this section explains some popular metrics under the umbrella of the two mentioned models.

6.2.1 **Similarity-based evaluation**

Probably, the most widely recognized score is Receiver Operating Characteristic (ROC) analysis. It is usually summarized in terms of Area Under the ROC Curve (AUC). This section addresses ROC analysis of two continuous maps. One map is the ground truth (i.e., a human fixation density map) and the second is the predicted results (i.e., saliency map). Traditionally, continuous maps are treated as a binary classifier to discriminate salience of each pixel. The procedure applies a threshold operation to one of the maps and computes the quality of classification given the other one. To obtain a percent salient curve, one can threshold the ground truth in order to preserve a constant percentage of image pixels. However, it is possible to apply a systematic moving threshold on the saliency map like traditional ROC analysis in signal detection theory. Le Meur & Baccino (2013) discussed the issue in details.

The Kullback-Leibler (KL) divergence is a metric of dissimilarity between two probability density functions. It is defined as

$$KL(Q, P) = \sum_k p_k \log\left(\frac{q_k}{p_k}\right),$$

where $P$ and $Q$ are discrete distributions with probability distribution functions $p_k$ and $q_k$, respectively. It has a lower bound of zero which is achieved when the two maps are identical. Le Meur et al. (2006) analyzed similarity of saliency maps and their reference fixation density maps using KL-divergence.

Judd et al. (2012) introduced a similarity score, $s$, to measure the similarity of continuous maps. It treats each map as a discrete distribution. It scales them to sum to
one and computes the sum of the minimum values at each point in the distributions; mathematically \( s = \sum_k \min(p_k, q_k) \). A score of one indicates absolute similarity and zero indicates total dissimilarity. They also motivated the use of Earth Mover’s Distance (EMD) (Pele & Werman 2008, 2009) which is a regional measure of similarity between two probability distributions; in this case, saliency map and its corresponding ground truth (i.e., human fixation density map).

Another popular metric is correlation coefficient (CC) between a human fixation density map \((H)\) and saliency map of a model \((S)\), \(CC(H, S) = \sum_{xy}(H(x, y) - \mu_H)(S(x, y) - \mu_S) / \sqrt{\sigma_H^2 \sigma_S^2} \) where \(\mu\) and \(\sigma^2\) are the mean and variance of the values in these maps and \(\sum_{xy}\) is the covariance matrix. It is a similarity metric with a well-defined upper-bound of one. In other words, two identical maps result in the correlation coefficient value of one. The major concern considering the similarity-based use of metrics is that they are easily affected by the center bias.

### 6.2.2 Fixation-based evaluation

Contrary to the similarity-based evaluation, which employ a fixation density map, fixation-based evaluation utilizes the fixation locations directly. This section starts with a variant of AUC which analyzes the salience map using fixation locations (e.g., Tatler et al. 2005, Bruce & Tsotsos 2006). It considers the fixation locations as a positive set and some uniformly random chosen points from the images as a negative set. Afterwards, it treats the saliency map as a binary classifier to separate positive and negative samples. A concern regarding computing AUC as described is that it produces a high score for a central trivial Gaussian blob (Zhang et al. 2008, Zhao & Koch 2011). Nonetheless, a slightly different variant, shuffled AUC (sAUC), tries to tackle the problem by making the negative set from all fixations of all images except the positive set of the image of interest (Zhang et al. 2008). It is worth noting that there are two ways of computing AUC and two ways of computing sAUC depending on the sampling strategy (e.g., Itti 2005, Ehinger et al. 2009, Judd et al. 2009, Parkhurst & Niebur 2003, Borji et al. 2013a). Supplement of Paper III provides an argument on them.

To assess salience performance, Itti & Baldi (2005a,b) utilized the KL-divergence as a measure to compare histograms of saliency sampled at eye fixations and that sampled at random locations. Contrary to the aforementioned similarity-based use of KL-divergence, an algorithm is good if it produces a high KL value because it is expected that the histogram of salience should be different from a random histogram. To
avoid center bias effect, (Zhang et al. 2008) proposed a shuffling scheme similar to sAUC.

Normalized Scanpath Saliency (NSS) (Peters et al. 2005) measures the average of response values of a saliency model, \( S \), at human fixation locations, \((x_i, y_i)\). In mathematical terms, \( \text{NSS} = \frac{1}{N} \sum_{i=1}^{N} \frac{(S(x_i, y_i) - \mu_S)}{\sigma_S} \), where \( N \) is the number of fixations for each image, \( \mu \) and \( \sigma \) are mean and standard deviation. A large positive value of NSS indicates a good salience model meanwhile \( \text{NSS} \leq 0 \) indicates that the model performs no better than picking a random position. The importance of this metric is that it facilitates the use of fixation order to perform through observer-based analysis. Nonetheless, it is usually a matter of less concern in salience modeling and often neglected.

### 6.3 Scanpath analysis and salience models

Not only humans are correlated in terms of the locations they fixate, but they also agree somewhat in the order of their fixations (Privitera & Stark 2000). Scanpath analysis establishes a relationship between two sequences of fixations. It often requires taking a number of factors, such as the temporal dimension or the alignment procedure, into account. Consequently, they are somehow complicated and few models aimed to predict scanpath sequence. Intrigued to measure and compare models in terms of their saccade generation ability, which is desired in some applications, e.g., advertising, Paper III proposed a measure of scanpath evaluation.

For each image, the proposed metric initially derives some clusters from human fixation map and codes the scanpath of each subject into a string using these clusters. Afterwards for the corresponding saliency map, it runs the IOR mechanism to generate a fixation sequence. It uses the Needleman-Wunch (Needleman & Wunsch 1970) to compare the scanpath of a model with an observer’s scanpath. Eventually, it reports the average score over all subjects. Algorithm 1 summarizes the details of the proposed evaluation measure.

To make clusters, the algorithm employs mean-shift which is a non-parametric clustering technique with no assumption on the shape and the number of clusters. Furthermore, it is robust toward outliers. It approximates maxima of a density function from discrete samples of that function. It is defined as an iterative multivariate kernel density estimation (Fukunaga & Hostetler 1975) in which the kernel bandwidth, denoted band-width in the algorithm 1, influences the shape and quality of clusters. The proposed
algorithm chooses the appropriate bandwidth such that the maximum transition between fixation clusters occur. In other words, it identifies the maximum interacting regions in an image using fixations.

Algorithm 1  Scanpath evaluation

- **PHASE 1:** generating human scanpath and clusters
  **Input:** subjects’ fixations.
  **Output:** fixation clusters & human scanpath strings.
  1. `BW ← List of all possible band-width values`
  2. `for each band-width ∈ BW do`
  3.  Compute a clustering using the mean-shift algorithm
  4.  Compute interaction between clusters
  5. `end for`
  6.  Chose the band-width with the highest interaction rate
  7.  Repeat the clustering for the chosen band-width and store clusters
  8.  Assign a unique character to each cluster
  9. `for each subject do`
  10.  Generate a string for scanpath by finding the closest cluster centers
  11.  Store the string sequence
  12. `end for`
  13.  Repeat the above process for all the images in the database

- **PHASE 2:** Model evaluation on a given test image
  **Input:** model generated saliency maps.
  **Output:** model’s overall score.
  1. `for each image do`
  2.  Take a model saliency map
  3.  Load the sequence strings and clusters corresponding to that image
  4. `for each subject do`
  5.  Generate an equal length sequence running IOR
  6.  Generate a string sequence by finding the closest cluster
  7.  Assign a matching score to the two strings
  8. `end for`
  9.  Save the average score over all subjects
  10. `end for`
  11. Report the average matching score over all images

6.4 Benchmarks

There are two widely renowned benchmarks available in the context of salience modeling. Borji et al. (2013a) compared 36 models over three datasets, they exploited three metrics of NSS, CC, and sAUC to compare saliency models. The emphasize is on the sAUC metric and neutralizing the center bias. On the other hand, Judd et al. (2012) stress on a dataset with unavailable ground truth, which consists of 300 images,
Table 1. Benchmark of Judd et al. (2012), the performance results of some example models, including Paper I.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>(s)</th>
<th>EMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.922</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Judd et al. (2009)</td>
<td>0.811</td>
<td>0.506</td>
<td>3.13</td>
</tr>
<tr>
<td>CovSal (Erdem &amp; Erdem 2013)</td>
<td>0.8056</td>
<td>0.5018</td>
<td>3.1092</td>
</tr>
<tr>
<td>Paper I</td>
<td>0.8033</td>
<td>0.4952</td>
<td>3.3488</td>
</tr>
<tr>
<td>GBVS (Harel et al. 2007)</td>
<td>0.801</td>
<td>0.472</td>
<td>3.574</td>
</tr>
<tr>
<td>Gaussian</td>
<td>0.783</td>
<td>0.451</td>
<td>3.719</td>
</tr>
<tr>
<td>AIM (Bruce &amp; Tsotsos 2006)</td>
<td>0.751</td>
<td>0.39</td>
<td>4.236</td>
</tr>
<tr>
<td>Torralba et al. (2006)</td>
<td>0.684</td>
<td>0.343</td>
<td>4.715</td>
</tr>
<tr>
<td>SUN (Zhang et al. 2008)</td>
<td>0.672</td>
<td>0.34</td>
<td>5.088</td>
</tr>
</tbody>
</table>

To prevent models from fitting themselves to the underlying bias, it utilizes EMD, similarity score \((s)\), and AUC. Paper III elaborates the benchmark of Borji et al. (2013a) by introducing new insights. This study advocates that there is no single benchmark and measure of evaluation for salience modeling. In other words, each metric shed light on one/some aspect of a model and it is the responsibility of the researchers to provide an unbiased interpretation of the results. To provide an example on how the choice of metrics matters, the performance of the model of Paper I is here demonstrated in the two aforementioned benchmarks. The scores based on Judd et al. (2012) are the following: AUC=0.8033, \(s = 0.4952\), EMD=3.3488. Table 1 provides the ranking order of the method compared to several models, including human, a model to demonstrate how \(n\) observers predict other observers, and Gaussian, which is a Gaussian kernel at the center of image. As depicted, Paper I has a good place in the chart ranking fourth after the human. In essence, this benchmark reveals that the model of Paper I performs comparably better than Gaussian providing somewhat similar to ground truth saliency maps. Meanwhile, it has a high detection rate based on AUC score.

Let’s consider the same models using \(s\)AUC metric provided by the benchmark of Borji et al. (2013a). Table 2 provides the performance of \(s\)AUC metric. As depicted, \(s\)AUC reveals a different performance ranking. Interestingly, the methods reported, by Judd et al. (2009), performing worse than trivial Gaussian saliency are assessed to perform better. It tells one that these models are accurate when the center bias is neutralized. One can somehow interpret it as how accurate these models are on the
Table 2. sAUC score for some selected models including, Paper I on four different datasets of Toronto (Bruce & Tsotsos 2006), NUSEF (Ramanathan et al. 2010), MIT (Judd et al. 2009), and (Kootstra et al. 2008); rank reports on how a method performs on average on all the datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>Toronto</th>
<th>NUSEF</th>
<th>MIT</th>
<th>Kootstra</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human (Bruce &amp; Tsotsos 2006)</td>
<td>0.73</td>
<td>0.66</td>
<td>0.75</td>
<td>0.62</td>
<td>-</td>
</tr>
<tr>
<td>AWS (Garcia-Diaz et al. 2009a)</td>
<td>0.72</td>
<td>0.64</td>
<td>0.69</td>
<td>0.62</td>
<td>1</td>
</tr>
<tr>
<td>AIM (Bruce &amp; Tsotsos 2006)</td>
<td>0.69</td>
<td>0.64</td>
<td>0.68</td>
<td>0.59</td>
<td>2.5</td>
</tr>
<tr>
<td>Torralba et al. (2006)</td>
<td>0.69</td>
<td>0.63</td>
<td>0.67</td>
<td>0.59</td>
<td>3</td>
</tr>
<tr>
<td>Judd et al. (2009)</td>
<td>0.58</td>
<td>0.61</td>
<td>0.66</td>
<td>0.59</td>
<td>4</td>
</tr>
<tr>
<td>SUN (Zhang et al. 2008)</td>
<td>0.67</td>
<td>0.61</td>
<td>0.65</td>
<td>0.56</td>
<td>5.3</td>
</tr>
<tr>
<td>Paper I</td>
<td>0.64</td>
<td>0.56</td>
<td>0.65</td>
<td>-</td>
<td>6.5</td>
</tr>
<tr>
<td>GBVS (Harel et al. 2007)</td>
<td>0.65</td>
<td>0.59</td>
<td>0.64</td>
<td>0.56</td>
<td>6.5</td>
</tr>
<tr>
<td>Gaussian</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>-</td>
</tr>
</tbody>
</table>

off-center of the image. Nonetheless, this is just one perspective of these models. Considering the model of Paper I, sAUC reveals that it is a moderate model once the center bias is neutralized. Nonetheless, it is expected because it utilizes location prior.

Considering the scanpath prediction measure, the method of Paper I performs the best on Toronto dataset and acceptable on MIT dataset. The other well-performing methods are GBVS and Judd (see, Paper III). It underlines the scanpath prediction power of some models which are not necessarily the best considering other metrics.

The lesson learnt from this example is that no single metric is enough for judging all the aspects of a salience model. Therefore, Paper III advocates different metrics to benchmark models. It even goes further and studies the performance of models in the task of image category recognition.

6.5 Discussion

The bias in the dataset has several aspects. To date, in salience modeling, the focus has been on a small, but crucial, set of them, i.e., center bias and border effect. Nonetheless, there is evidence that there are more affecting factors involved. For instance, Niu et al. (2012) demonstrated how the presence of emotional content violates expected eye movement patterns. In fact, it is expected that the degree and type of the emotion conveyed in an image affect the ground truth and performance of salience models. Intrigued, Paper III assessed some of the models in the presence of emotional stimuli.
and neutral stimuli. The models express a performance drop when exposed to emotional stimuli, e.g., the sAUC score of AWS model is 0.59 for emotional stimuli and 0.69 for neutral stimuli (see, Paper III for details). While it promises room for future work, it alerts the requirement of careful attention to stimuli and subjects due to the volatile effect of a stimulus on the observer, e.g., emotional differences between male and female observers (Lithari et al. 2010).

It is worth noting to underline that several metrics are necessary to clarify different aspects of models in a benchmark. Furthermore, researchers should consider the bounds of metrics. Often, a metric with well-defined lower-bound and upper-bound is preferable, e.g., sAUC (Borji et al. 2013a) has well-defined bounds and is demonstrated to be of importance (Riche et al. 2013). The scanpath evaluation metrics take into account the order of fixation sequences of each observer. Thus, they are suitable for the detailed analysis of the ability of replicating scanpath given a model. In this respect, the emphasize of the scanpath metric presented in Paper III is on the interaction between the regions of a scene, which makes it suitable for interaction analysis. Nonetheless, the average of score somehow provides some intuition to the overall performance of a model and its presence is complementary in a benchmark.

Another perspective to discuss is the performance of models in a computer vision applications. In this case, one shall compare the salience model with the methods and techniques designed for that application in order to demonstrate the usefulness of the model. For instance, the model of Paper VI was compared with several background subtraction techniques and reported successful to rank 3rd in the background modeling competition and 4th among eleven techniques (Vacavant et al. 2012).

In conclusion, the results of Paper III indicates a small gap (but statistically significant) between models and human, which highlights a significant progress in salience modeling. Nevertheless, it alerts benchmarks to watch for fitting of models to the data. Hence, gathering of new datasets is recommended. In the domain of benchmarks, a future trend can be designing standard and more challenging tasks and considering other features and influencing factors such as scanpath information, populations of observers, or stimulus category.
7 Summary

This study investigated visual saliency modeling and eye movements in the computer vision domain. It tackled several perspectives by introducing salience models, investigating their applications, and examining the role of eye movements. Here, the contributions are summarized and future works are elaborated.

This thesis introduced a Bayesian center-surround salience model. The model utilizes a non-parametric kernel density estimation under two assumptions of singleton center and a surround consists of a sufficient small number of pixels. Later, it extended the proposed model to handle video stimuli. This model is capable of incorporating any kind of prior information, which facilitates top-down task related manipulation.

The study developed the local similarity number (LSN) model for tracking application. Similar to the Bayesian center-surround, the LSN relies on center-surround differences. It successfully employed LSN to target representation in a mean-shift tracking framework. To find an unsupervised solution to the challenge of background subtraction, it introduced temporal motion model and spatio-temporal motion model. It grounds them on the significance of motion cue by utilizing the concept of flicker. Assessments reported their success in the assigned application.

Motivated by the role of eye movements in human perception, it implemented a model of saccade generation to replicate saccadic eye movements. It employed the model to introduce a fixation prediction scheme (i.e., a salience model). The model somehow incorporates the oculomotor bias into the saliency modeling.

Eye movements have several characteristics which can be exploited for various purposes. This study investigated the combination of features extracted from them and salience in recognition of a scene. Eventually, it resulted in a simple scene recognition task to benchmark the performance of salience models. Furthermore, it utilized eye movement features and saliency in the task of emotion recognition. The findings reveal that salience is not as successful as features extracted from eye movements and portrays that the statistics extracted from eye movements is a strong cue in the recognition of a scene valence.

In the context of benchmarks, this study introduced a scanpath evaluation criteria, in which the novelty lies on considering the interactive clusters of fixations that represent interactive regions of a scene. It provides complementary scanpath performance
information to available benchmarks. Also, it is a useful asset to detailed analysis of 
scanpath prediction power of a model.

Despite the aforementioned contributions, there are still challenges in this area to be 
tackled. To date, there is no standard data set that covers broad perspectives of data to 
facilitate multidisciplinary investigation. For instance, the interaction of stimuli and 
observer is neglected in the datasets, e.g., the effect of image content on observers 
viewing strategy as well as observers’ experience and background such as education, 
gender, ethnic, etc.

This study kept the application of salience as simple as possible in order to assess 
the usefulness of the concept of salience in algorithms. Consequently, the presented 
models can be further extended to fit better into the application. For instance, in the 
domain of tracking, tracker recovery using salience models can still be investigated. 
Nonetheless, it requires a more complicated tracking framework.

This study investigated the usefulness of models that take into account the role of 
eye movements. The results prompt the investment in affordable biologically-motivated, 
though complicated, models that comprise overt attention or oculomotor behavior. Thus, 
extending the proposed saccade generation model to a full attention model is a potential 
future direction of research, likewise, is adapting it to handle videos. Similarly, an 
interesting area to be studied in the future is the role of eye movements characteristics 
and their contribution to salience models which is supported by the contributions of this 
study in the area of scene and valence recognition.
References


Online pattern recognition and machine learning techniques for computer-vision: Theory and applications.


Borji A, Sihite DN & Itti L (2012c) An object-based bayesian framework for top-down visual attention.


Filipe S & Alexandre L (2013) From the human visual system to the computational models of visual attention: a survey. Artificial Intelligence Review 1–47.


Yarbus AL (1967) Eye Movements and Vision. PLENUM PRESS.


Original publications


Reprinted with permission from Springer (I, V, VI, VII), Elsevier (II) and IEEE (III).

Original publications are not included in the electronic version of the dissertation.
490. Aapoja, Aki (2014) Enhancing value creation of construction projects through early stakeholder involvement and integration
492. Sliz, Rafal (2014) Analysis of wetting and optical properties of materials developed for novel printed solar cells
495. Mehtonen, Saara (2014) The behavior of stabilized high-chromium ferritic stainless steels in hot deformation
496. Majava, Jukka (2014) Product development: drivers, stakeholders, and customer representation during early development
497. Myllylä, Teemu (2014) Multimodal biomedical measurement methods to study brain functions simultaneously with functional magnetic resonance imaging
498. Tamminen, Satu (2014) Modelling the rejection probability of a quality test consisting of multiple measurements
499. Tuovinen, Lauri (2014) From machine learning to learning with machines: remodeling the knowledge discovery process
503. Tuhkala, Marko (2014) Dielectric characterization of powdery substances using an indirectly coupled open-ended coaxial cavity resonator

Book orders:
Granum: Virtual book store
http://granum.uta.fi/granum/
Hamed Rezazadegan Tavakoli

VISUAL SALIENCY AND EYE MOVEMENT: MODELING AND APPLICATIONS