FEDERAL UNIVERSITY OF TECHNOLOGY – PARANÁ
GRADUATE PROGRAM IN ELECTRICAL AND COMPUTER ENGINEERING

MAIKO MIN IAN LIE

AN EFFICIENT STRATEGY FOR ESTIMATION OF VISUALLY SALIENT REGIONS IN IMAGES

MASTER'S THESIS

CURITIBA
2018

MAIKO MIN IAN LIE

AN EFFICIENT STRATEGY FOR ESTIMATION OF VISUALLY SALIENT REGIONS IN IMAGES

Thesis presented to the Graduate Program in Electrical and Computer Engineering (CPGEI) of the Federal University of Technology – Paraná (UTFPR), in partial fulfillment of the requirements for the degree of Master in Computer Engineering.

Advisor: Hugo Vieira Neto
Co-advisor: Gustavo Benvenutti Borba

CURITIBA
2018

Dados Internacionais de Catalogação na Publicação

L716e 2018

Lie, Maiko Min Ian
An efficient strategy for estimation of visually salient regions in images / Maiko Min Ian Lie.-- 2018.
69 f.: il.; 30 cm

Disponível também via World Wide Web.
Texto em inglês com resumo em português.
Dissertação (Mestrado) - Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial. Área de Concentração: Engenharia de Computação, Curitiba, 2018.
Bibliografia: f. 62-69

1. Percepção visual. 2. Atenção - Simulação por computador. 3. Visão por computador. 4. Processamento de imagens - Técnicas digitais. 5. Algoritmos computacionais. 6. Sistemas de reconhecimento de padrões. 7. Métodos de simulação. 8. Engenharia de computador. 9. Engenharia elétrica - Dissertações. I. Vieira Neto, Hugo, orient. II. Borba, Gustavo Benvenutti, coorient. III. Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial. IV. Título.

CDD: Ed. 23 -- 621.3

Biblioteca Central do Câmpus Curitiba – UTFPR
Bibliotecária: Luiza Aquemi Matsumoto CRB-9/794

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Diretoria de Pesquisa e Pós-Graduação

TERMO DE APROVAÇÃO DE DISSERTAÇÃO Nº 790

A Dissertação de Mestrado intitulada **“An Efficient Strategy for Estimation of Visually Salient Regions in Images”** defendida em sessão pública pelo(a) candidato(a) **Maiko Min Ian Lie**, no dia 28 de março de 2018, foi julgada para a obtenção do título de Mestre em Ciências, área de concentração Engenharia de Computação, e aprovada em sua forma final, pelo Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial.

BANCA EXAMINADORA:

Prof(a). Dr(a) Hugo Vieira Neto - Presidente (UTFPR)
Prof(a). Dr(a). Mylène Christine Queiroz de Farias - (UnB)
Prof(a). Dr(a). William Robson Schwartz - (UFMG)
Prof(a). Dr(a). Bogdan Tomoyuki Nassu - (UTFPR)

A via original deste documento encontra-se arquivada na Secretaria do Programa, contendo a assinatura da Coordenação após a entrega da versão corrigida do trabalho.

Curitiba, 28 de março de 2018.

ACKNOWLEDGMENTS

The work presented in this thesis would not be possible without the contributions of several people. First of all, I would like to thank my advisor, prof. Hugo Vieira Neto, for providing thorough and thoughtful feedback for my writing since I was an undergraduate student. Besides being a source of inspiring work, he was the first person to associate my writing with the word "scientific". As a struggling undergraduate student, the suggestion that I might have what it takes to contribute to scientific research someday meant a lot to me. I still keep a printed copy of that old report with his annotations as a memento — it reminds me to never take a thoughtful review for granted. I have no doubt that his words had an important role in my choice of pursuing a career in science. For that I will always be grateful.

I would also like to thank my co-advisor, prof. Gustavo Benvenutti Borba, with whom I worked closely and learned a lot from during my undergraduate research days, which is when I first had contact with the subject of this thesis. During the time we worked together, I learned many valuable lessons. However, the one which will stay with me the most is the role of positive feedback on the growth of an aspiring researcher. When you care deeply about your work, much of your time is spent in polishing subtle but important aspects, such as writing style, visual composition, typography, and such minutiae. Most of the time, this goes unnoticed, which is a success in a way, since the purpose is clarity over embellishment. Still, there are no words to describe how gratifying it is when someone does notice and appreciate such details. If this is brought up accompanied by so many free beers that you increase your tolerance to alcohol significantly, all the better. I have no way to return the patience, free beers, attention, free beers, opportunities, free beers, and friendship with anything besides these few words. Thank you.

During the last few years, most of my days were spent at the *Imaging and Electronic Instrumentation Laboratory* (LABIEM), where I had the opportunity to meet very interesting people to which I am grateful for making the hard work much more tolerable. Among these people are João Pedro Curti, André Lucas Zanellato, Bianca Alberton Visineski, Cristian Bortolini Ferreira, Germano Rosa Figueiredo, Mauren Abreu de Souza, Nelson Garcia de Paula, Valfredo Pila Jr., Hellen Mathei Della Justina, and Flavio Henrique Galon. This list would not be complete without prof. Humberto Remigio Gamba, who created this place in which we spend so much time of our lives. Run-

ning this place and ensuring that we have the necessary conditions to perform our jobs is not an easy matter, especially when one is constantly overwhelmed with as many administrative responsibilities as he is.

I cannot say that my social skills improved during the last few years, but I did get the chance to socialize a little, in part because the people from the laboratory next corridor (LAPIS) were kind enough to invite me to chat and have some cake every once in a while. So thank you Eduardo Tondin Ferreira Dias and Ricardo Fantin da Costa.

While I had many colleagues, I had very few close collaborators, so a special thank you goes to Andriy Guilherme Krefer, for the very insightful conversations and hard work during the few opportunities in which we got to work together. After all, there are not many people with whom I can claim to have watched thriller movies in the lab at 4 am, due to being locked up on campus doing late night data analysis. By the way, if I am not mistaken, the first time I got my name printed on a scientific document was on the acknowledgments section of his thesis, so the debt is paid :-)

It cannot go unnoticed that I was fortunate enough to have had an excellent thesis committee, to whom I am very grateful. Prof. Mylène Christine Queiroz de Farias gave thoughtful feedback from a psychovisual perspective, backed by experience with visual perception in video quality assessment. Prof. William Robson Schwartz provided a just as thoughtful assessment from a visual pattern recognition perspective, backed by experience with challenging smart surveillance tasks. Prof. Bogdan Tomoyuki Nassu is someone I have known for years, since I took his *Introduction to Computer Vision* class as an undergraduate. We have always shared an admiration for the insightful aspects of computer vision, which resulted in very entertaining conversations. So interesting, in fact, that I could not help but take his class again as a graduate student. As expected from our previous interactions, his feedback as a thesis committee member did not disappoint in the least.

A less personal acknowledgment, but not less important, goes to the *Brazilian Coordination for Improvement of Higher Education Personnel* (CAPES), for the fellowship that allowed me to dedicate myself exclusively to research during the last two years. I would also like to acknowledge the *Graduate Program in Electrical and Computer Engineering* (CPGEI), for the financial support with the travel expenses for my conference presentations as well as for the participation of the external thesis committee members.

A large aspect of the art of creating artificial vision, or any neurobiologically inspired application, is to select the right subset and to determine the best way to translate those hints into enabling elements.

JOHN K. TSOTSOS

ABSTRACT

LIE, Maiko Min Ian. An Efficient Strategy for Estimation of Visually Salient Regions in Images. 2018. 69 f. Master's Thesis, Graduate Program in Electrical and Computer Engineering, Federal University of Technology – Paraná. Curitiba, 2018.

The information incident on the human visual system is bound by a selection mechanism, known as *visual attention*. This mechanism is responsible for restricting incoming visual information to a smaller and potentially important subset for further processing, enabling the visual system to respond rapidly, despite the enormous amount of information to which it is subject. Computer vision systems often employ reproductions of this mechanism in order to reduce visual search space, since this strategy can lead to substantial improvement in efficiency. This thesis addresses the problem of efficient computation of visual attention, particularly the case of *salient region detection*. A strategy based on joint upsampling of coarse-scale saliency estimates is presented for that purpose. This approach allows leveraging both the advantages of coarse-scale estimation (reduction of computational cost, abstraction of unnecessary details) and fine-scale edge information (high accuracy). Based on the highly redundant data and spatially-varying importance of content in images of real-world scenes, two visual saliency formulations are presented for coarse-scale estimation in the proposed strategy. The first approach operates on a pixel level, based on random color distances. The second approach operates on a patch level, based on the reconstruction error computed from the Principal Component Analysis of the image boundaries. The efficacy and efficiency of the proposed strategy are demonstrated through assessment on the ASD, MSRA10K, ECSSD, and DUT-OMRON datasets. Comparison with other seven state-of-the-art methods in terms of precision, recall, F-measure, and execution time demonstrate that the proposed strategy is highly competitive, achieving one of the best trade-offs between accuracy and execution time.

Keywords: Visual attention, saliency detection, computer vision.

Resumo

LIE, Maiko Min Ian. Uma Estratégia Eficiente para Estimação de Regiões Visualmente Salientes em Imagens. 2018. 69 f. Dissertação, Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial (CPGEI), Universidade Tecnológica Federal do Paraná. Curitiba, 2018.

A informação incidente no sistema visual humano é limitada por um mecanismo de seleção, conhecido como *atenção visual*. Este mecanismo é responsável por restringir a informação visual incidente a um subconjunto menor e potencialmente importante para processamento adicional, permitindo que o sistema visual responda rapidamente apesar da enorme quantidade de informação ao qual normalmente está sujeito. Sistemas de visão computacional empregam reproduções deste mecanismo para redução de espaço visual, visto que essa estratégia pode levar a substanciais ganhos em eficiência. Esta dissertação trata do problema de computação eficiente de atenção visual, em particular o caso de *detecção de regiões salientes*. Uma estratégia com base em sobreamostragem conjunta (*joint upsampling*), de estimativas de saliência em baixa resolução é apresentada com esse propósito. Isso permite explorar tanto as vantagens de estimativa em baixa-resolução (redução de custo computacional, abstração de detalhes desnecessários) quanto as de bordas em alta-resolução (alta acurácia). Com base na alta redundância de dados e importância espacialmente-variável no conteúdo de imagens de cenas reais, duas formulações de saliência visual são apresentadas para estimativa em baixa resolução na estratégia proposta. A primeira opera em nível de pixel, baseada em distâncias de cor aleatórias. A segunda opera em nível de *patch*, baseada em erro de reconstrução por bases obtidas através de Análise de Componentes Principais nas margens da imagem. A eficácia e eficiência da estratégia proposta são demonstradas através de avaliação nos bancos de imagens ASD, MSRA10K, ECSSD, e DUT-OMRON. Uma comparação com outros sete métodos do estado-da-arte em termos de precisão, abrangência, *F-measure* e tempo de execução demonstra que a estratégia proposta é altamente competitiva, alcançando uma das maiores relações custo-benefício entre acurácia e tempo de execução.

Palavras-chave: Atenção visual, detecção de saliência, visão computacional.

List of Figures

List of Tables

Contents

1 Introduction

The human visual system is subject to a massive amount of input data, such that it is unfeasible to entirely process it in detail. It has been argued that, if unbounded, the problem of visual search, and perhaps visual perception in general, is computationally intractable (TSOTSOS, 1990). Yet, most people are capable of efficiently performing a wide range of visual tasks even in visually complex environments. The key for this efficiency is that, in fact, *human vision is bounded* by an information reduction mechanism. This mechanism, known as *visual attention*, prevents the overload of the human visual system by allocating its processing resources only to potentially important parts of its input. In other words, it performs a visual search space reduction. This thesis addresses the problem of efficient computation of visual attention.

1.1 Background

While several experimental studies have investigated the operation of human visual attention, a fundamental contribution, and perhaps the most influential, is the *Feature Integration Theory* by Treisman and Gelade (1980). It hypothesized that elementary properties of the scene are registered early, automatically and in parallel across the visual field, and that their conspicuity is subsequently used to select regions for allocation of attention. This theory heavily influenced the neurally plausible architecture proposed by Koch and Ullman (1985), which laid out an approach for saliency-based visual attention whose aspects are adopted to this day. Despite the significant importance of this architecture, it was not its conceptual model, but arguably its later computational implementation by Itti, Koch and Niebur (1998) that popularized the adoption of visual attention in technical applications, and established it as a major research theme in computer vision. Their work demonstrated not only that a neurally plausible feature integration approach is computationally practical, but that it can effectively predict human performance on visual search tasks with images of real-world scenes.

An often overlooked aspect from these early studies is of particular importance to the current popularity of visual attention models in computer vision. This aspect is the very fortunate choice by Koch and Ullman (1985) of formalizing the encoding of early-feature conspicuity as a *saliency map*, although its notions were already present in

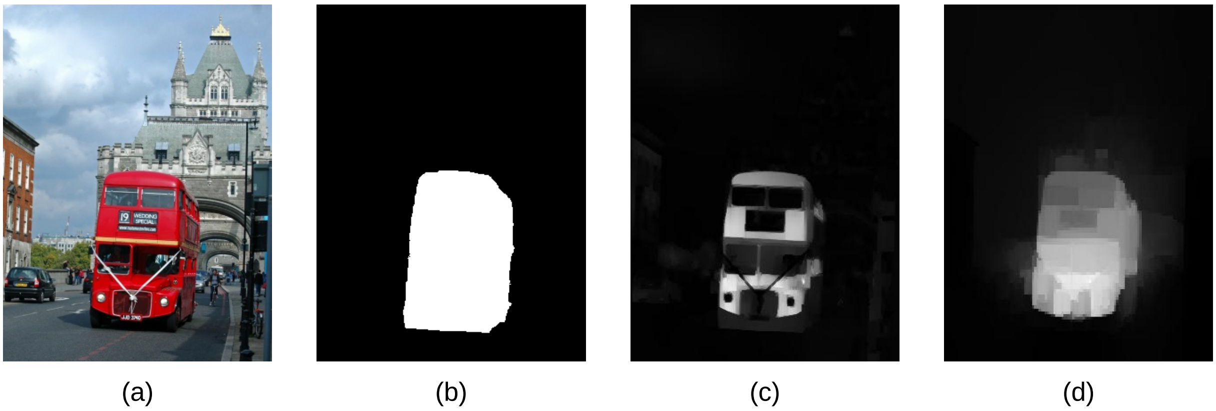

Figure 1: Computation of bottom-up visual attention (i.e., *saliency map*). **(a)** Input image. **(b)** Ground truth (i.e., human labeling). **(c)–(d)** Saliency maps using the pixel-level and patch-level approaches proposed in this thesis, respectively. A saliency map highlights visually distinctive regions of the scene, which are likely to attract visual attention in the absence of explicit tasks.

the earlier *master map* proposed by Treisman and Gelade (1980). Computationally, the saliency map is simply a grayscale image with the intensity at each location describing its corresponding conspicuity (Figure 1). This allows straightforward inclusion into vision applications and easy assessment. It is also versatile in the sense that it is general enough to encode both bottom-up (i.e., purely stimulus-based) and top-down (i.e., semantic) aspects of attention (BRAUN; KOCH; DAVIS, 2001). More importantly, despite the origins of the concept, a saliency map does not impose any particular visual attention architecture. The implication is that, at least technically, any strategy for feature conspicuity estimation may be employed for bottom-up visual attention, as long as it outputs a useful saliency map. In fact, the terms *bottom-up visual attention* and *saliency detection* are often used interchangeably, the latter being arguably more common.

1.2 Motivation and scope

Given the large number of approaches for feature conspicuity computation, each one leading to significantly different saliency maps, bottom-up visual attention models account for a substantial part of current visual attention research. An extensive benchmark by Borji et al. (2015) assessed more than 30 saliency detection models only in the period between 1998 and 2014. These models span a wide range of formulations, including graph-theoretical (JIANG et al., 2013a), frequency-domain (HOU; ZHANG, 2007), probabilistic (ALPERT et al., 2012), among many others. Most of them differ basically in aspects such as feature set (e.g., color, intensity, orientation), locality (e.g., lo-

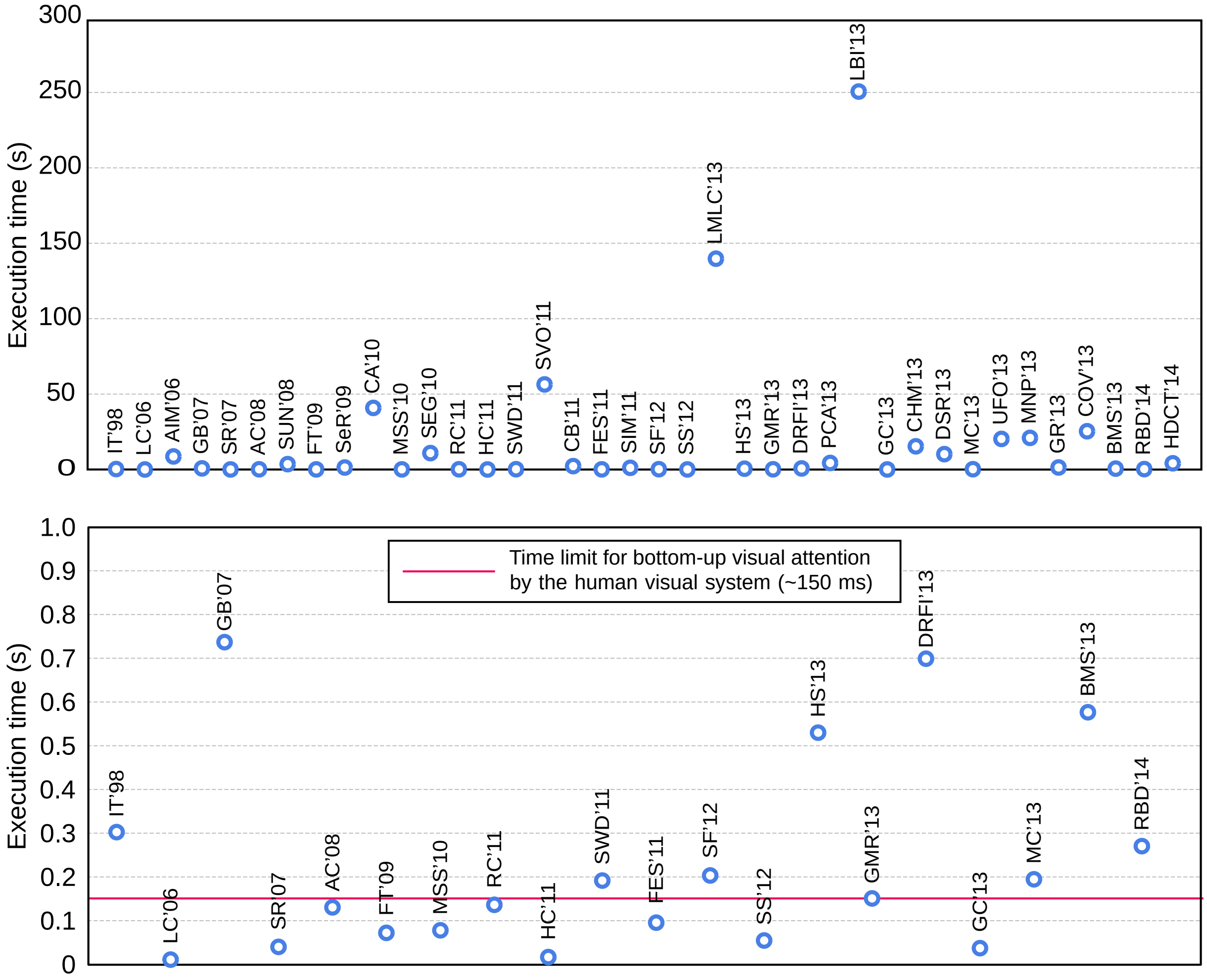

Figure 2: Average execution time of state-of-the-art saliency detection methods for an RGB image with 400×300 pixels, according to the benchmark by Borji et al. (2015). **Top:** 38 methods published from 1998–2014. **Bottom:** selection of algorithms with execution time under one second. The approximate time limit for bottom-up visual attention by the human visual system, 150 ms according to Theeuwes (2010), is indicated in red. Only 11 of the 38 methods perform within this time frame. The execution times reported are for a Xeon E5645 2.4 GHz CPU desktop with 8 GB RAM.

cal, global), scale (e.g., coarse, fine, multiple), granularity (e.g., pixel, patch, segments) and learning (e.g., unsupervised, supervised, reinforced).

While the increased interest in visual attention modeling is encouraging, a problem with many models being currently proposed is that, as they get increasingly sophisticated to achieve higher accuracy, in general, their computational performance decreases accordingly. In many cases, this severely limits their applicability, since saliency detection is usually not the task itself, but a pre-processing step prior to more elaborate processing. It is also worth noting that the original motivation for saliency detection is its effectiveness on the human visual system, in which it has been reported to take less than 150 ms (THEEUWES, 2010). For perspective, Figure 2 shows the average execution time of 38 saliency detection algorithms assessed by Borji et al. (2015), computed for an RGB image with 400×300 pixels. Only 20 of the 38 methods execute under a second, showing that around half of them are unlikely to perform in real-time, possibly delaying instead of accelerating further processing stages. Moreover, only 11 of those methods execute under the time frame estimated for bottom-up visual attention by the human visual system, several of them at the cost of drastic simplifications that severely limit their accuracy, some of which are discussed in this thesis. Of course, this is just a rough comparison since there is no hard threshold on the execution time required for a method to be useful. Still, this offers some perspective on the relevance of more efficient saliency detection models, which is the main subject of this work.

Regarding scope, this thesis is restricted to salient region detection with *intrinsic cues* (BORJI; ITTI, 2013) in the *the unsupervised setting*. In other words, only visual cues from the image itself (e.g., color) are used for saliency modeling, as opposed to *extrinsic cues*, such as additional saliency maps (i.e., *co-saliency detection setting* (ZHANG et al., 2018)) or manually labeled data (i.e., *supervised setting*). In this sense, while there are more accurate methods in the literature (e.g., Jiang et al. (2013b), Kim et al. (2014), Kim et al. (2016), Wang et al. (2017)), including those employing recently popular deep neural network models (e.g., Li et al. (2017a), Li et al. (2017b), Hou et al. (2017)), they are mostly supervised and out of the scope of this thesis.

1.3 Thesis statement

This work addresses the problem of efficient saliency detection. For this purpose, a completely bottom-up approach is proposed, which is unsupervised and makes minimal assumptions about the input image. In order to achieve short execution time without significantly sacrificing accuracy, a principled strategy is adopted, which leverages the properties of data redundancy and spatially-varying perceptual importance in images of real-world scenes by means of joint upsampling of coarse-scale saliency estimates. Thus, the central thesis of this work is as follows:

> *Most images of real-world scenes present highly redundant data and spatially-varying perceptual importance. This can be leveraged to design efficient and effective algorithms for estimation of visually salient regions in images. Redundancy can be exploited for efficiency by modeling saliency in terms of a subset of the image, while spatially-varying importance can be exploited for efficacy by biasing from where this subset is selected.*

The effectiveness and efficiency of the proposed approach are demonstrated by quantitative assessment on the ASD, MSRA10K, ECSSD, and DUT-OMRON datasets, in terms of precision, recall, F-measure, and execution time. Comparison with state-of-the-art approaches demonstrate that the proposed approach is highly competitive, achieving one of the best trade-offs between accuracy and execution time.

1.4 Contributions

The main contributions of the work presented in this thesis are:

- **An efficient salient region detection model.** The proposed model is capable of computing accurate salient regions by joint upsampling coarse-scale saliency estimates. This approach outputs accurate region silhouettes without relying on pre-segmentation, leading to a significant lower computational burden than most methods in the literature. Besides avoiding pre-segmentation, the key to the efficiency of this approach is that it operates mostly at coarse-scale, restricting computation of high-resolution data to only when it is absolutely required, i.e., when assigning saliency to regions.

- **A pixel-level saliency estimation function.** The visual saliency of a pixel is modeled as its color distance to a randomized color summary of the image. This approach is computationally efficient, presenting linear complexity, and can lead to more accurate saliency maps than previous similar approaches (i.e., pixel-level based on a color summary or random sampling) when employed within the proposed joint upsampling model. Moreover, this estimation approach can be trivially extended with location-based cues, such as a boundary prior.

- **A patch-level saliency estimation function.** The visual saliency of a patch is modeled as the residual of its reconstruction from a PCA (Principal Component Analysis) basis extracted from patches at the image boundaries. In contrast to previous work on PCA reconstruction for saliency detection, this estimate is employed within the joint upsampling model, such that additional processing stages previously employed are unnecessary. This leads to a modest decrease in accuracy compared to previous work, but a massive increase in performance.

1.5 Publications

The following publications resulted from the work presented in this thesis:

Journal papers

- LIE, M. M. I.; BORBA, G. B.; VIEIRA NETO, H.; GAMBA, H. R. Joint Upsampling Random Color Distance Maps for Fast Salient Region Detection. *Pattern Recognition Letters*, Elsevier, 2017. In press.

Conference proceedings papers

- LIE, M. M. I.; BORBA, G. B.; VIEIRA NETO, H.; GAMBA, H. R. Fast Saliency Detection Using Sparse Random Color Samples and Joint Upsampling. In: *Proceedings of the Conference on Graphics, Patterns and Images*. São José dos Campos, SP, Brazil: IEEE, 2016. p. 217–224.

- LIE, M. M. I.; VIEIRA NETO, H.; BORBA, G. B.; GAMBA, H. R. Automatic Image Thumbnailing Based on Fast Visual Saliency Detection. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web*. Teresina, PI, Brazil: ACM, 2016. p. 203–206.

- LIE, M. M. I.; VIEIRA NETO, H.; BORBA, G. B.; GAMBA, H. R. Progressive Saliency-Oriented Object Localization Based on Interlaced Random Color Distance Maps. In: *Proceedings of the Latin American Symposium on Robotics*. Curitiba, PR, Brazil: IEEE, 2017. p. 1–6.

2 Background

This chapter introduces the main subject of this work, namely visual attention and its computational modeling (Section 2.1). Additionally, an overview of low-dimensional image representation is presented (Section 2.2) to substantiate one of the main assumptions in the strategy proposed in this thesis, that saliency detection can be efficiently and accurately computed from a small subset of pixels from the original image. The chapter closes with a description of two techniques used in the implementation of the proposed strategy, Joint Upsampling (Section 2.3), and Principal Component Analysis (Section 2.4).

2.1 Visual attention

2.1.1 Theoretical background

While historical accounts (ITTI; REES; TSOTSOS, 2005) trace the concept of visual attention back to as far as Descartes (1649), a more relevant and illustrative starting point is the experimental work by Yarbus (1967). In his experiment, subjects were presented with a painting — *An Unexpected Visitor* by Repin (1884), shown in Figure 3. Meanwhile, an eye-tracking device recorded their eye movements, allowing posterior inspection of viewing patterns across the scene. Under the described experimental setup,

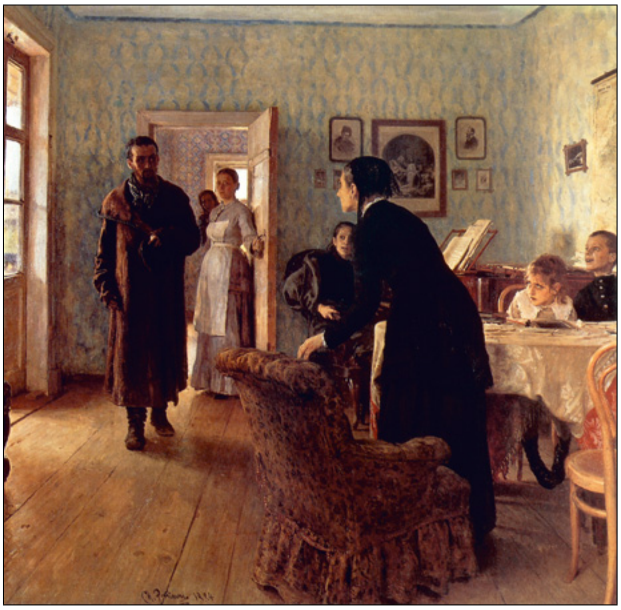

Figure 3: *An Unexpected Visitor*, painting by Repin (1884). This painting was used as visual input in the eye movement experiments by Yarbus (1967).

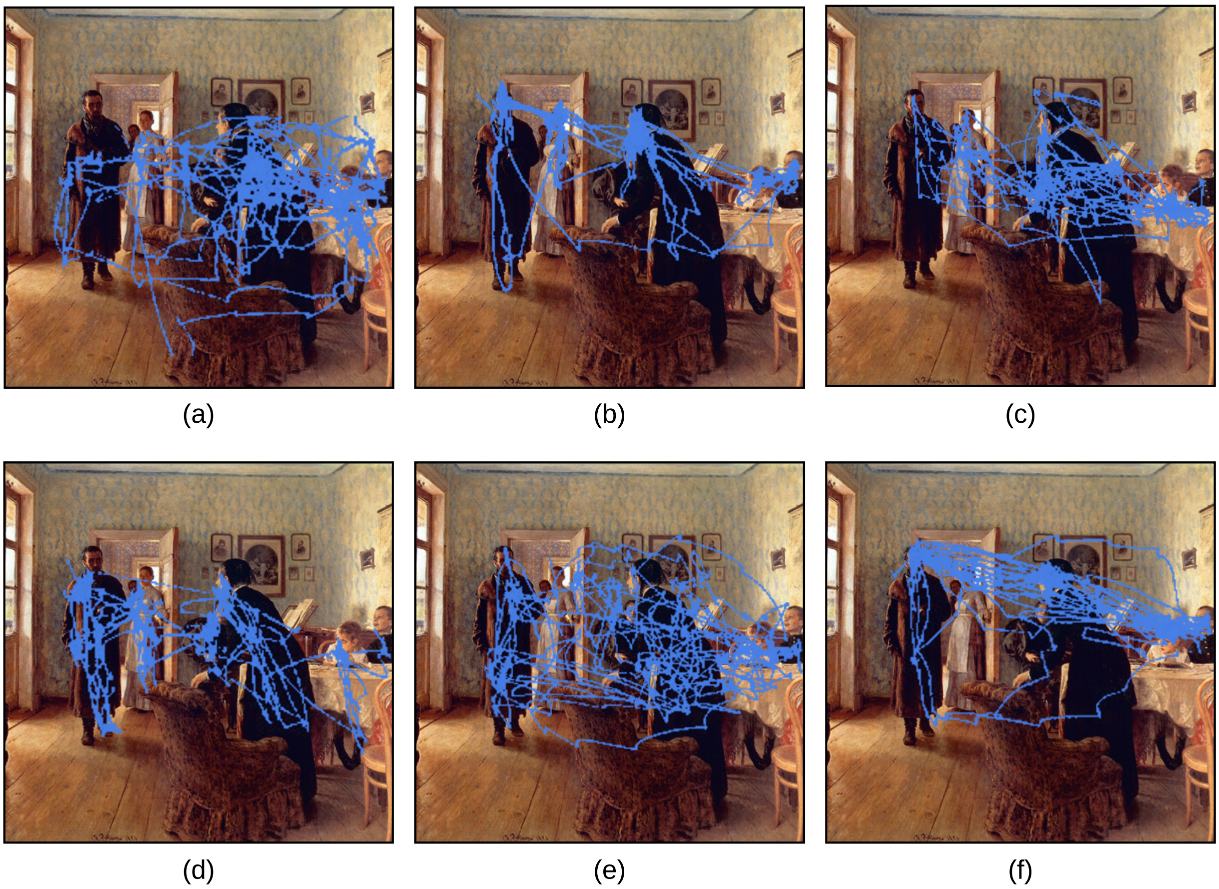

Figure 4: Eye movement experiment by Yarbus (1967). Each figure corresponds to eye movements recorded from subjects instructed to perform the following tasks: **(a)** estimate the material circumstances of the family, **(b)** estimate the age of the people, **(c)** estimate what the family was doing before the arrival of the "unexpected visitor", **(d)** remember the clothing worn by the people, **(e)** remember the position of the people and objects in the room, **(f)** estimate how long the "unexpected visitor" had been away from the family. Each record corresponds to three minutes. The eye movement records were manually superimposed for illustration purposes and might not be perfectly accurate.

the subjects were instructed to perform several cognitive tasks, such as estimating the material circumstances of the family, estimating the age of the people, and remembering the position of the people and objects in the room. The eye movement recordings revealed strikingly distinctly eye movement patterns for each task. These patterns are shown in Figure 4 (a–f) for the six different tasks of the experiment. The results indicate that the visual field is not sampled passively or arbitrarily — *eye gaze is actively allocated to potentially useful information*. This allocation of cognitive processing resources to a small portion of visual stimuli has been attributed to the mechanism of *visual attention*.

Visual attention has been intensively investigated — its aspects have been attributed to specific brain regions (FINK et al., 1996; ITTI; KOCH, 2001), and several psychological theories have been proposed to model its operation (for a computationally oriented review, see Tsotsos (2011)). This mechanism performs information reduction, and is believed to be largely responsible for preventing an overload of cognitive processing in the brain. Two types of selection comprise the operation of visual attention. The first is reflexive and stimulus-driven, or *bottom-up*, the second is voluntary and task-driven, or *top-down* — there are indications that both interact in a non-trivial manner (CONNOR; EGETH; YANTIS, 2004). At this point, it is important to clarify the difference between *eye gaze*, *visual attention* and *visual saliency*. *Eye gaze* is a coordinated motion of the eyes and head, and is largely guided by process of *visual attention* (BORJI; ITTI, 2013). This process, in turn, is comprised of *top-down* and *bottom-up* factors, the latter of which is believed to be driven by distinctiveness of low-level visual features, that is, *visual saliency* (NOTHDURFT, 2000; THEEUWES, 2010).

Similarly to the human visual system, computer vision systems possess limited processing capacity, leading to a substantial interest in the computational modeling of visual attention. Despite some moderate success (e.g., Oliva et al. (2003)), modeling top-down attention has proved elusive for practical computer vision systems (ITTI; KOCH, 2001). This is a consequence of its dependence on prior knowledge and expectations, which are not always available, often not transferable between different tasks, and lead to more complex mental processes (FRINTROP, 2006). Bottom-up attention, on the other hand, lends itself to simple modeling and is generally applicable, since it is based exclusively on low-level scene features. For this reason it is arguably one of the major subjects in current computational visual attention research. This thesis is concerned exclusively with bottom-up visual attention.

2.1.2 Computational modeling

Bottom-up visual attention is computed from visual saliency. In other words, a scene location is more likely to attract involuntary attention if it distinguishes itself from its surroundings in terms of attributes such as orientation, intensity, and color. This approach was popularized by Itti, Koch and Niebur (1998), which implemented and demonstrated the efficacy of the biologically-plausible model by Koch and Ullman (1985). Their implementation selects candidate regions to attend based on center-surround saliency in multiple scales, and demonstrated remarkable predictive power with respect to human eye movements for images of real-world scenes. Due to this suc-

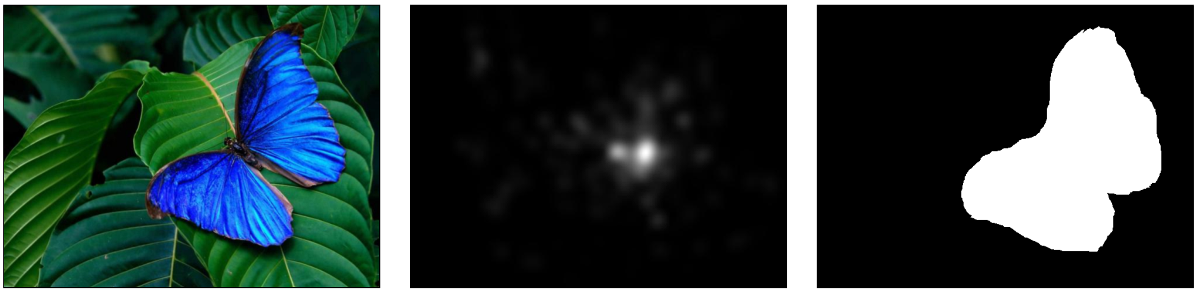

Figure 5: Comparison of different types of saliency maps. **Left:** Input image. **Middle:** Fixation prediction. **Right:** Salient region detection. Fixation prediction highlights locations that are more likely to attract eye gaze, and results in a blurry clusters of points. Salient region detection, on the other hand, entirely highlights regions that are more likely to attract eye gaze. The images are from the Imgsal dataset (LI et al., 2013a).

cess, saliency detection has been widely adopted in computer vision applications, such as image compression (OUERHANI et al., 2001), video quality assessment (ĆULIBRK et al., 2011) and content-based image retrieval (MARQUES et al., 2006), among several others (NGUYEN; ZHAO; YAN, 2018).

It is important to distinguish between the two main types of saliency detection, since they output saliency maps with significantly different aspects, and consequently lead to different algorithm design choices. The first type is *fixation prediction*, which is concerned with computing saliency maps with higher intensity in image locations that human viewers are more likely to fixate at. Since fixations occur at relatively precise locations, saliency maps for fixation prediction highlight sparse clusters, which are usually blurred since it was shown that this improves prediction accuracy in most cases (HOU; HAREL; KOCH, 2012). The second type is *salient region detection*, which is concerned with computing saliency maps that highlight salient objects or regions in the scene entirely. In contrast to the blurry clusters of points in fixation prediction, this type of saliency map highlights entire regions, which are usually computed from segmentation algorithms. Figure 5 presents an example of each type of saliency map for comparison. Since salient region detection is more predominant in computer vision applications, it is the type of saliency map with which this thesis is concerned.

Most salient region detectors are roughly based on two stages (PERAZZI et al., 2012): image abstraction and saliency assignment. The former decomposes the image into perceptually homogeneous regions, to reduce the number of visual elements and discard unnecessary details, while the latter assigns a saliency value to each region, often based on its visual feature uniqueness. While there are methods that do not perform image abstraction and assign saliency at a pixel level (e.g., Achanta et al. (2009)), this approach has lost favor, since it does not scale as well as region-based methods (CHENG et al., 2015). There are, however, methods that employ pixel-level estimation in combination with region-level estimation (e.g., Li et al. (2013b)).

2.2 LOW-DIMENSIONAL IMAGE REPRESENTATION

One might be tempted to assume that more visual data leads to more accurate image analysis. This is not necessarily true. When it comes to visual perception, typical scenes present *highly redundant data* and *spatially-varying importance* — both aspects can enable substantial reduction in visual data dimensionality. *Redundancy* implies that a part of the information can be ignored without impact on the accuracy of visual tasks. *Varying importance* implies that, from the information that does affect accuracy, a subset can be chosen such that ignoring it decreases accuracy the least. These characteristics can be leveraged to reduce the dimensionality of image content, and consequently improve the efficiency of computer vision processes.

2.2.1 Redundancy in natural images

The role of redundancy on visual perception has been discussed at least since Attneave (1954), and demonstrated for natural images in an experiment by Kersten (1987), in which the redundancy of missing pixels was estimated based on the ability of human subjects to predict each of their values. Redundancy is not restricted to pixel-level — Zontak and Irani (2011) analyzed the redundancy of patches in single natural images, and showed that patches tend to reoccur in close proximity of each other, with probability decaying rapidly with distance from the patch. In fact, redundancy also occurs in scale, as demonstrated in the example-based image super-resolution method by Glasner, Bagon and Irani (2009), which dispenses with external datasets, relying only on patches extracted from the same image at several scales. The redundancy of natural images over scale is such that, combined with the tolerance that the human visual system presents to degradations in image resolution, it has been reported that humans need only 32×32 pixels to achieve an 80% recognition rate on scene recognition tasks (TORRALBA; FERGUS; FREEMAN, 2008). Figure 6 (a) shows an example of patch redundancy in space and scale.

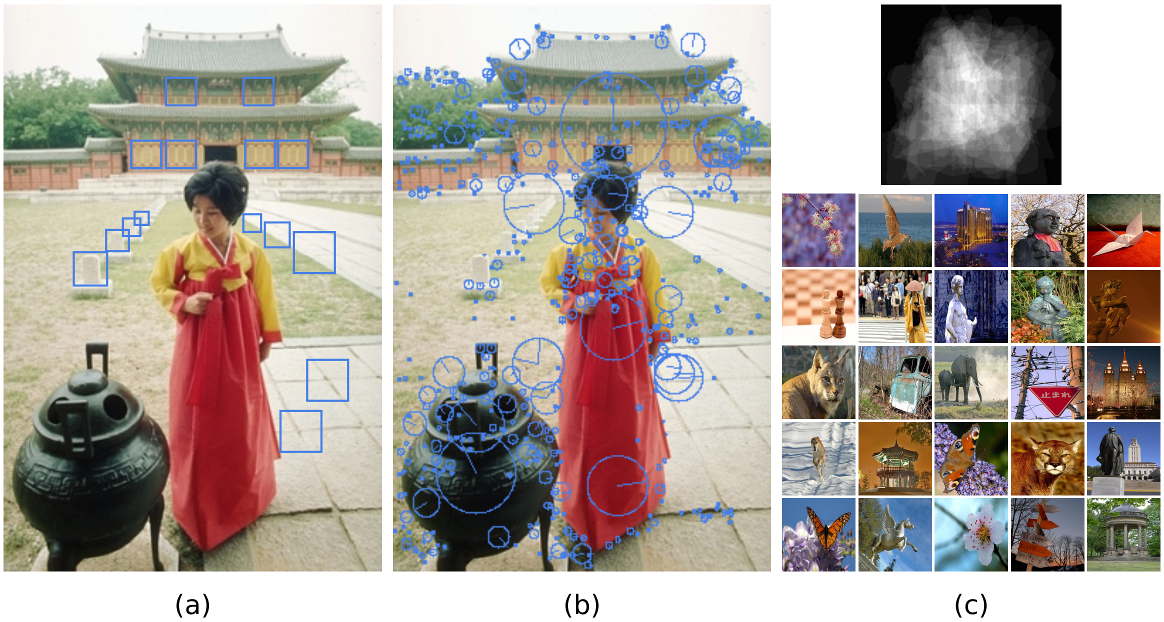

Figure 6: Examples of redundancy and spatially-varying importance in images. **(a)** Patch recurrence on space and scale. **(b)** Keypoints using the SIFT method. The circle indicates the scale at which the keypoint was detected and the line is its dominant gradient orientation. **(c)** Average saliency map of 25 images randomly sampled from the MSRA10K dataset (CHENG et al., 2015) — there is a clear bias towards the center. The ground truths of the dataset were used as saliency maps and were resized to the same size prior to averaging.

2.2.2 Spatially-varying importance of image content

The spatially-varying importance of image content was discussed by Brady (1987) from the perspective of constraints to visual processes. In this interpretation, corners and other curvature maxima were called *"seeds of perception"*, i.e., locations that provide more reliable parameterization of visual processes by imposing tighter constraints. This concept is arguably the origin of what are currently known as *"interest points"* or *"keypoints"* (SCHMID; MOHR; BAUCKHAGE, 2000), although current methods are more sophisticated than simple corner detectors, and are designed to detect points or blobs that present robustness to photometric and geometric transformations (TUYTELAARS; MIKOLAJCZYK, 2008). Keypoint-based image representation is used, for instance, in the bag-of-visual-words approach for image classification (CSURKA et al., 2004), which encodes images as sets of patches extracted from keypoints. Despite being very influential, this approach was subject to a later study by Nowak, Jurie and Triggs (2006), which showed that random sampling is more effective than keypoint detection for this particular application. Figure 6 (b) shows an example of keypoint detection using the SIFT (Scale Invariant Feature Transform) method (LOWE, 2004).

In the context of visual attention, while keypoint detection has been employed for fixation prediction (e.g., Oliveira, Rocha Neto and Gomes (2016)), general location cues are more common, the most notable being *center prior*, i.e., the center of the image is more likely to be salient, and *boundary prior*, i.e., the image boundaries are more likely to be background. While these two may seem equivalent, boundary prior is more general, since salient objects can appear off the center, for instance due to the one-third composition principle from photography, and still not overlap with the boundary (WEI et al., 2012). Figure 6 (c) shows the average saliency map of 25 images (mostly photographs) sampled randomly from the MSRA10K dataset (CHENG et al., 2015), demonstrating center-bias. It is worth noting that center-bias is not a property of visual content itself, but a consequence of the framing imposed during image acquisition, for instance, by the photographer.

2.3 JOINT UPSAMPLING

The Gaussian low-pass filter computes a weighted average of the values inside its support, such that the weights decay with distance from its center. It is one of the most commonly employed spatial filters in image processing. The visual effect of Gaussian filtering is ″smoothing“, for this reason it is also employed in computational photography for removal of small details (e.g., for abstraction and denoising). Decomposing an image into successively less detailed layers by this approach is known as a Gaussian pyramid decomposition (ADELSON et al., 1984), which is perhaps the most common multi-scale representation technique in the literature.

For some applications, blurring is required to reduce the amount of small details in the image, but edges need to be preserved to avoid distorting shape information. In these cases, traditional Gaussian filtering is not enough — since the filter operates only in space, it cannot account for edges. A solution is to filter in both space and range, in other words, compute based not only on *geometric closeness* but also on *photometric similarity* between the pixels (TOMASI; MANDUCHI, 1998). In this manner, the output of the filter is attenuated not only with distance from the center of the support but also in the vicinity of edges. Filters that operate with this approach are known as *edge-preserving smoothing* (EPS) filters. Figure 7 presents an example comparing the operation of traditional and edge-preserving smoothing.

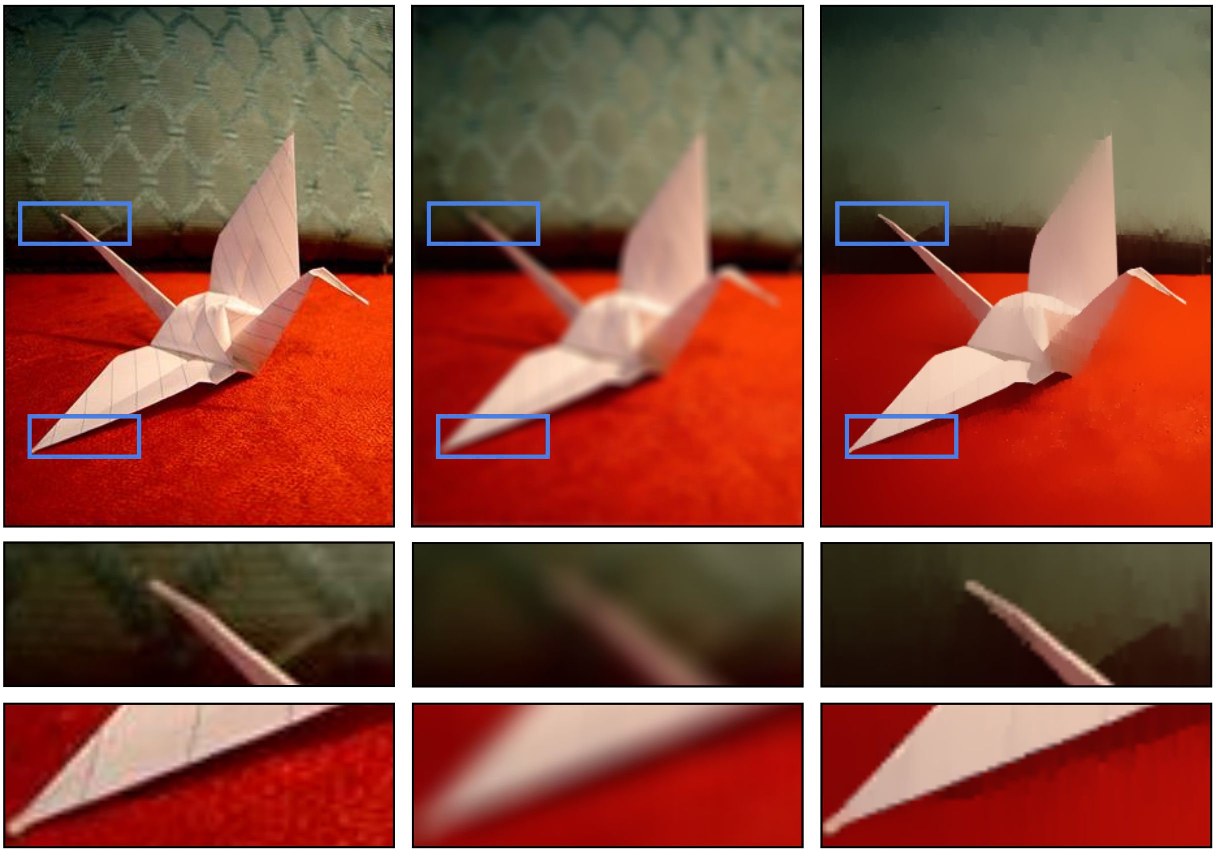

Figure 7: Comparison of smoothing filters. **Left:** Input image. **Middle:** Traditional Gaussian smoothing. **Right:** Edge-preserving smoothing. The regions enclosed by blue rectangles are displayed zoomed-in for more detailed comparison. The original image presents details such as the texture of the fabric and the pattern on the background. Gaussian filtering is capable of removing most of the details, but blurs the edges in the process. Edge-preserving smoothing is capable of removing a similar amount of detail while retaining sharp edges.

In this work, edge-preserving smoothing is performed using the *Fast Global Smoother* (FGS) proposed by Min et al. (2014), motivated by its computational performance and relatively easy parameterization. The FGS formulates edge-preserving smoothing as a solution to the following 1D minimization problem for each row and column in the image:

$$J(u) = \sum_{n} \left((u_n - f_n)^2 + \lambda \sum_{i \in \mathcal{N}(n)} w_{n,i}(g)(u_n - u_i)^2 \right), \tag{1}$$

where f, g and u correspond to rows or columns of the *input*, *guide* and *output* images, respectively. While the f provides the data to be smoothed, g defines the edges within this content is to be smoothed. Equation 1 is defined for $n \in [1..L]$, where L is the width of f when solving for rows, and the height of f for columns. $\mathcal{N}$ is the pair of neighbor pixels of n, λ is a parameter defining the *smoothness* of the output, and $w_{n,i}(g)$ is a function that defines the *similarity* of the pixels n and i in the image g:

$$w_{n,i}\left(g\right) = exp\left(\frac{-||g_n - g_i||}{\sigma_c}\right), \tag{2}$$

where σ_c is the *range parameter*. The parameters values were set empirically as $\sigma_c = 0.03$ and $\lambda = 100$.

Filtering in both space and range provides an interesting possibility: smoothing content from one image within edges from another image (i.e., $f \neq g$ in Equation 1). The main appeal of this approach is that the content to be smoothed can be computed at a low resolution, while the edges may come from a full-resolution image. Prior to filtering, the low-resolution content in f must be resized to match the full-resolution size of g, which can be efficiently done through nearest-neighbor interpolation. This provides a computationally efficient approach for achieving full-resolution output despite restricting more costly processing operations to low resolution. This approach is called *joint upsampling*, and its effectiveness was demonstrated for tasks such as colorization, tone-mapping, and depth from stereo (KOPF et al., 2007).

Joint upsampling is a central component of the strategy proposed in this thesis, and is employed to efficiently achieve accurate, near full-resolution, saliency maps from coarse-scale estimates. As will be shown later, employing joint upsampling allows leveraging full-resolution edge information to achieve high accuracy in a more computationally efficient manner than the segmentation-based approach of most state-of-the-art methods, while leading to comparable and very competitive results.

2.4 Principal component analysis

Principal Component Analysis (PCA) is a statistical procedure commonly used for dimensionality reduction and exploratory data analysis (ABDI; WILLIAMS, 2010). It projects the data onto a basis (i.e., principal components) in which the variance is maximized. Let a matrix **M** describe a dataset, such that each row represents a data instance and each column describes a coordinate in feature space (i.e., data attribute, variable). The PCA of **M** can be performed by eigendecomposition of its covariance matrix. More precisely, the covariance matrix of **M** is computed as:

$$\mathbf{C} = \mathbf{M}\mathbf{M}^{\mathsf{T}}, \tag{3}$$

and then subject to eigendecomposition:

$$\mathbf{C}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}, \tag{4}$$

where **U** is a matrix containing the eigenvectors in its columns, and **Λ** is a diagonal matrix containing the eigenvalues associated with the eigenvectors in **U**. In the context of PCA, the columns of **U** are the *principal components* and the elements in the diagonal of **Λ** are their corresponding variances. The latter is often used to choose which principal components to retain, e.g., retain the first few that account for 95% of the total variance.

Besides dimensionality reduction, PCA can also be directly employed for pattern recognition. For instance, a principal component model can be computed for each class, such that data instances are classified according to how well they fit each class (WOLD, 1976). This approach has been effectively employed in applications such as face recognition (TURK; PENTLAND, 1991), novelty detection (VIEIRA NETO; NEHMZOW, 2007), object detection (MALAGÓN-BORJA; FUENTES, 2009; RAZAKARIVONY; JURIE, 2013), and pedestrian detection (CARVALHO et al., 2011). In this thesis, similarly to previous work (MARGOLIN; TAL; ZELNIK-MANOR, 2013; LI et al., 2013b), PCA is used to estimate the visual saliency of image regions in a simple and computationally versatile manner, albeit more efficiently.

3 Related Work

This chapter reviews some of the most relevant work related to the contributions of this thesis. Section 3.1 introduces the idea of estimating the visual saliency of a pixel as its color dissimilarity to the rest of the scene, and the efforts made to improve the efficiency of this approach. While effective on relatively simple scenes, pixel-level estimation presents some known limitations regarding scalability. Considering this, an overview of region-level saliency estimation methods is also presented, with emphasis on subspace-based methods, which is the region-level approach explored in this thesis. Section 3.2 summarizes the theoretical motivation for coarse-scale saliency estimation, including empirical results on human subjects, and how these have been employed in computational visual attention models.

3.1 Visual saliency modeling

3.1.1 Pixel-level estimation

When discussing visual saliency in digital images, it is reasonable to start with the model of saliency for a single pixel. Since perception of color difference is closely related to saliency, it has often been the central aspect in most saliency detection models, for instance, Zhai and Shah (2006) defined the saliency of a pixel p in terms of color distances as:

$$s(p) = \sum_{p_i \in \boldsymbol{\Omega}} \|f(p) - f(p_i)\|, \tag{5}$$

where f is the input image and $\boldsymbol{\Omega}$ is the set of all image locations. In other words, *saliency is defined as the accumulated color distance with respect to the entire image.* This distance is often computed in the CIELAB color space, due to its perceptual uniformity, i.e. in this space the Euclidean distance is approximately linear with respect to human visual perception (REINHARD et al., 2008). While reasonable and straightforward, this is approach is computationally inefficient — its computational complexity is $O(n^2)$ for an image with n pixels. If there are more pixels than colors in an image, performance can be improved by computing the saliency of each color instead of each pixel, since assigning precomputed saliency to each pixel can then be done with linear complexity. However, this is often not the case. For instance, even a 1920×1080 true-color image has approximately 2 million pixels, but more than 16 million possible colors. On the other hand, limiting this approach to luminance can lead to very fast computation, since it can be encoded in a single channel, spanning 256 values at most — which is much smaller than the number of pixels in most images.

Despite being efficient, restricting the model to luminance information compromises effectiveness in a non-negligible manner, encouraging alternative strategies to alleviate computational burden without sacrificing color information. For instance, Achanta et al. (2009) proposed adopting a color summary — in their approach *saliency is defined as the color distance to the average color of the image*. Since a single color difference is computed for each pixel, this approach executes in $O(n)$. However, despite performing well in simple datasets, the correlation between saliency and distance to the average color of the image does not scale to more complex datasets (YILDIRIM, 2015). Cheng et al. (2015) proposed returning to the strategy by Zhai and Shah (2006), but reducing the number of colors (12 colors per channel) using histogram-based quantization and ignoring rare colors. This enabled retaining color information while achieving computation time comparable to using only luminance information, but with a substantial increase in accuracy. For an image with n pixels and k colors, this approach executes in $O(k^2) + O(n) \approx O(n)$, assuming $n > k$.

Another relevant approach is the random sampling strategy by Vikram, Tscherepanow and Wrede (2012) that, motivated by the random scattering of receptive fields in the human visual system, estimates the saliency of a pixel as its color distance to the average color of the randomly generated windows (i.e., in terms of location and size) that contain it. It can be argued that this approach is a local variant of the approach by Achanta et al. (2009), despite presenting a significantly longer execution time, mainly due to the number of windows randomly generated ($0.2 \cdot n$ for an image with n pixels) and post-processing based on mean filtering. In this thesis, it will be shown that a simpler randomized color summary can be employed under the proposed strategy, resulting in superior trade-off between accuracy and execution time.

3.1.2 Region-level estimation

Saliency is not exclusively estimated at a pixel-level. In fact, recent models are mostly based on region-level estimation, with regions computed from segmentation algorithms such as *Mean shift* (COMANICIU; MEER, 2002) and *SLIC (Simple Linear Iterative Clustering) superpixels* (ACHANTA et al., 2012). There are, however, methods that do not employ segmentation, and simply adopt regular patches as regions (e.g., Parikh, Zitnick and Chen (2008), Borji and Itti (2012)), which is much more efficient, despite not being as accurate. Computing saliency from regions instead of individual pixels allows extraction of more informative features, as well as efficiency improvement, since there are significantly less regions than pixels in an image (BORJI et al., 2014).

While there are many frameworks in which to model region-level saliency, usually graph-based (e.g., Gopalakrishnan, Hu and Rajan (2009), Yang et al. (2013), Jiang et al. (2013a)), here emphasis is given to subspace-based methods, particularly those using Principal Component Analysis (PCA), which has been shown to be a popular and efficient approach to reveal the internal structure of data (MARGOLIN; TAL; ZELNIK-MANOR, 2013). Several studies have investigated the idea of projecting visual data onto a latent subspace prior to visual saliency estimation. Rajashekar, Cormack and Bovik (2003) computed the PCA of patches around fixations from eye-tracking data, comprised of viewing patterns from six human subjects on approximately 100 images, containing both natural and man-made scenes. The principal components of these patches were employed as low-level features for saliency computation — convolving images with as few as four principal components as filter kernels resulted in promising saliency maps for fixation prediction. Borji and Itti (2012) computed patch saliency as a combination of local and global dissimilarity, in which the input patches were first projected onto a dictionary learned from 1500 images of natural scenes.

Some approaches do not rely on external images. For instance, the method by Duan et al. (2011) projects the patches from the input image onto a basis computed from the input image itself. In this approach, patches are considered as vectors for dissimilarity computation, and PCA is employed simply as a dimensionality reduction and denoising mechanism. In contrast to this approach, Margolin, Tal and Zelnik-Manor (2013) proposed a strategy that resembles that by Achanta et al. (2009) (i.e., saliency as dissimilarity to a visual feature summary of the image), but for patches on the PCA subspace. Based on the observation that salient patches appear scattered in the subspace spanned by the principal components, the saliency of a patch is estimated as its dissimilarity to the average patch of the image in this subspace.

Another interesting strategy comes from a classification perspective (MALAGÓN-BORJA; FUENTES, 2009), in which a PCA subspace is computed for each class. Classification of new instances is done by projection into each subspace and assignment to the one that results in the smallest reconstruction error. This approach was explored in the saliency detection method by Li et al. (2013b), which assumes that image segments belong to one of two classes: background or salient region. In this manner, the saliency of a segment is estimated as its dissimilarity to segments in the image boundaries, which is computed as the reconstruction error from a basis extracted from boundary segments. However, this approach computes segments from SLIC superpixels (ACHANTA et al., 2012), which combined with this dissimilarity formulation leads to heterogeneous saliency maps on certain types of images. To address this issue, the authors combined several post-processing stages, including estimation of sparse reconstruction error, a propagation mechanism based on K-means, and multi-scale combination (LI et al., 2013b). As the thesis results will show, under the proposed joint upsampling strategy, PCA reconstruction error of simple regular patches can be employed for accurate region-level saliency detection without need for pre-segmentation or costly post-processing steps, resulting in high accuracy with very short execution time.

3.2 COARSE-SCALE SALIENCY ESTIMATION

Estimating saliency in coarse scale (i.e., low-resolution) is advantageous from mainly two perspectives. First, there is a substantial decrease in processing data, which leads to a very significant reduction in execution time. Second, there is evidence that attention is much more coarse-grained than visual resolution, suggesting that coarse-scale analysis might also be the more theoretically adequate (INTRILIGATOR; CAVANAGH, 2001). Regarding this perspective, Judd (2011) presented extensive experiments quantifying the extent to which human fixations are affected by reduction in image resolution. For a dataset of 168 natural images and 25 pink noise images, her results showed not only that fixations in low-resolution can predict fixations in high-resolution, but that 85% of accuracy can be achieved from a resolution as small as

64×64 pixels. The strategy presented in this thesis is largely motivated by these results.

While there are fixation prediction methods that operate at a coarse-scale (e.g., Hou and Zhang (2007), Harel, Koch and Perona (2007), Seo and Milanfar (2009)), this approach is mostly unexplored in salient region detection. A possible reason for this is that, in contrast to fixation prediction, salient region detection requires accuracy at the level of object contours, and consequently must rely on fine-scale processing. As explained previously, instead of downscaling the image, salient region detectors often adopt segmentation to reduce the number of image elements and alleviate computational burden. It will be demonstrated later in this thesis that it is possible to efficiently leverage the advantages of low-resolution saliency estimation and still achieve accurate results, without resorting to segmentation algorithms.

4 An Efficient Strategy for Estimation of Visually Salient Regions in Images

This chapter presents the main contribution of the thesis, an efficient salient region detection strategy based on joint upsampling of coarse-scale saliency estimates. The problem is mathematically stated to establish the notation (Section 4.1), and the proposed approach is introduced through a general overview, which describes its steps and rationale (Section 4.2). Then, a pixel-level saliency formulation based on randomized color distances is presented for coarse-scale estimation within the proposed strategy (Section 4.3). While employing this formulation with the proposed strategy improves over previous pixel-level approaches and is adequate for relatively simple scenes, it suffers from scalability limitations common to all pixel-level approaches. Considering this, a second saliency formulation is presented, which overcomes this limitation by operating at a patch level, based on PCA reconstruction errors (Section 4.4). The chapter closes with a description of implementation details, including choices of algorithms and minor parameterization choices (Section 5.1.4).

4.1 Problem statement

Let the function $f\colon \mathbf{\Omega} \rightarrow \mathbb{R}^3$, with domain $\mathbf{\Omega} = \{(x, y) \subset \mathbb{Z}^2 \,|\, 0 \leq x < W,\ 0 \leq y < H\}$, define a color image with dimensions $W \times H$. For each element in the domain $\mathbf{\Omega}$, f defines a tuple in $\mathbb{R}^3$ encoding its value in an arbitrary color space. The problem consists in defining a function $s \circ f\colon \mathbf{\Omega} \rightarrow \mathbb{R}$, which maps each element $(x, y) \in \mathbf{\Omega}$ to a real value describing its perceptual dissimilarity to a subset $\mathbf{\Omega_s} \subseteq \mathbf{\Omega}$, based on their color values as defined by f. In other words, given a color image f, s describes the perceptual dissimilarity of each of its elements with respect to a subset of the same image. The function s defines a *saliency map*, and its construction, which is the main subject of this work, is defined as the *saliency detection problem*.

4.2 Overview

The proposed strategy consists in two stages: coarse-scale saliency estimation and joint upsampling. Operating at a coarse scale allows a drastic reduction in the amount of processed data. Additionally, due to the amount of redundancy in real-world images (Section 2.2 – *Low Dimensional Image Representation*) and the scale of human visual attention (Section 3.2 – *Coarse-scale Saliency Estimation*), it can be argued that it is also the most theoretically adequate scale in which to process visual attention. While it is known that low-resolution visual data is highly informative, capable of providing enough information for tasks as complex as object and scene recognition (TORRALBA; FERGUS; FREEMAN, 2008), the fact that salient region detection requires accurate object contours remains.

Since the predominant approach for salient region detection consists in assigning saliency to segments, achieving saliency maps with accurate object contours is not an issue, given that segmentation decomposes the image into regions with boundaries that coincide with edges of the full-resolution input. In contrast to previous approaches, the proposed strategy avoids segmentation in favor of edge-preserving smoothing, which is significantly more efficient and can be employed for joint upsampling of coarse-scale estimates. In this manner, it is possible to leverage the advantages of both coarse-scale saliency estimation (i.e., reduction in computational cost, agreement with experimental evidence, abstraction of unnecessary detail) and fine-scale edge information (i.e., high accuracy). Since fine-scale information is used only when absolutely required, most of the computation is performed in coarse scale, leading to higher computational efficiency.

To illustrate the difference in efficiency between edge-preserving smoothing and oversegmentation, Table 1 presents the average execution time of the Fast Global Smoother (FGS) (MIN et al., 2014) — the algorithm employed for joint upsampling in the proposed approach — and the three most common algorithms employed for oversegmentation in salient region detection, considering the methods in the benchmark by Borji et al. (2015): *Mean shift* (COMANICIU; MEER, 2002), *Efficient Graph-Based Segmentation* (EGBS) (FELZENSZWALB; HUTTENLOCHER, 2004), *Simple Linear Iterative Clustering* (SLIC) (ACHANTA et al., 2012). FGS performs in less than half of the execution time of SLIC, which is the fastest among the compared segmentation algorithms.

In the next sections, two saliency formulations are proposed for coarse-scale estimation in the proposed strategy. The first one operates at a pixel level, based on distances to a randomized color summary of the image. It will be shown that aspects

Table 1: Execution time of image abstraction algorithms. FGS has the shortest execution time, less than half of the time taken by the second fastest algorithm (SLIC). The algorithms were executed using their default parameters, using an input image with 400×300 pixels, on an Intel Core i7-860 2.80 GHz CPU with 4 GB RAM.

Method	Mean shift	EGBS	SLIC	FGS
Execution time (s)	0.90	0.13	0.11	0.04

of the saliency map that might be degraded by estimating on a smaller amount of data are efficiently compensated by joint upsampling. This formulation presents improvements over previous similar approaches, but cannot overcome the scalability limitations of exclusively pixel-level methods. Considering this, the second formulation estimates saliency at a region-level, based on patch reconstruction error when projected on a PCA basis computed from the image boundaries. This approach enables the aforementioned advantages of the proposed strategy, while scaling much better than its pixel-level counterpart.

4.3 Randomized color distance maps

Similarly to the work by Achanta et al. (2009), and following the *color uniqueness hypothesis*, the pixel-level saliency formulation proposed in this thesis defines the saliency of a pixel as the distance of its color to a color summary of the image. However, instead of the average color of the image, a randomized subset of the image is adopted as summary instead. In this manner, the saliency of a pixel p is defined as:

$$s(p) = \sum_{p_i \in \mathbf{\Omega_s}} \|f(p) - f(p_i)\|, \tag{6}$$

where f is the input image and $\mathbf{\Omega_s}$ is a random subset of all image locations. The subset $\mathbf{\Omega_s}$ is resampled each time Equation 6 is computed, and, similarly to previous models (e.g., Achanta et al. (2009), Vikram, Tscherepanow and Wrede (2012), Cheng et al. (2015)), f is assumed to have been converted from RGB to the CIELAB color space before saliency estimation, to leverage its perceptual uniformity.

Equation 6 is almost identical to Equation 5, differing only in the adoption of a randomly selected subset $\mathbf{\Omega_s}$ instead of all image locations $\mathbf{\Omega}$. This approach leverages a principle of randomized algorithms, which states that it is possible to estimate fea-

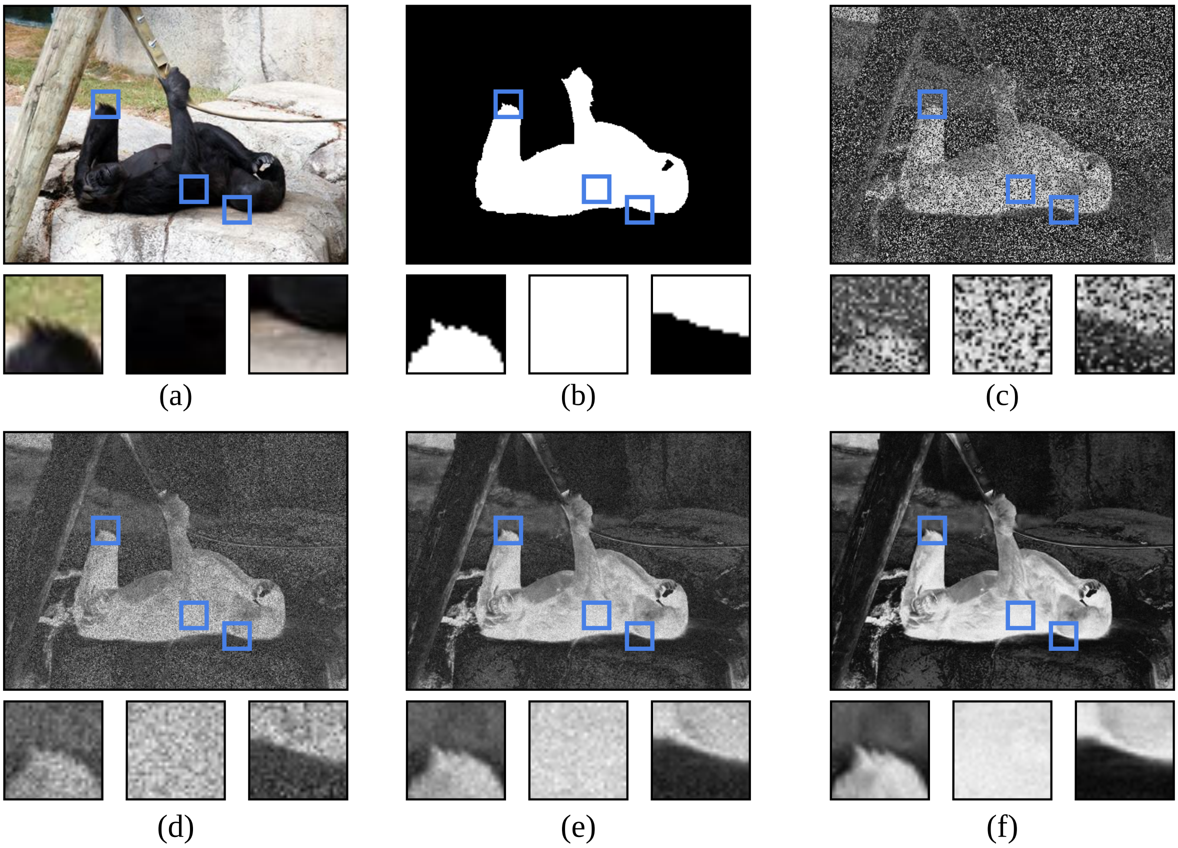

Figure 8: Effect of the subset size $|\mathbf{\Omega_s}|$ on the randomized color distance map. **(a)** Input image. **(b)** Ground truth. **(c)** $|\mathbf{\Omega_s}| = 1$. **(d)** $|\mathbf{\Omega_s}| = 10$. **(e)** $|\mathbf{\Omega_s}| = 100$. **(f)** $|\mathbf{\Omega_s}| = 1000$. The regions enclosed by blue squares are displayed zoomed-in for more detailed comparison. The output presents a noisy aspect, which is attenuated for larger values of $|\mathbf{\Omega_s}|$ — however, the salient region is already emphasized for a subset size as small as a single pixel.

tures of the entire population in a computationally inexpensive manner from a small sample (MOTWANI; RAGHAVAN, 1996). For an image with n pixels, if $\mathbf{\Omega_s}$ is kept small (i.e., $|\mathbf{\Omega_s}| \ll n$), Equation 6 can be computed in $O(|\mathbf{\Omega_s}|\, n) \approx O(n)$. As Figure 8 shows, this can be reasonably expected, since saliency regions are evident even using a subset size as small as $|\mathbf{\Omega_s}| = 1$, despite larger values being desirable to avoid a noisy output. However, increasing $|\mathbf{\Omega_s}|$ is not an efficient approach, since execution time increases in proportion to its value. As will be shown later, keeping $|\mathbf{\Omega_s}|$ small and correcting the output using joint-upsampling is much more efficient. To distinguish the result of Equation 6 from a saliency map, it is called a *randomized color distance map*.

Sampling $\mathbf{\Omega_s}$ randomly from the entire image is already a reasonably effective approach. However, its accuracy can be substantially improved by leveraging the spatially-varying importance of image content. This is straightforward to incorporate into the model. *Boundary prior*, the assumption that image boundaries are more likely

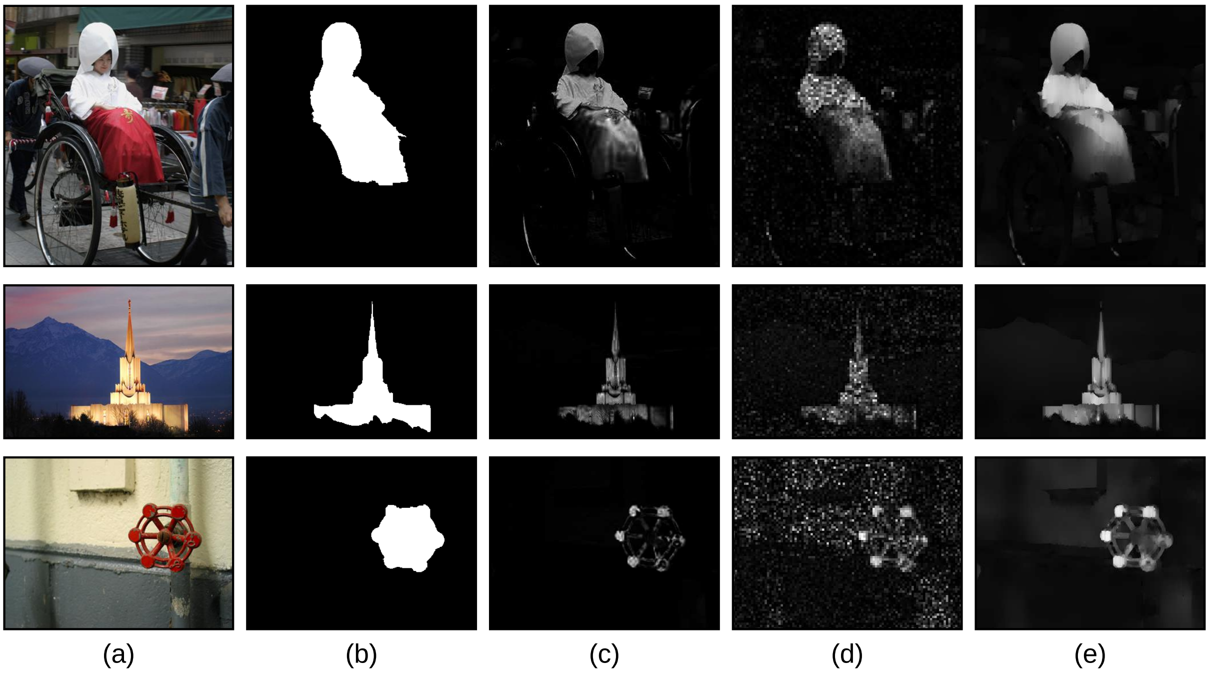

Figure 9: Comparison of randomized color distance maps. **(a)** Input image. **(b)** Ground truth. **(c)** Saliency estimation with $|\mathbf{\Omega_s}| = 1000$, from full-resolution input. **(d)** Saliency estimation with $|\mathbf{\Omega_s}| = 3$, from coarse-scale (20% of full-resolution) input, unfiltered. **(e)** Saliency estimation with $|\mathbf{\Omega_s}| = 3$, from coarse-scale (20% of full-resolution) input, joint-upsampled with the full-resolution input.

to belong to the background, can be incorporated into the model simply by restricting the sampling of $\mathbf{\Omega_s}$ to the image boundaries. In other words, for each pixel location $p_i = (x_i, y_i)$, instead of randomly sampling the coordinates x_i and y_i from the interval $[1 \mathinner{..} L]$, they are randomly sampled from $[1 \mathinner{..} BL] \cup [(L - BL) \mathinner{..} L]$, where L is the image width for x_i and height for y_i, while B is the *boundary ratio*, a parameter that defines the proportion of the image dimensions to adopt as boundary size for the prior. Adopting $B = 0.5$ disregards boundary prior, since it indicates that two opposing boundaries take half of the image each, setting the entire image as boundary.

A comparison of randomized color distance maps is shown in Figure 9. Comparing the results obtained with $|\mathbf{\Omega_s}| = 1000$ on the full-resolution input and with $|\mathbf{\Omega_s}| = 3$ on the input downsampled to 20% of the full-resolution (Figures 9 (c) and (d), respectively) show that the latter contains most of the information of the former, despite presenting a noisy aspect. Joint upsampling the downsampled randomized color distance map leads to a full resolution input, with even more homogeneous salient regions than can be obtained by only increasing $|\mathbf{\Omega_s}|$.

In addition to the higher quality output, the joint upsampling approach is much more efficient. The downsampled randomized color distance map alone was computed in 0.03 seconds (on an Intel Core i7-860 2.80 GHz CPU with 4 GB RAM), while computing it with joint upsampling takes on average 0.06 seconds. On the other hand, adopting $|\mathbf{\Omega_s}| = 1000$ on the full-resolution input took on average 17.29 seconds, and the results are still inferior (e.g., the salient regions output are not as homogeneous). In order to reduce execution time, RGB to CIELAB conversion is performed only on the downsampled image, rather than in the full-resolution image, since it is only needed for perceptually uniform color distance computation. In this manner, joint upsampling is guided by the edges of the RGB input image. While edges in the CIELAB color space might be more perceptually meaningful, performing this conversion on the full-resolution image increases execution time substantially, without any perceptible increase in accuracy.

4.4 Patch reconstruction error from a boundary basis

Similarly to the work by Li et al. (2013b), the region-level saliency formulation proposed in this thesis defines the saliency of a patch as its reconstruction error from a boundary basis. The input image is decomposed into non-overlapping patches, and those that are located at the boundaries are selected to form a basis computed using PCA (Figure 10). Each patch is then reconstructed using this basis, and their saliency is estimated as the resulting reconstruction error, since a large error implies dissimilarity to the boundary, which is equivalent to saliency under the *boundary prior* hypothesis.

More precisely, let f be a color image with $W \times H$ pixels. To account for coarse-scale analysis, f is resized to $L \times L$ such that $L < W, H$. A set of V visual feature channels are extracted from the resized image, namely HSV and CIELAB color channels, since they yielded the most accurate results. Only color features are used, following the *color uniqueness hypothesis*. The resized image is decomposed into non-overlapping $k \times k$ patches, which are unfolded into Vk^2-dimensional vectors, accounting for all feature channels. In this manner, if $\mathbf{P_B}$ is the set of non-overlapping patches extracted from the image boundaries, then its covariance matrix is computed as:

$$\mathbf{C} = \mathbf{P_B} \mathbf{P_B^\top}, \tag{7}$$

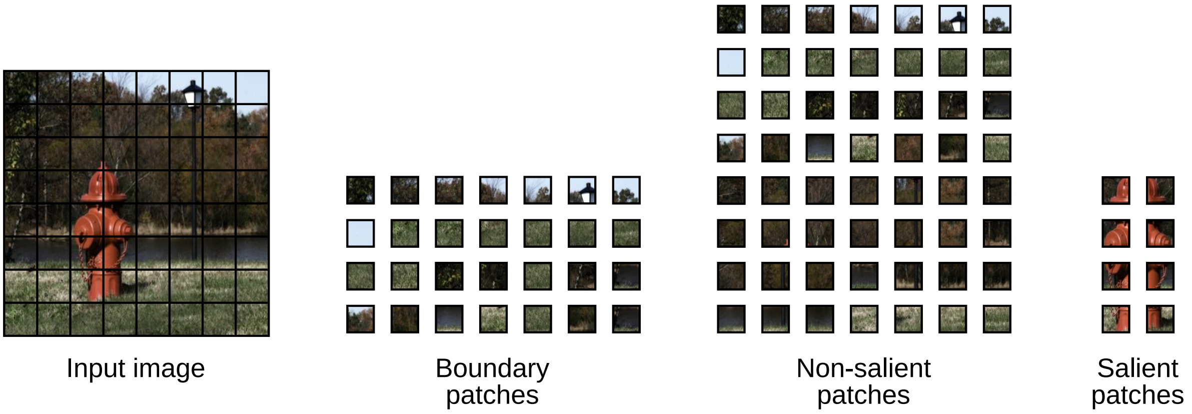

Figure 10: Visual dissimilarity between salient and boundary patches. Boundary patches are generally more similar to non-salient than salient patches. Thus, dissimilarity to boundary patches can be adopted as a cue for saliency estimation. While the set of non-salient patches contains all boundary patches for this simple example, this is not always the case and is not an assumption of the model. Since the formulation is based on PCA reconstruction error, it tolerates a certain amount of overlap between salient and boundary patches.

and subject to eigendecomposition:

$$\mathbf{C}\mathbf{U_B} = \mathbf{U_B}\mathbf{\Lambda_B}, \tag{8}$$

which provides the matrices of eigenvectors $\mathbf{U_B}$ and eigenvalues $\mathbf{\Lambda_B}$, corresponding to a basis (i.e., principal components) of $\mathbf{P_B}$ and the corresponding variances, respectively. The saliency of a patch is then estimated as its reconstruction error, normalized between $[0\,..\,255]$, when computed from this basis:

$$s(\mathbf{p}) = \|\mathbf{p} - \mathbf{U_B}\mathbf{U_B^T}\mathbf{p}\|. \tag{9}$$

Some parameters are fixed due to design choices, and are consequently not included in the assessment in Chapter 5 (*Experimental Results*). The parameters and their values are: the *image size for patch extraction* $L = 64$ and *patch size* $k = 8$. The former is motivated by experiments indicating that it is approximately the smallest resolution at which human fixations become relatively consistent (JUDD; DURAND; TORRALBA, 2010), and technical results indicating its effectiveness in coarse-scale saliency estimation (e.g., Hou and Zhang (2007), Seo and Milanfar (2009)). The latter is motivated by efficiency, since PCA is computationally intensive on large patches.

These fixed parameter values constrain the initial saliency estimation to 8×8, which is too coarse-grained for direct joint upsampling to full-resolution to yield ac-

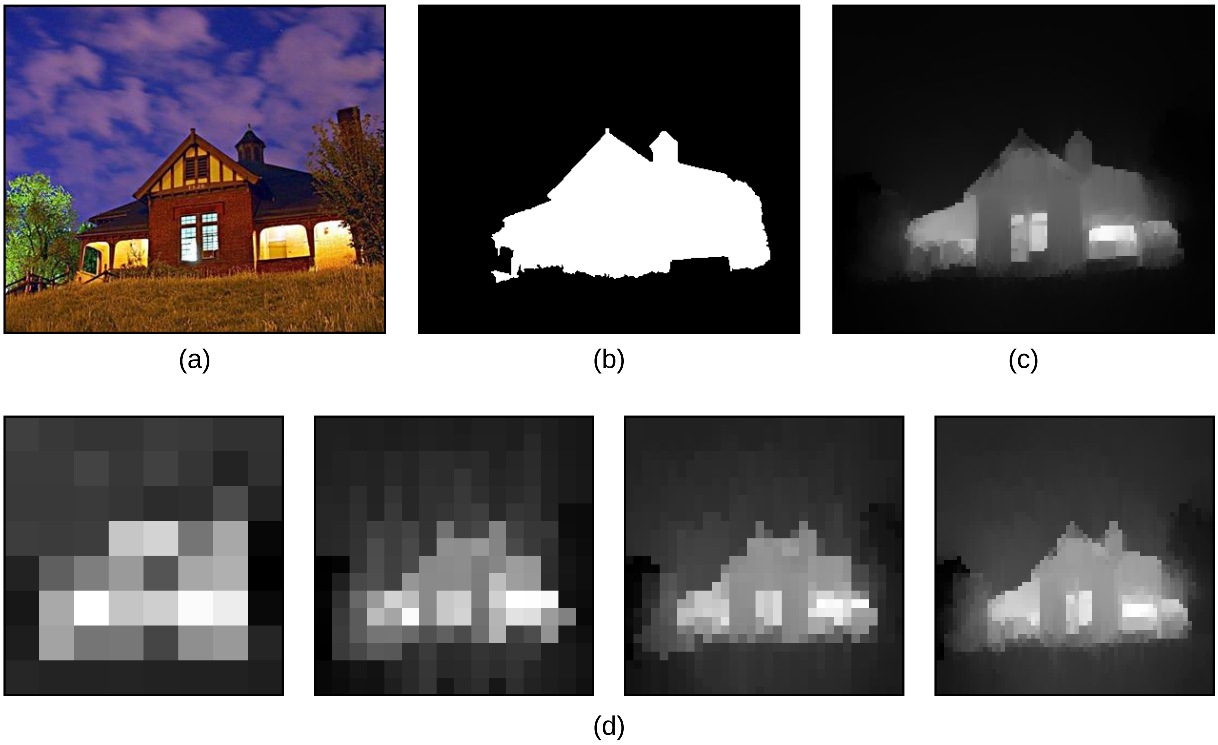

Figure 11: Multiscale joint upsampling of coarse-scale patch saliency estimation. **(a)** Input image. **(b)** Ground truth. **(c)** Saliency map. **(d)** Intermediate saliency estimation steps — from left to right: 8×8 patch reconstruction error, joint upsampling to 16×16, 32×32 and 64×64. While the initial saliency estimate presents a relatively heterogeneous aspect, through gradual joint upsampling it is efficiently processed into highly accurate and homogeneous regions.

curate results. Considering this, a multi-scale approach is adopted instead, in which the output of joint upsampling at a scale is used as input for the next scale to achieve gradual upsampling. This approach leads to more accurate results, and is in agreement with previous work on edge-preserving smoothing (PARIS et al., 2009), which shows that iterated filtering is more effective than simply adjusting range and smoothness parameters (see Section 2.3). An example of the process of multi-scale joint upsampling of coarse-scale patch saliency estimation shown in Figure 11.

5 Experimental Results

This chapter presents an experimental assessment of the proposed strategy. The experimental setup is described (Section 5.1) and the parameters of both pixel-level and patch-level saliency formulations is presented under the proposed strategy (Section 5.2). Then, a comparative assessment, both quantitative and qualitative, is presented with respect to several state-of-the-art methods (Section 5.3).

5.1 Setup

Assessment of saliency detection methods is made based on the quality of the saliency maps they output and the computational effort they require. The quality of a saliency map is measured by its accuracy with respect to a ground truth, which is an image obtained from human labeling and considered as the ideal output. Computational effort is measured by the average execution time on a common computer system. Most experiments were performed on an Intel Core i7-860 2.80 GHz CPU with 4 GB RAM, running the Windows 7 Professional (32-bit) operating system. The single exception is the parameter assessment for the joint upsampled *randomized color distance map*, presented in Section 5.2.1, which was performed on an Intel Xeon E5-2620 2.0 GHZ CPU with 24 GB RAM, running the Windows 10 Professional (64-bit) operating system, due to technical issues with the previous computer. Since the interest is in the *behavior* of execution time as the parameters vary, and not its absolute value, the assessment is unaffected by the difference in performance between the computer systems.

5.1.1 Metrics

Accuracy is measured in terms of *precision*, *recall*, and *F-measure*, which are standard metrics in salient region detection assessment (BORJI et al., 2014). *Precision* and *recall* are defined as:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \tag{10}$$

where TP (true positives) are salient pixels correctly detected as such, FN (false negatives) are salient pixels detected as background and FP (false positives) are background pixels detected as salient. Since saliency maps are usually given in shades of gray, and these metrics are for binary values, the maps are thresholded for each value in the [0 .. 255] interval. The accuracy of a method on an image is summarized as the precision-recall curve for all thresholds in this interval, while the accuracy for an entire dataset is summarized as the average precision-recall curve for all images.

Besides the precision-recall curve, accuracy can also be summarized by the F-measure, which is the weighted harmonic mean of precision and recall, that is:

$$F_\beta = (1+\beta^2)\frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall}, \tag{11}$$

where β is used to emphasize the effect of precision or recall. Since many authors consider precision more important than recall for saliency detection, it is common to adopt $\beta^2 = 0.3$ (e.g., Achanta et al. (2009), Li et al. (2013b), Cheng et al. (2015)). In order to keep the experiments more easily comparable to the literature, this work also follows this choice. While the precision-recall curve is computed for all thresholds in [0 .. 255], F-measure is computed for a single adaptive threshold, *twice the average saliency of the image*, following the widely adopted assessment approach by Achanta et al. (2009).

5.1.2 Datasets

The datasets adopted in the assessment are described as follows, along with sample images of each (Figure 12).

- **ASD:** Also known as MSRA1K (ACHANTA et al., 2009), this dataset contains 1000 images sampled from the MSRA database (LIU et al., 2007), with their respective contour-accurate ground truths. These ground truths were manually extracted from bounding boxes labeled by three human subjects. The images in this dataset were collected mostly from internet forums and search engines, have approximately 400×300 pixels, and contain mostly a single, relatively large, salient object.

- **MSRA10K:** Similarly to the ASD dataset, the images in this dataset were sampled from the MSRA database, and had their contour-accurate ground truths extracted in the same manner. However, this selection is substantially larger, containing 10,000 images with their respective ground truths (CHENG et al., 2015). Also similarly to ASD, the images in this dataset have approximately 400×300 pixels and contain a single, relatively large, salient object.

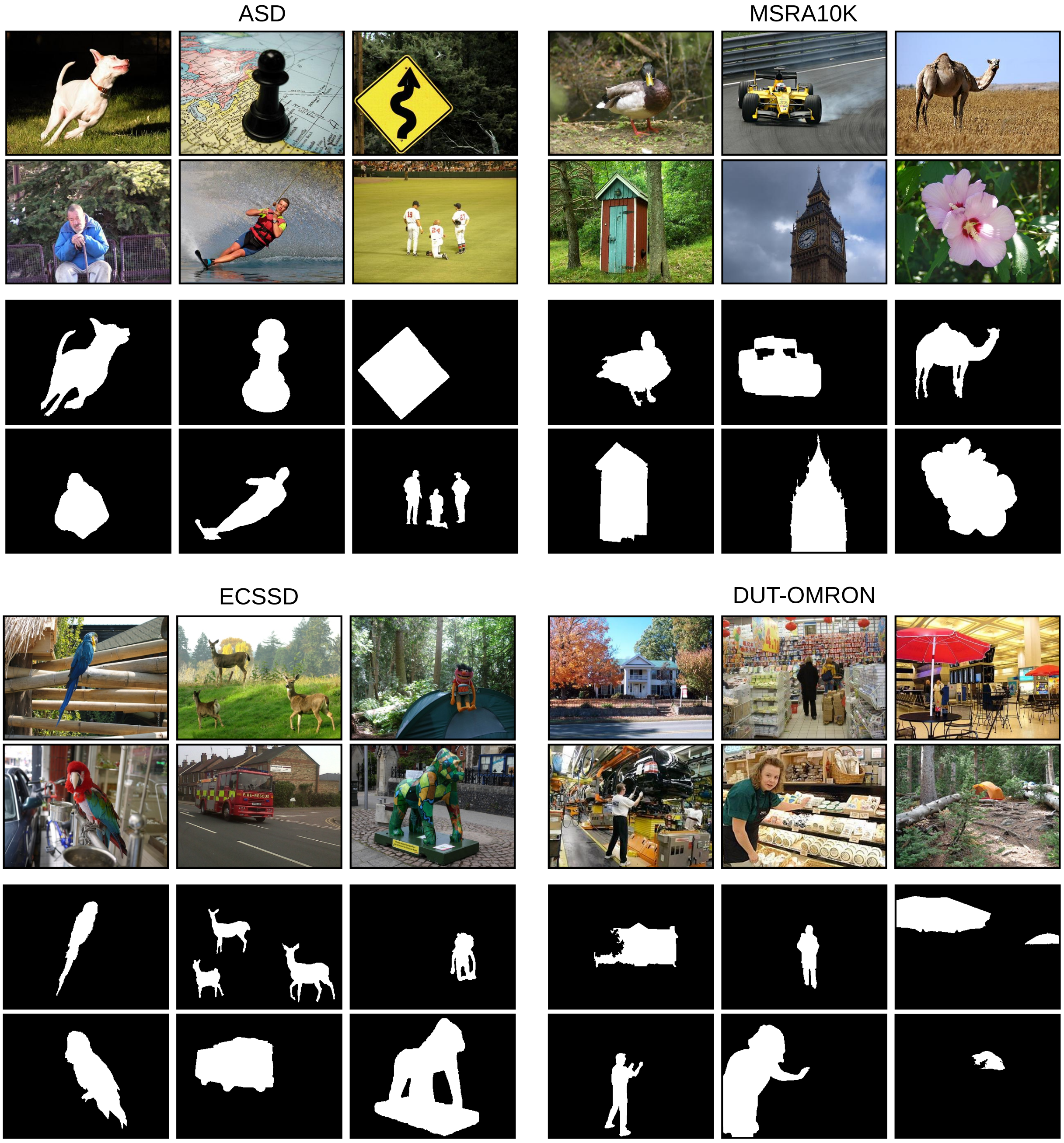

Figure 12: Examples of scenes depicted in the datasets, along with their ground truths. The datasets are displayed, roughly, in order of increasing detection difficulty, from top-left to bottom-right.

- **ECSSD:** This dataset contains 1000 images with their respective contour-accurate ground truths. The images in this dataset have approximately 400×300 pixels, contain relatively complex background, and possibly multiple salient objects (SHI et al., 2016). The ground truths were extracted from labeling data from five human subjects.

- **DUT-OMRON:** This dataset contains 5,168 images with their respective ground truths. All images have approximately 400×300 pixels, and one or more salient regions over relatively complex backgrounds (YANG et al., 2013). Contour-accurate, bounding box, and fixation prediction ground truths are provided — these were extracted from labeling and eye-tracking data from five human subjects per image, from a group of 25 volunteers.

5.1.3 Compared methods

The proposed strategy was assessed and compared to seven other state-of-the-art saliency detection methods, namely *Spectral Residual* (SR) (HOU; ZHANG, 2007), *Frequency-tuned* (FT) (ACHANTA et al., 2009), *Context-aware* (CA) (GOFERMAN; ZELNIK-MANOR; TAL, 2010), *Random Center-surround* (RCS) (VIKRAM; TSCHEREPANOW; WREDE, 2012), *PCA Saliency* (PCAS) (MARGOLIN; TAL; ZELNIK-MANOR, 2013), *Absorbing Markov Chain* (AMC) (JIANG et al., 2013a), and *Dense and Sparse Reconstruction* (DSR) (LI et al., 2013b). Besides relevance, the criteria for selection of these methods were mostly number of citations (currently, SR, CA, and FT have between 1,700 and 2,700 citations on Google Scholar), similarity to the proposed approach (RCS is based on random sampling, PCAS is based on a PCA subspace, and DSR is based on patch reconstruction), and performance (AMC is the fastest among the most accurate methods in the benchmark by Borji et al. (2015)).

5.1.4 Implementation details

The proposed saliency estimation approaches were implemented in MATLAB, using the Image Processing Toolbox (IPT). Due to efficiency and numerical stability concerns, PCA was performed using *Singular Value Decomposition* (SVD), since it avoids the explicit computation of the covariance matrix (YANG et al., 2004). For edge-preserving smoothing, the original C++ source code for the Fast Global Smoother from its authors (MIN et al., 2014) was used, through a MEX interface. All saliency maps output using the proposed strategy are subject to gamma correction with $\gamma = 2$. This is done merely for aesthetic reasons and does not impact on accuracy or execution time in any perceptible manner, consequently, it is considered as an implementation detail and therefore is not included as a parameter in the assessment.

5.2 Parameter assessment

5.2.1 Randomized color distance map

The randomized color distance map has three parameters: *subset size* $|\mathbf{\Omega_s}| \in [1\,..\,n]$, *downsize scale* $D \in (0, 1]$, and *boundary ratio* $B \in (0, 0.5]$. Each parameter was assessed while keeping the remaining fixed at default values, to isolate their effects. For D and B, the default values are $D = 1.0$ (i.e., no downsampling) and $B = 0.5$ (i.e., no boundary ratio). Since there is no such obvious default value for $|\mathbf{\Omega_s}|$, it was determined based on its accuracy and execution time for several values. As the first row of Figure 13 shows, accuracy saturates for $|\mathbf{\Omega_s}| \approx 40$ across all datasets, however, increasing $|\mathbf{\Omega_s}|$ to a value larger than approximately 10 is not cost-effective, since it improves accuracy only marginally while increasing execution time significantly. Considering this, $|\mathbf{\Omega_s}| = 10$ is adopted as default subset size.

Downsize scale also has a significant impact on execution time, and consequently must be set to the smallest value possible. As the second row of Figure 13 shows, accuracy remains relatively unchanged for most scales. The only sharp change in accuracy occurs from $D = 0.1$ to $D = 0.2$ — from this value upwards accuracy improves marginally (and not always monotonically), while execution time increases significantly. Considering this fact, downsize scale is set as $D = 0.2$, since it is enough to achieve most of the possible accuracy range on all datasets.

Boundary ratio does not impact execution time, since it merely defines the area from which $\mathbf{\Omega_s}$ is sampled. As the third row of Figure 13 shows, adopting $B = 0.5$, which is equivalent to disabling boundary ratio, results in the lowest accuracy. This suggests that boundary prior always improves accuracy. Boundary ratio is set to $B = 0.2$, since it results in the highest F-measure for all datasets.

5.2.2 Patch reconstruction from a boundary basis

The joint upsampled patch reconstruction approach has two parameters: *set of joint upsampling scales* $\mathbf{s} \subseteq \{16{\times}16,\ 32{\times}32,\ 64{\times}64,\ 128{\times}128\}$, and *fraction of principal components retained* $f_{eig} \in (0.1, 1.0]$. The proposed approach joint upsamples patch saliency estimates across several scales. However, it is not necessary to joint upsample to full-resolution, nor it is necessary to include all intermediate scales. Considering this, due

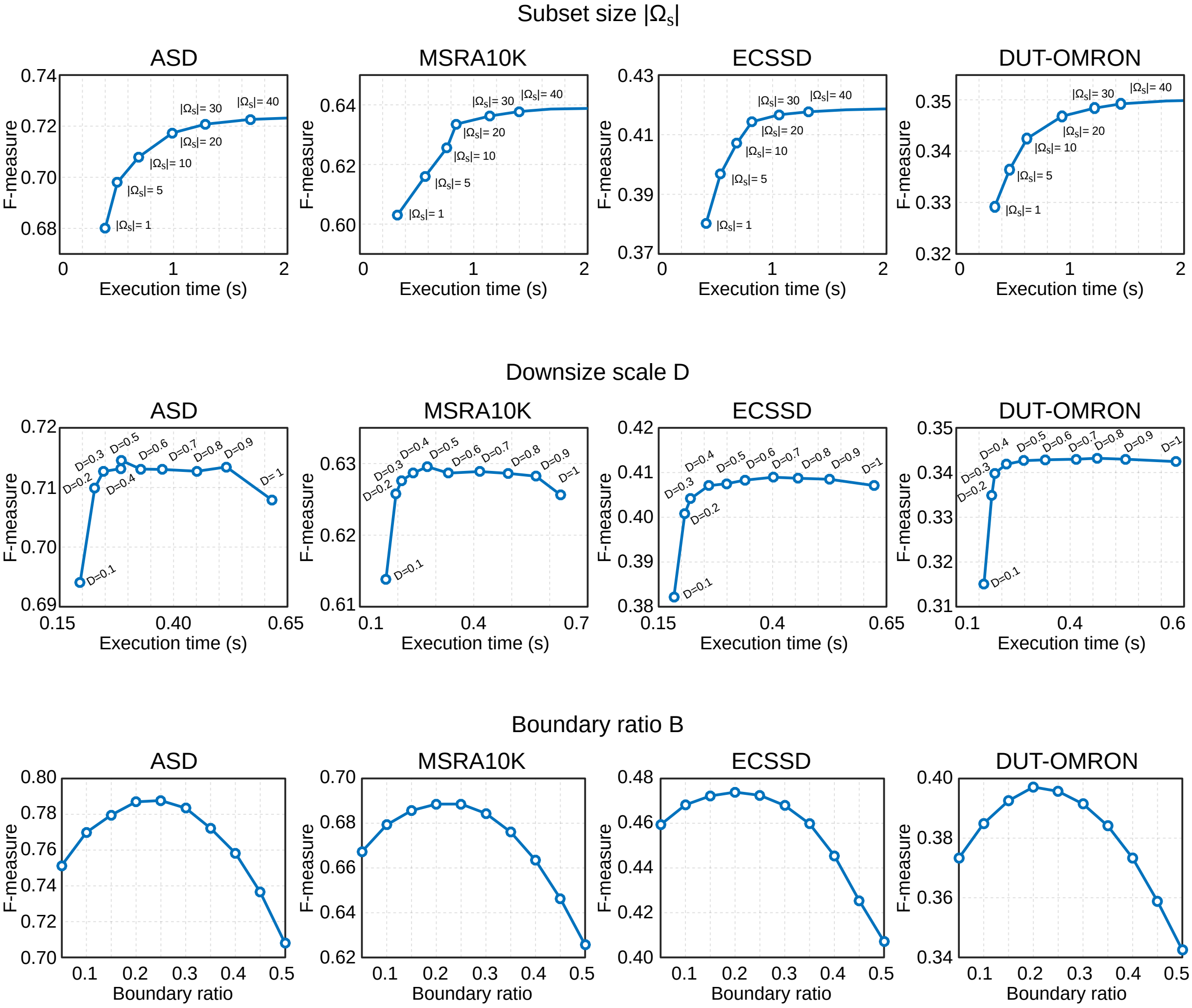

Figure 13: Parameter assessment — joint upsampled randomized color distance map. **(a)** Subset size $|\mathbf{\Omega_s}|$ ($D = 1.0$, $B = 0.5$). Small values offer the best trade-off. For $|\mathbf{\Omega_s}| > 10$, execution time increases significantly with only marginal accuracy improvement. **(b)** Downsize scale D ($|\mathbf{\Omega_s}| = 10$, $B = 0.5$). The best trade-off occurs for $D = 0.2$. Larger values incur significant computational cost for almost no accuracy improvement. **(c)** Boundary ratio B ($|\mathbf{\Omega_s}| = 10$, $D = 1.0$). Since any valid boundary ratio higher than 0.5 (i.e., no boundary prior) improves accuracy, boundary prior is always advantageous. The best trade-off occurs for $B = 0.2$. The computation was performed on an Intel Xeon E5-2620 2.0 GHZ CPU with 24 GB RAM.

to efficiency concerns, after the largest scale adopted for **s**, the output is uniformly upsampled to full-resolution using nearest neighbor interpolation. Note that **s** is specified in number of pixels instead of a ratio (cf. parameter D in Section 5.2.1), since the coarse patch estimation dimensions are known and fixed.

Table 2 shows an assessment of different scale combinations in terms of F-measure and average execution time. As expected, iterating through intermediate scales leads to higher accuracy than joint upsampling directly to the finest scale, as

Table 2: Performance of joint upsampling for different scale combinations. F-measure (average for $f_{eig} \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$) was computed with $\beta^2 = 0.3$, following Achanta et al. (2009). For each number of scales, the most accurate combination is indicated in bold typeface. Entries are listed from top to bottom in order of increasing number of combined scales. The computation was performed on an Intel Core i7-860 2.80 GHz CPU with 4GB RAM.

Scales				**F-measure**				**Average exec. time (ms)**
16×16	32×32	64×64	128×128	ASD	MSRA10K	ECSSD	DUT-OMRON	
●	○	○	○	0.71	0.65	0.52	0.40	50.60
○	●	○	○	**0.72**	**0.67**	**0.53**	**0.40**	**50.60**
○	○	●	○	0.70	0.65	0.52	0.39	51.22
○	○	○	●	0.67	0.63	0.51	0.37	53.22
●	●	○	○	0.75	0.68	0.54	0.41	58.94
●	○	●	○	**0.75**	**0.69**	**0.54**	**0.41**	**59.38**
●	○	○	●	0.75	0.69	0.54	0.41	61.80
○	●	●	○	0.74	0.68	0.54	0.41	59.84
○	●	○	●	0.74	0.68	0.54	0.41	61.78
○	○	●	●	0.72	0.67	0.53	0.40	62.15
●	●	●	○	0.76	0.69	0.54	0.41	67.62
●	●	○	●	**0.76**	**0.69**	**0.55**	**0.41**	**69.65**
●	○	●	●	0.76	0.69	0.55	0.41	70.22
○	●	●	●	0.75	0.69	0.55	0.41	71.32
●	●	●	●	**0.77**	**0.69**	**0.55**	**0.41**	**78.96**

does including additional scales. It is worth noting that each additional scale leads to an additional joint upsampling, which is the more significant step in terms of execution time, as can be verified by the fact that the difference in execution time due to increasing scale (e.g., single-scale from 32×32 to 64×64) is more subtle than due to increase in number of scales (e.g., one to two scales). Moreover, including finer scales does not necessarily improve accuracy. Interestingly, the best combinations for each number of scales are consistent across datasets. These results suggest the most adequate scale combinations according to the number allowed by the requirements of the application. While the differences in execution time might seem small, they lead to significantly different frame processing rates. In the comparative assessment, three scales ($\mathbf{s} = \{16\times16,\ 32\times32,\ 64\times64\}$) are used, since with this number of scales the execution time is significantly shorter than with four scales, while achieving almost the same accuracy.

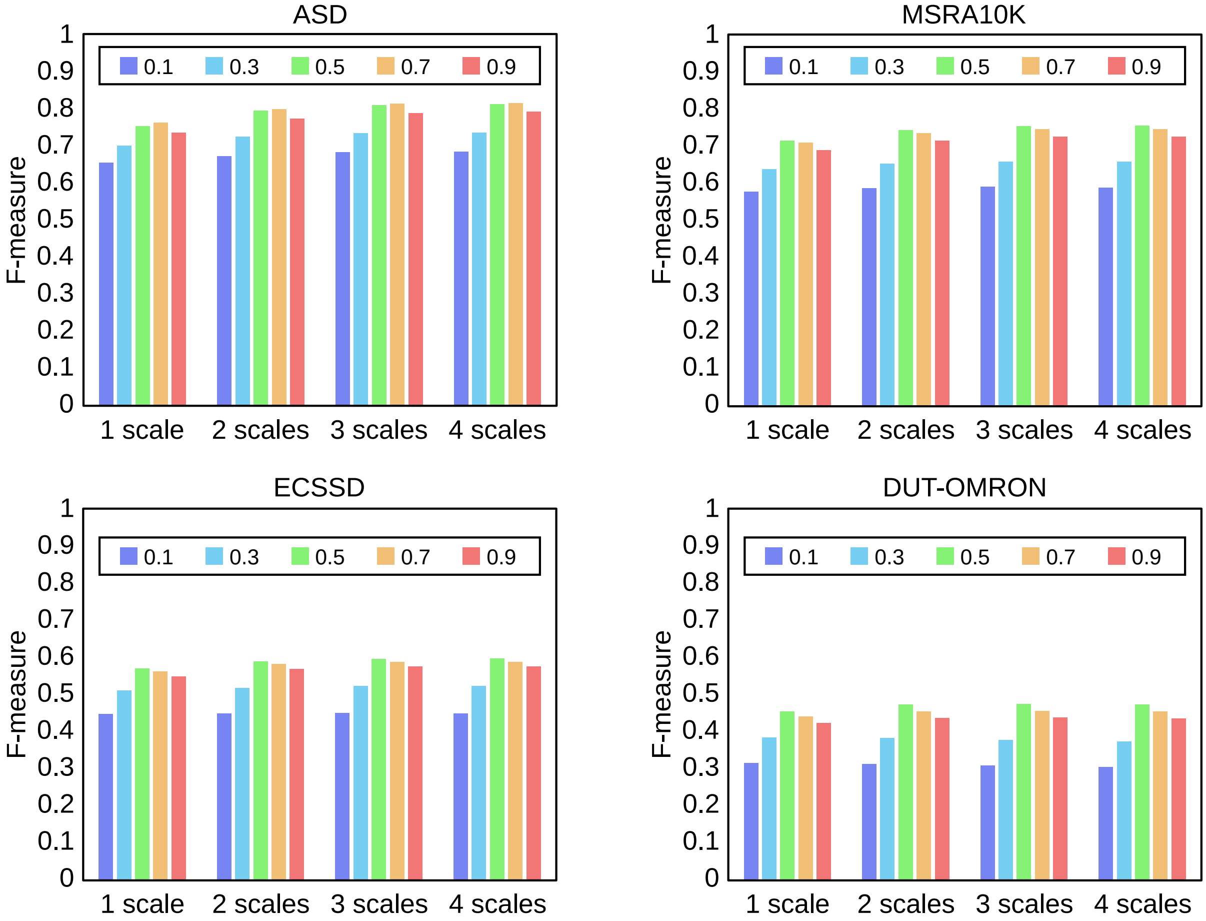

Figure 14: F-measure ($\beta^2 = 0.3$) for different fractions $f_{eig} \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ kept from the entire set of eigenvectors.

PCA allows reconstruction with minimum squared error from the eigenvectors (i.e., principal components) associated with the largest eigenvalues (ABDI; WILLIAMS, 2010). In this manner, it is common to reduce data dimensionality by discarding some of the eigenvectors according to some general rule-of-thumb (HALL; MARSHALL; MARTIN, 2000). However there is no guarantee that these are adequate for perception-oriented applications such as saliency estimation, in which perfect reconstruction is not crucial. Considering this, the proposed approach was assessed by keeping different fractions f_{eig} of the total eigenvectors. The results, presented in Figure 14 report the F-measure for each number of scales, for which only the most accurate combinations were chosen, as described in Table 2.

The fraction of eigenvalues that results in the highest accuracy is consistently $f_{eig} = 0.5$, across both datasets and number of scales. The exception is on the ASD dataset, in which accuracy is slightly higher for $f_{eig} = 0.7$. However, this advantage is marginal and disappears as the number of scales increases. The results are consis-

Table 3: F-measure ($\beta^2 = 0.3$) and average execution time (for a 400×300 image) of the compared methods. The three most accurate methods in each dataset are indicated in bold. Parameters for JSAL-pixel: $|\mathbf{\Omega_s}| = 10$, $D = 0.2$, $B = 0.2$. Parameters for JSAL-patch: $f_{eig} = 0.5$, **s** is set as the most accurate combination for each number of scales n_s according to Table 2. The computation was performed on an Intel Core i7-860 2.80 GHz CPU with 4GB RAM.

Method	**Average exec. time (s)**	**F-measure**			
		ASD	MSRA10K	ECSSD	DUT-OMRON
JSAL-patch					
$n_s = 1$	0.05	0.75	0.71	0.57	0.45
$n_s = 2$	0.06	0.79	0.74	0.59	**0.47**
$n_s = 3$	0.07	**0.81**	0.75	**0.60**	0.47
$n_s = 4$	0.08	0.81	**0.76**	0.60	0.47
JSAL-pixel	0.06	0.79	0.70	0.48	0.40
DSR	5.67	**0.85**	**0.81**	**0.69**	**0.53**
AMC	0.18	**0.89**	**0.84**	**0.70**	**0.53**
PCAS	5.49	0.80	0.75	0.58	0.46
RCS	0.73	0.66	0.62	0.52	0.39
CA	38.41	0.56	0.58	0.43	0.36
FT	0.06	0.67	0.59	0.38	0.31
SR	0.01	0.46	0.49	0.22	0.19

tent with the observation by previous authors (HYVÄRINEN; HURRI; HOYER, 2009, p. 104) that, for natural images, principal components associated with low variance encode mostly noise. Considering this, the fraction $f_{eig} = 0.5$ is adopted in the comparative assessment.

5.3 Comparison to the state-of-the-art

5.3.1 Quantitative assessment

The precision-recall performance of the proposed strategy (referred to as *JSAL* in the assessment), with both pixel-level (randomized color distance map) and patch-level (patch reconstruction error from a boundary basis) saliency estimation, is presented in Figure 15, along with that of seven state-of-the-art methods, for comparison. The F-measure of all compared methods as well as their average execution time is presented in Table 3.

Overall, the proposed method presents competitive accuracy to the best per-

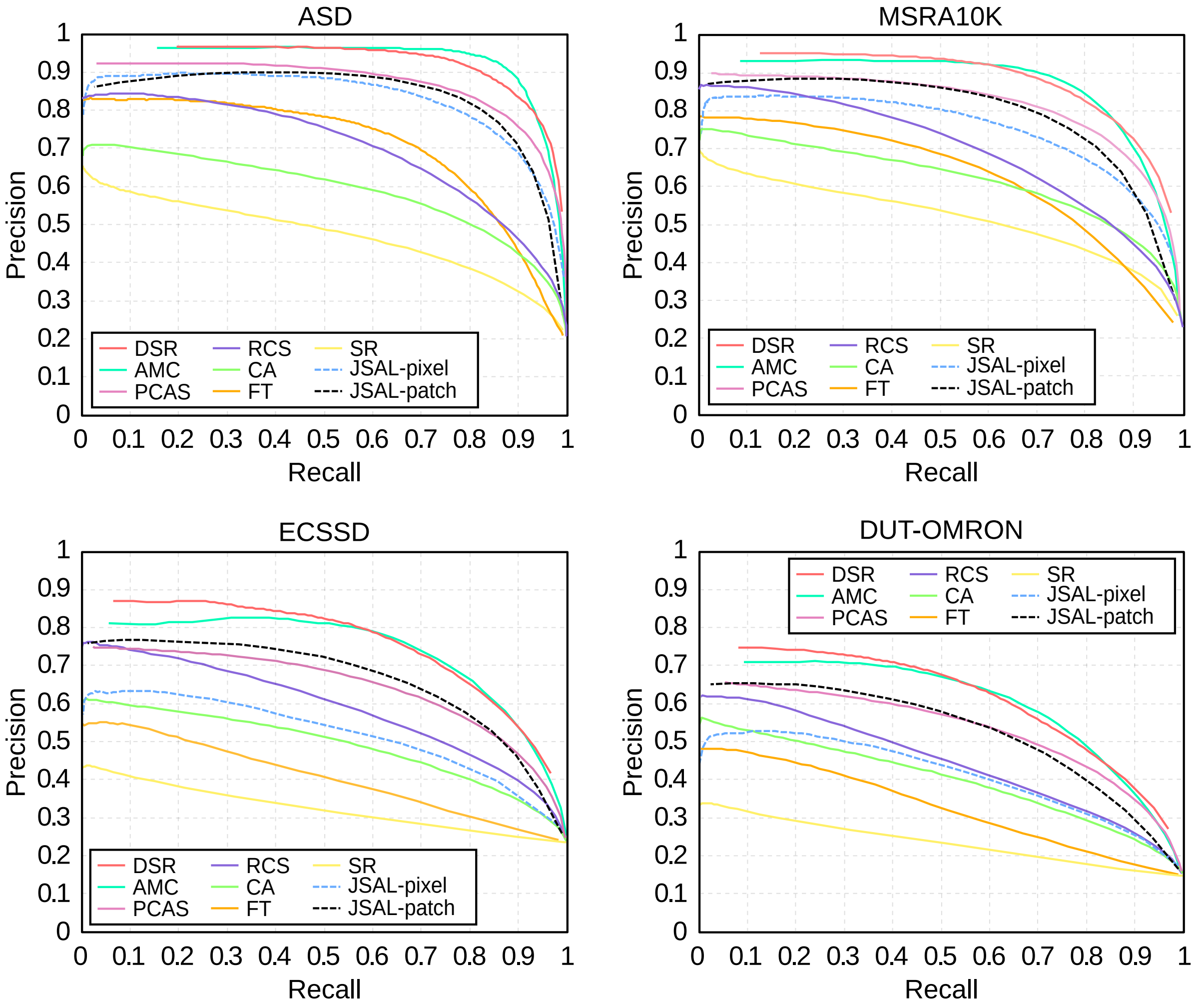

Figure 15: Precision-recall curves for the compared methods. Parameters for JSAL-pixel: $|\mathbf{\Omega_s}| = 10$, $D = 0.2$, $B = 0.2$. Parameters for JSAL-patch: $\mathbf{s} = \{16{\times}16,\ 32{\times}32,\ 64{\times}64\}$, $f_{eig} = 0.5$.

forming methods in the state-of-the-art. JSAL-pixel presents superior performance to all pixel-level methods (i.e., RCS, CA, FT, SR), except RCS, which is more accurate in the ECSSD and DUT-OMRON datasets. This might be due to its relatively more local estimation approach of RCS compared to other pixel-level approaches, which can result in increased robustness to more complex scenes, such as those in the ECSSD and DUT-OMRON datasets. However, this comes at the cost of an execution time more than 10 times longer than JSAL-pixel (Table 3). The decrease in accuracy of all pixel-level methods from the ASD to the MSRA10K dataset is already expected, as Cheng et al. (2015) reported experiments suggesting that pixel-level methods tend to scale worse than region-based methods.

While JSAL-pixel becomes less competitive in the more complex ECSSD and DUT-OMRON datasets, JSAL-patch is consistently competitive on all datasets, pre-

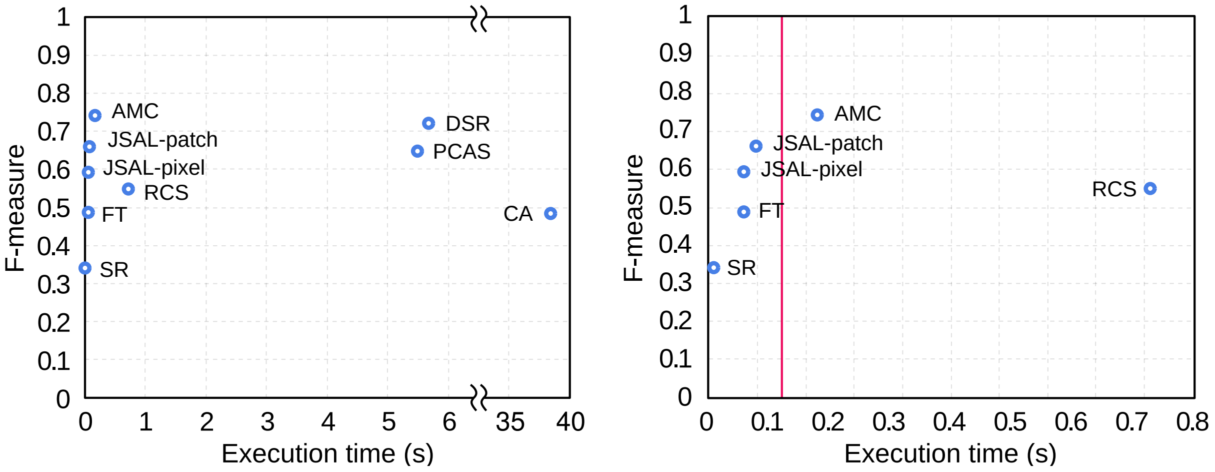

Figure 16: Trade-off between accuracy (F-measure, $\beta^2 = 0.3$) and execution time. **Left:** All compared methods. **Right:** Methods that perform under one second per 400×300 image. The closer to the top left the better. The F-measure indicated accounts for the average across all datasets (i.e., ASD, MSRA10K, ECSSD, DUT-OMRON). The red line indicates the approximate time limit for bottom-up visual attention by the human visual system according to Theeuwes (2010). The computation was performed on an Intel Core i7-860 2.80 GHz CPU with 4GB RAM.

senting one of the three highest accuracies on all of them (Table 3). This is largely due to its region-level approach, which estimates saliency patch-wise, and is consequently not subject to the scalability limitations of pixel-level methods. The accuracy of JSAL-patch is moderately lower than AMC and DSR, but in terms of execution time, it is 2.5 and 81 times faster, respectively. The superior efficiency over DSR is particularly significant, considering the similarity between the approaches. Compared to PCAS, which presents the precision-recall curve most similar to JSAL-patch, the proposed approach also has a large efficiency advantage — it executes 78 times faster.

Two methods present execution time shorter or equivalent to the JSAL-pixel and JSAL-patch: SR and FT. SR is the fastest method among all assessed (0.01 second per 400×300 image), which is mainly because it is estimated at a coarse scale (∼64×64) and not subject to any post-processing besides Gaussian filtering. This severely compromises accuracy for salient region detection, and consequently results in the lowest accuracy on all datasets, by a large margin. FT presents execution time equal to JSAL-pixel and 0.01 second shorter than JSAL-patch. However, while it can be argued that its accuracy is almost competitive in the ASD and MSRA10K datasets, it does not scale well to more complex scenes, presenting the second lowest accuracy on the ECSSD and DUT-OMRON datasets.

Analyzing the precision-recall curves alone might be misleading with respect to the quality of the methods. While the F-measure and execution time of the compared methods are presented in Table 3, a graphical summarization of the trade-off between accuracy and efficiency can reveal these qualities more explicitly. Figure 16 presents such representation in a scatterplot, in which points closer to the top left of the plot present better trade-off. The proposed methods, along with AMC, achieve the best trade-offs, with JSAL-patch presenting an arguably equivalent trade-off to AMC, since it presents a modest decrease in accuracy while executing in less than half of its execution time. Despite the lower precision-recall curves of JSAL-pixel on the two harder datasets (ECSSD and DUT-OMRON), considering the averaged accuracy on all datasets and execution time, it clearly presents one of the best trade-offs — the best among the pixel-based methods.

The rule-of-thumb presented in the motivation of this thesis, regarding the time limit for bottom-up visual attention in the human visual system ($\sim$150 ms, see Section 1.2, Figure 2), is revisited in Figure 16 (right) to give a better perspective on the adequacy of the time frame required by the compared methods. Considering the methods with highest accuracy, while AMC manages to perform close to this time frame, DSR and PCAS take several seconds to process a single image and do not present a competitive trade-off. JSAL-patch and JSAL-pixel are the most accurate methods within the indicated time limit, corroborating the thesis statement that the proposed strategy is capable of effective and efficient visual saliency detection.

5.3.2 Qualitative assessment

Examples of saliency maps computed from the compared methods are presented in Figures 17 and 18, for comparison. As mentioned previously, in general, methods that operate exclusively at pixel-level perform reasonably well in simple scenes, but do not scale well as complexity increases. This can be verified, for instance, in the saliency maps presented in the third column of Figure 17 and second column of Figure 18. In the first case, the input image depicts a relatively simple scene, with predominantly homogeneous background, and a salient region with very distinctive colors enclosed by unambiguous edges. The pixel-level methods assessed, namely SR, FT, CA, RCS, and JSAL-pixel, are capable of detecting the salient region reasonably well. On the second case, however, the input image is more cluttered, and pixel-level methods output saliency maps that are significantly less accurate.

The lack of accuracy on more complex scenes can be attributed to some relatively consistent qualitative aspects of the output for each method. For instance, CA and SR overemphasize edges, something that is more severe in the latter case, since it operates in a coarser scale. FT emphasizes undesired textures, besides detecting boundaries as salient in several cases. RCS, despite scaling better than other pixel-level methods on the ECSSD and DUT-OMRON datasets, outputs low-quality saliency maps, with blurry and inhomogeneous regions. It is possible to notice that its superior scalability compared to other pixel-level methods is mostly a result of its heavily center-biased output, which emphasizes the center of the saliency map almost irregardless of the object boundaries. In some cases RCS even emphasizes the content in the center of the image *except* the salient region (e.g., second and seventh columns of Figure 18). Most of the aforementioned aspects do not occur in the output of JSAL-pixel. It does not emphasize edges, small-scale textures, or boundary distractors. The first two are due to joint-upsampling, which "spreads" the coarse-scale estimate inside region boundaries. The latter is due to the boundary prior, which allows modeling the boundary as background in an implicit manner, i.e., when selecting color samples. In this manner, it is possible to avoid heavily biasing towards the center, as RCS does.

The limitations of pixel-level methods can be at least mitigated by region-level analysis. While this is evident from the substantially higher accuracy presented by the region-level methods (i.e., DSR, AMC, PCAS, JSAL-patch) with respect to pixel-level methods (Figure 15), it is even more evident through quantitative analysis. For instance, in the second and fifth columns of Figure 18, it can be noticed that the output of region-level methods is much more homogeneous. This reduction in "visual clutter" reduces the detection of background regions as salient and improves the uniformity of the detection inside salient regions.

Among region-level methods, PCAS presents the most heterogeneous output, besides emphasizing edges. These are characteristics of pixel-level methods, and occur because PCAS outputs a combination of both pixel-level and region-level saliency maps. AMC outputs homogeneous and accurate saliency maps, however, since it is formulated as a graphical model based on propagation from the boundary towards the center of the image, it tends to emphasize the center of the image when the transition from the boundaries is smooth (e.g. third and fifth columns of Figure 17, and third column of Figure 18). DSR also outputs homogeneous and accurate saliency maps, however, since it employs explicit center-bias and multiscale superpixel decomposition, it might output artifacts on object boundaries (e.g., first columns of Figures 17 and 18), create inexistent texture (e.g., fourth column of Figure 17), or overemphasize the center of the image (e.g., third column of Figure 18).

In contrast to these region-level methods, JSAL-patch does not rely on superpixel segmentation, but still achieves not only competitive accuracy, but also saliency maps with comparable, and in several cases superior, quality. Despite not being as accurate as AMC and DSR, JSAL-patch is robust to some of their limitations. For instance, by adopting a boundary prior instead of an explicit center-bias, it avoids overemphasizing the center of the image. Additionally, PCAS, AMC, and DSR adopt the SLIC algorithm for segmentation, which outputs a fixed number of uniformly spaced segments that can lead to the artifacts (e.g., DSR in the first columns of Figure 17 and 18) and mosaic-like output (e.g., AMC in third and fifth columns in Figure 17). Since JSAL-patch avoids superpixel segmentation in favor of joint upsampling, content is simply smoothed inside edges and such disadvantages are avoided. Moreover, by adopting a simpler model, JSAL-patch provides better detection in some cases that might confuse more complex algorithms (e.g., AMC and DSR in the fourth and fifth columns in Figure 17).

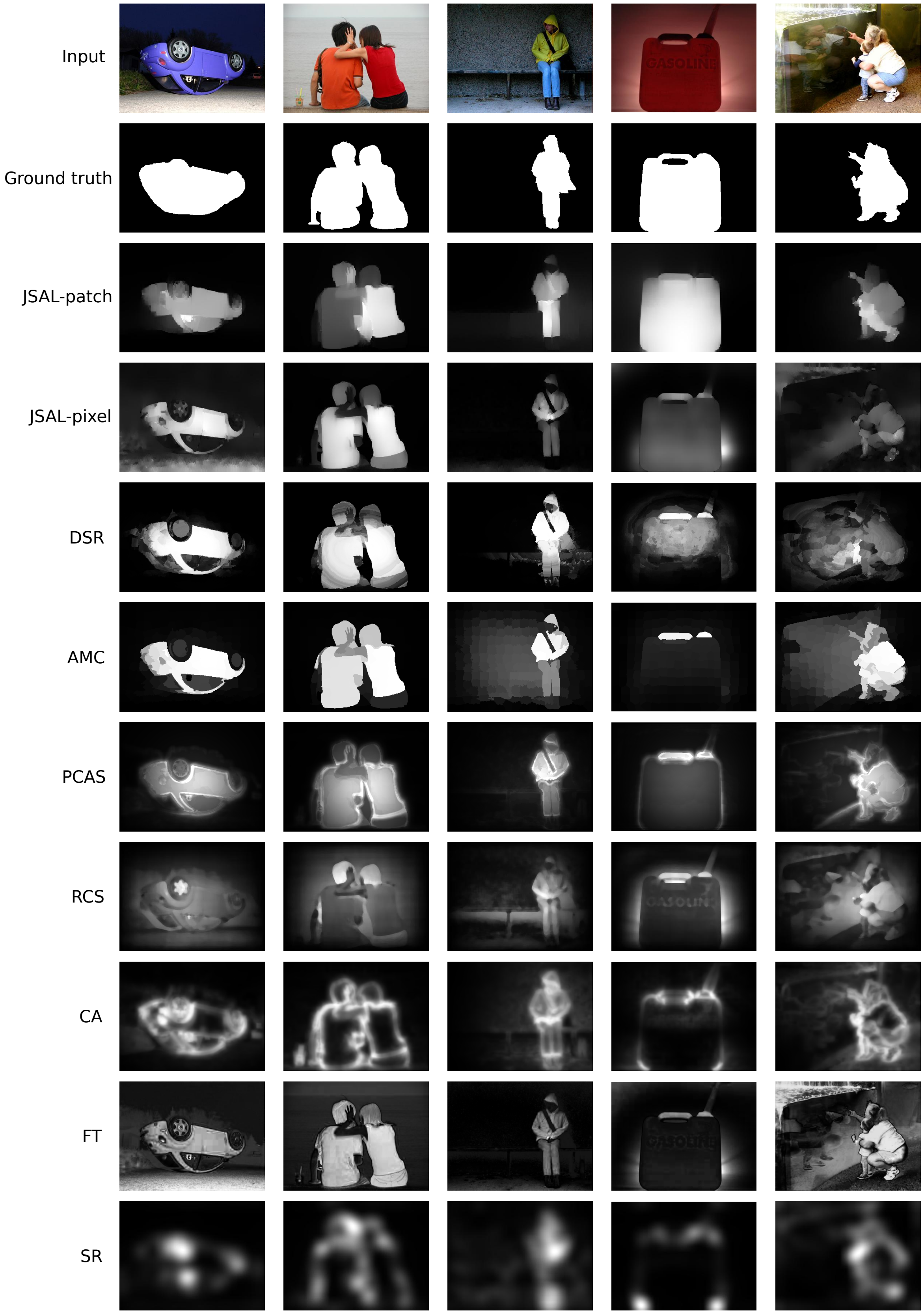

Figure 17: Saliency maps computed from the compared methods. Except for JSAL-pixel and FT, all pixel-level methods (i.e., RCS, CA, SR) overemphasize borders or small regions. On some cases, small details (fourth column) lead even region-level methods to output inaccurate saliency maps — for this image, JSAL-patch is the only one to detect the salient region correctly.

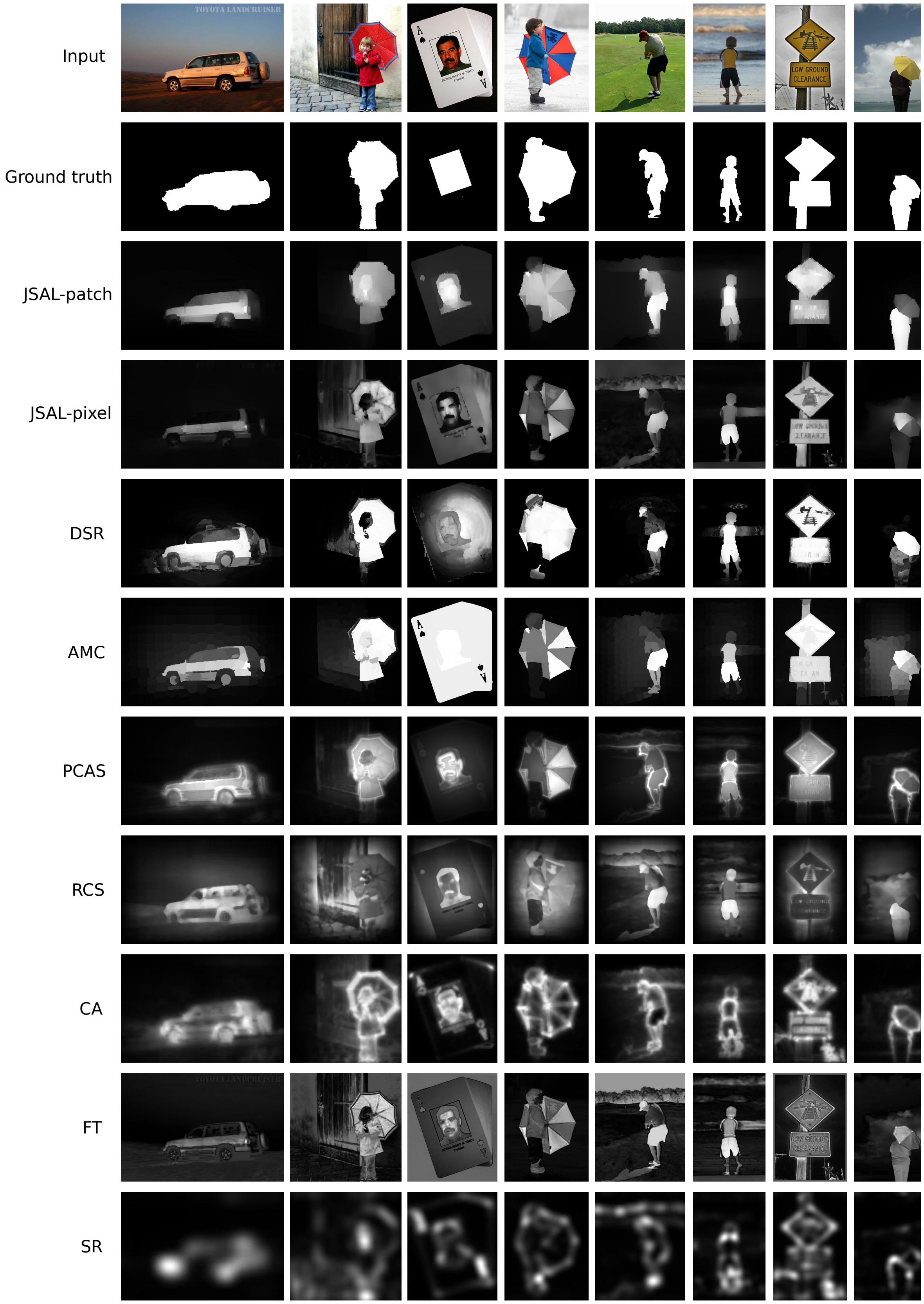

Figure 18: Examples of joint upsampled patch saliency estimates (continued). JSAL-pixel outputs cleaner saliency maps than other pixel-level methods, while JSAL-patch manages to correctly detect even some challenging cases, in which other region-level methods fail (third column) or overemphasize details (fifth column).

6 Conclusions

This thesis began with the motivation that, despite recent advances, most saliency detection methods are currently being designed only to perform accurately, with few concerns regarding efficiency. In bottom-up visual attention systems, computational efficiency is paramount. This mechanism is assumed to precede most processes in the visual system, such that inefficiency in its operation can hinder the operation of later stages.

However, efficiency alone is clearly not enough. Accuracy is also necessary, since there is no point in employing an inaccurate process, regardless of how efficient it is. Taking this into account, and based on the principles presented in Section 2.2 (*Low-dimensional Image Representation*), as well as experimental results on human visual attention (INTRILIGATOR; CAVANAGH, 2001), it was argued that estimation of visual saliency in coarse-scale can be not only efficient, due to the reduced amount of data compared to the full-resolution input, but also the most adequate scale for this purpose. However, coarse-scale estimation alone is not compatible with the accuracy requirements of saliency maps for salient region detection. Thus, an efficient joint upsampling approach was proposed, which enables leveraging both the advantages of coarse-scale saliency estimation (i.e., efficiency, agreement with experimental evidence, abstraction of unnecessary detail) and fine-scale edge information (i.e., high accuracy). This approach was designed to provide good trade-off between accuracy and efficiency, which was demonstrated by a comparative assessment with other seven state-of-the-art methods on four major datasets. Two saliency formulations were proposed for computation of coarse-scale estimates in the presented strategy, one operating at pixel-level and the other at patch-level, both achieving accuracy and efficiency among the top performing methods assessed. The former presented simple implementation and fine-grained adjustability, being an adequate choice for applications that require short execution time and are based on relatively simple scenes (e.g., Lie et al. (2016), Lie et al. (2017)). The latter presented superior accuracy and scalability, as demonstrated by its performance among the three most accurate on all datasets, and its trade-off between accuracy and efficiency, which was the second best overall and the best within the time frame expected for bottom-up attention by the human visual system.

During the development of this thesis, several questions and improvements ideas appeared. Some of them present potential for future work and are summarized as follows:

- **Spatially-variant resolution.** In this thesis, coarse-scale was computed simply as uniform downsampling — this is not biologically accurate. It is known that the acuity of the primate retina is not uniform, its higher is in the center and decays rapidly towards the periphery (KANDEL; SCHWARTZ; JESSELL, 1995). This suggests that exploring a spatially-varying resolution approach might be interesting, not only for increased biological-plausibility, but for analysis at a possibly more appropriate level of detail. This approach has been employed in several computer vision applications (BOLDUC; LEVINE, 1998), including visual attention (TRAVER; BERNARDINO, 2010).

- **Scene decomposition and pixel-level estimation.** Similarly to previous work (CHENG et al., 2015), the experiments in this thesis demonstrated that pixel-level methods do not scale to more complex scenes. However, some pixel-level approaches achieved high accuracy on simpler datasets, while performing with remarkably short execution time. This suggests that if computationally efficient methods for decomposing an image into simpler scenes are available, it might be possible to leverage pixel-level methods as efficient components in more elaborate visual attention models. This is a promising approach, considering that recent approaches for estimation of "objectness" of image regions, which could be used for image decomposition, have reported impressive processing rates (e.g., 300 fps by Cheng et al. (2014)).

- **Alternative to simple PCA subspace.** Based on its simplicity and effectiveness in previous work, the proposed patch-level saliency estimation approach operated on a PCA subspace, which was computed from simple *Singular Value Decomposition* (SVD). There are very elegant approximation strategies that are more efficient (HALKO; MARTINSSON; TROPP, 2011). While such approximations are interesting from a numerical point of view, they could be further investigated under the context of the tolerance of visual perception to inaccuracies.

References

ABDI, H.; WILLIAMS, L. J. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 2, n. 4, p. 433–459, 2010.

ACHANTA, R.; HEMAMI, S.; ESTRADA, F.; SÜSSTRUNK, S. Frequency-Tuned Salient Region Detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, USA: IEEE, 2009. p. 1597–1604.

ACHANTA, R.; SHAJI, A.; SMITH, K.; LUCCHI, A.; FUA, P.; SÜSSTRUNK, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 34, n. 11, p. 2274–2282, Nov 2012.

ADELSON, E. H.; ANDERSON, C. H.; BERGEN, J. R.; BURT, P. J.; OGDEN, J. M. Pyramid Methods in Image Processing. *RCA Engineer*, RCA, v. 29, n. 6, p. 33–41, 1984.

ALPERT, S.; GALUN, M.; BRANDT, A.; BASRI, R. Image Segmentation by Probabilistic Bottom-Up Aggregation and Cue Integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 34, n. 2, p. 315–327, Feb 2012.

ATTNEAVE, F. Some Informational Aspects of Visual Perception. *Psychological Review*, APA, v. 61, n. 3, p. 183–193, 1954.

BOLDUC, M.; LEVINE, M. D. A Review of Biologically Motivated Space-Variant Data Reduction Models for Robotic Vision. *Computer Vision and Image Understanding*, Elsevier, v. 69, n. 2, p. 170–184, 1998.

BORJI, A.; CHENG, M.-M.; JIANG, H.; LI, J. Salient Object Detection: A Survey. *arXiv preprint arXiv:1411.5878*, 2014.

BORJI, A.; CHENG, M. M.; JIANG, H.; LI, J. Salient Object Detection: A Benchmark. *IEEE Transactions on Image Processing*, IEEE, v. 24, n. 12, p. 5706–5722, 2015.

BORJI, A.; ITTI, L. Exploiting Local and Global Patch Rarities for Saliency Detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Providence, RI, USA: IEEE, 2012. p. 478–485.

BORJI, A.; ITTI, L. State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 35, n. 1, p. 185–207, 2013.

BRADY, M. Seeds of Perception. In: *Proceedings of the Alvey Vision Conference*. Cambridge, UK: The Alvey Vision Club Committee, 1987. p. 1–8.

BRAUN, J.; KOCH, C.; DAVIS, J. L. *Visual Attention and Cortical Circuits*. London, England: MIT Press, 2001. (Bradford Books).

CARVALHO, G. V.; MORAES, L. B.; CAVALCANTI, G. D. C.; REN, T. I. A Weighted Image Reconstruction Based on PCA for Pedestrian Detection. In: *Proceedings of the Joint Conference on Neural Networks*. San Jose, CA, USA: IEEE, 2011. p. 2005–2011.

CHENG, M. M.; MITRA, N. J.; HUANG, X.; TORR, P. H. S.; HU, S.-M. Global Contrast Based Salient Region Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 37, n. 3, p. 569–582, March 2015.

CHENG, M.-M.; ZHANG, Z.; LIN, W.-Y.; TORR, P. H. S. BING: Binarized Normed Gradients for Objectness Estimation at 300 fps. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, 2014.

COMANICIU, D.; MEER, P. Mean Shift: a Robust Approach Toward Feature Space Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 24, n. 5, p. 603–619, May 2002.

CONNOR, C. E.; EGETH, H. E.; YANTIS, S. Visual Attention: Bottom-Up Versus Top-Down. *Current Biology*, Cell Press, v. 14, n. 19, p. R850–R852, 2004.

CSURKA, G.; DANCE, C.; FAN, L.; WILLAMOWSKI, J.; BRAY, C. Visual Categorization with Bags of Keypoints. In: *Proceedings of the European Conference on Computer Vision (Workshop on Statistical Learning in Computer Vision)*. Prague, Czech Republic: Springer, 2004.

ĆULIBRK, D.; MIRKOVIĆ, M.; ZLOKOLICA, V.; POKRIĆ, M.; CRNOJEVIĆ, V.; KUKOLJ, D. Salient Motion Features for Video Quality Assessment. *IEEE Transactions on Image Processing*, IEEE, v. 20, n. 4, p. 948–958, 2011.

DESCARTES, R. *Passions of the Soul*. Indianapolis, IN, USA: Hackett Publishing Company, Inc., 1649. (Hackett Classics). Translated version by Stephen H. Voss, 1989.

DUAN, L.; WU, C.; MIAO, J.; QING, L.; FU, Y. Visual Saliency Detection by Spatially Weighted Dissimilarity. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Colorado Springs, CO, USA: IEEE, 2011. p. 473–480.

FELZENSZWALB, P. F.; HUTTENLOCHER, D. P. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, Springer, v. 59, n. 2, p. 167–181, 2004.

FINK, G. R.; HALLIGAN, P. W.; MARSHALL, J. C.; FRITH, C. D.; FRACKOWIAK, R. S. J.; DOLAN, R. J. Where in the Brain Does Visual Attention Select the Forest and the Trees? *Nature*, Nature Publishing Group, v. 382, n. 6592, p. 626–628, 1996.

FRINTROP, S. *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search (Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

GLASNER, D.; BAGON, S.; IRANI, M. Super-Resolution from a Single Image. In: *Proceedings of the IEEE Conference on Computer Vision*. Kyoto, Japan: IEEE, 2009. p. 349–356.

GOFERMAN, S.; ZELNIK-MANOR, L.; TAL, A. Context-Aware Saliency Detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. San Francisco, CA, USA: IEEE, 2010.

GOPALAKRISHNAN, V.; HU, Y.; RAJAN, D. Random Walks on Graphs to Model Saliency in Images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, USA: IEEE, 2009. p. 1698–1705.

HALKO, N.; MARTINSSON, P.-G.; TROPP, J. A. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, SIAM, v. 53, n. 2, p. 217–288, 2011.

HALL, P.; MARSHALL, D.; MARTIN, R. Merging and Splitting Eigenspace Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 22, n. 9, p. 1042–1049, 2000.

HAREL, J.; KOCH, C.; PERONA, P. Graph-Based Visual Saliency. In: *Advances in Neural Information Processing Systems*. Vancouver, BC, Canada: MIT Press Cambridge, 2007. p. 545–552.

HOU, Q.; CHENG, M.-M.; HU, X.; BORJI, A.; TU, Z.; TORR, P. Deeply Supervised Salient Object Detection with Short Connections. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, Hawaii: IEEE, 2017. p. 5300–5309.

HOU, X.; HAREL, J.; KOCH, C. Image Signature: Highlighting Sparse Salient Regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 34, n. 1, p. 194–201, 2012.

HOU, X.; ZHANG, L. Saliency Detection: A Spectral Residual Approach. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Minneapolis, MN, USA: IEEE, 2007. p. 1–8.

HYVÄRINEN, A.; HURRI, J.; HOYER, P. O. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. 1st. ed. Longon, UK: Springer, 2009.

INTRILIGATOR, J.; CAVANAGH, P. The Spatial Resolution of Visual Attention. *Cognitive Psychology*, v. 43, n. 3, p. 171–216, 2001.

ITTI, L.; KOCH, C. Computational Modelling of Visual Attention. *Nature Reviews. Neuroscience*, Nature Publishing Group, v. 2, n. 3, p. 194, 2001.

ITTI, L.; KOCH, C.; NIEBUR, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, Los Alamitos, CA, USA, v. 20, n. 11, p. 1254–1259, 1998.

ITTI, L.; REES, G.; TSOTSOS, J. K. *Neurobiology of Attention*. San Diego, CA, USA: Elsevier Science, 2005.

JIANG, B.; ZHANG, L.; LU, H.; YANG, C.; YANG, M.-H. Saliency Detection via Absorbing Markov Chain. In: *Proceedings of the IEEE Conference on Computer Vision*. Sydney, NSW, Australia: IEEE, 2013. p. 1665–1672.

JIANG, H.; WANG, J.; YUAN, Z.; WU, Y.; ZHENG, N.; LI, S. Salient Object Detection: A Discriminative Regional Feature Integration Approach. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.]: IEEE, 2013. p. 2083–2090.

JUDD, T. *Understanding and Predicting Where People Look in Images*. PhD Thesis — Massachusetts Institute of Technology, 2011.

JUDD, T.; DURAND, F.; TORRALBA, A. Fixations on Low Resolution Images. *Journal of Vision*, ARVO, v. 11, n. 4, p. 1–14, 2010.

KANDEL, E. R.; SCHWARTZ, J. H.; JESSELL, T. M. *Essentials of Neural Science and Behavior*. Norwalk, CT, USA: Appleton & Lange, 1995.

KERSTEN, D. Predictability and Redundancy of Natural Images. *Journal of the Optical Society of America A*, OSA, v. 4, n. 12, p. 2395–2400, 1987.

KIM, J.; HAN, D.; TAI, Y.-W.; KIM, J. Salient Region Detection via High-Dimensional Color Transform. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: [s.n.], 2014. p. 883–890.

KIM, J.; HAN, D.; TAI, Y.-W.; KIM, J. Salient Region Detection via High-Dimensional Color Transform and Local Spatial Support. *IEEE Transactions on Image Processing*, IEEE, v. 25, n. 1, p. 9–23, 2016.

KOCH, C.; ULLMAN, S. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Human Neurobiology*, Springer, v. 4, n. 4, p. 219–227, 1985.

KOPF, J.; COHEN, M. F.; LISCHINSKI, D.; UYTTENDAELE, M. Joint Bilateral Upsampling. *ACM Transactions on Graphics*, ACM, New York, NY, USA, v. 26, n. 3, jul. 2007.

LI, G.; XIE, Y.; LIN, L.; YU, Y. Instance-level Salient Object Segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, Hawaii: IEEE, 2017. p. 247–256.

LI, H.; CHEN, J.; LU, H.; CHI, Z. CNN for Saliency Detection with Low-Level Feature Integration. *Neurocomputing*, Elsevier, v. 226, p. 212–220, 2017.

LI, J.; LEVINE, M. D.; AN, X.; XU, X.; HE, H. Visual Saliency Based on Scale-Space Analysis in the Frequency Domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 35, n. 4, p. 996–1010, April 2013.

LI, X.; LU, H.; ZHANG, L.; RUAN, X.; YANG, M.-H. Saliency Detection via Dense and Sparse Reconstruction. In: *Proceedings of the IEEE Conference on Computer Vision*. Washington, DC, USA: IEEE, 2013. p. 2976–2983.

LIE, M. M. I.; BORBA, G. B.; VIEIRA NETO, H.; GAMBA, H. R. Fast Saliency Detection Using Sparse Random Color Samples and Joint Upsampling. In: *Proceedings of the Conference on Graphics, Patterns and Images*. São José dos Campos, SP, Brazil: IEEE, 2016. p. 217–224.

LIE, M. M. I.; BORBA, G. B.; VIEIRA NETO, H.; GAMBA, H. R. Joint Upsampling Random Color Distance Maps for Fast Salient Region Detection. *Pattern Recognition Letters*, Elsevier, 2017. In press.

LIE, M. M. I.; VIEIRA NETO, H.; BORBA, G. B.; GAMBA, H. R. Automatic Image Thumbnailing Based on Fast Visual Saliency Detection. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web*. Teresina, PI, Brazil: ACM, 2016. p. 203–206.

LIE, M. M. I.; VIEIRA NETO, H.; BORBA, G. B.; GAMBA, H. R. Progressive Saliency-Oriented Object Localization Based on Interlaced Random Color Distance Maps. In: *Proceedings of the Latin American Symposium on Robotics*. Curitiba, PR, Brazil: IEEE, 2017. p. 1–6.

LIU, T.; SUN, J.; ZHENG, N.-N.; TANG, X.; SHUM, H.-Y. Learning to Detect A Salient Object. In: *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*. Minneapolis, MN, USA: IEEE, 2007. p. 1–8.

LOWE, D. G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, Springer, v. 60, n. 2, p. 91–110, 2004.

MALAGÓN-BORJA, L.; FUENTES, O. Object Detection Using Image Reconstruction with PCA. *Image and Vision Computing*, Elsevier, v. 27, n. 1, p. 2–9, 2009.

MARGOLIN, R.; TAL, A.; ZELNIK-MANOR, L. What Makes a Patch Distinct? In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA: IEEE, 2013. p. 1139–1146.

MARQUES, O.; MAYRON, L. M.; BORBA, G. B.; GAMBA, H. R. Using Visual Attention to Extract Regions of Interest in the Context of Image Retrieval. In: *Proceedings of the Annual Southeast Regional Conference*. New York, NY, USA: ACM, 2006. p. 638–643.

MIN, D.; CHOI, S.; LU, J.; HAM, B.; SOHN, K.; DO, M. N. Fast Global Image Smoothing Based on Weighted Least Squares. *IEEE Transactions on Image Processing*, IEEE, v. 23, n. 12, p. 5638–5653, 2014.

MOTWANI, R.; RAGHAVAN, P. Randomized Algorithms. *ACM Computing Surveys*, ACM, v. 28, n. 1, p. 33–37, 1996.

NGUYEN, T. V.; ZHAO, Q.; YAN, S. Attentive Systems: A Survey. *International Journal of Computer Vision*, Springer, v. 126, n. 1, p. 86–110, 2018.

NOTHDURFT, H.-C. Salience from Feature Contrast: Additivity Across Dimensions. *Vision Research*, Elsevier, v. 40, n. 10-12, p. 1183–1201, 2000.

NOWAK, E.; JURIE, F.; TRIGGS, B. Sampling Strategies for Bag-of-Features Image Classification. In: *Proceedings of the European Conference on Computer Vision*. Graz, Austria: Springer, 2006. p. 490–503.

OLIVA, A.; TORRALBA, A.; CASTELHANO, M. S.; HENDERSON, J. M. Top-Down Control of Visual Attention in Object Detection. In: *Proceedings of the International Conference on Image Processing*. Barcelona, Spain: IEEE, 2003. v. 1, p. I–253.

OLIVEIRA, S. A.; ROCHA NETO, A. R.; GOMES, J. P. Towards Fixation Prediction: A Nonparametric Estimation-Based Approach through Key-Points. In: *Proceedings of the Brazilian Conference on Intelligent Systems*. Recife, PE, Brazil: IEEE, 2016. p. 391–396.

OUERHANI, N.; BRACAMONTE, J.; HÜGLI, H.; ANSORGE, M.; PELLANDINI, F. Adaptive Color Image Compression Based on Visual Attention. In: *Proceedings of the Conference on Image Analysis and Processing*. Palermo, Italy: IEEE, 2001. p. 416–421.

PARIKH, D.; ZITNICK, C. L.; CHEN, T. Determining Patch Saliency Using Low-Level Context. In: *Proceedings of the European Conference on Computer Vision*. Marseille, France: Springer, 2008. p. 446–459.

PARIS, S.; KORNPROBST, P.; TUMBLIN, J.; DURAND, F. Bilateral Filtering: Theory and Applications. *Foundations and Trends in Computer Graphics and Vision*, Now Publishers, Inc., v. 4, n. 1, p. 1–73, 2009.

PERAZZI, F.; KRÄHENBÜHL, P.; PRITCH, Y.; HORNUNG, A. Saliency Filters: Contrast Based Filtering for Salient Region Detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Providence, RI, USA: IEEE, 2012. p. 733–740.

RAJASHEKAR, U.; CORMACK, L. K.; BOVIK, A. C. Image Features that Draw Fixations. In: *Proceedings of the International Conference on Image Processing*. Barcelona, Spain: IEEE, 2003. v. 3, p. 313–316.

RAZAKARIVONY, S.; JURIE, F. Small Target Detection Combining Foreground and Background Manifolds. In: *Proceedings of the IAPR Conference on Machine Vision Applications*. Kyoto, Japan: IAPR, 2013. p. 1–4.

REINHARD, E.; KHAN, E. A.; AHMET; AKYÜZ, O.; JOHNSON, G. M. *Color Imaging: Fundamentals and Applications*. Wellesley, MA: AK Peters, 2008.

REPIN, U. E. *An Unexpected Visitor*. 1884. Oil on canvas, 160.5 cm×167.5 cm.

SCHMID, C.; MOHR, R.; BAUCKHAGE, C. Evaluation of Interest Point Detectors. *International Journal of Computer Vision*, Springer, v. 37, n. 2, p. 151–172, 2000.

SEO, H. J.; MILANFAR, P. Static and Space-Time Visual Saliency Detection by Self-Resemblance. *Journal of Vision*, ARVO, v. 9, n. 12, p. 1–27, 2009.

SHI, J.; YAN, Q.; XU, L.; JIA, J. Hierarchical Image Saliency Detection on Extended CSSD. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 38, n. 4, p. 717–729, 2016.

THEEUWES, J. Top-Down and Bottom-Up Control of Visual Selection. *Acta Psychologica*, Elsevier, v. 135, n. 2, p. 77–99, 2010.

TOMASI, C.; MANDUCHI, R. Bilateral Filtering for Gray and Color Images. In: *Proceedings of the IEEE Conference on Computer Vision*. Washington, DC, USA: IEEE, 1998. p. 839–846.

TORRALBA, A.; FERGUS, R.; FREEMAN, W. T. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 30, n. 11, p. 1958–1970, 2008.

TRAVER, V. J.; BERNARDINO, A. A Review of Log-Polar Imaging for Visual Perception in Robotics. *Robotics and Autonomous Systems*, Elsevier, v. 58, n. 4, p. 378–398, 2010.

TREISMAN, A. M.; GELADE, G. A Feature-integration Theory of Attention. *Cognitive Psychology*, Elsevier, v. 12, n. 1, p. 97–136, 1980.

TSOTSOS, J. K. Analyzing Vision at the Complexity Level. *Behavioral and Brain Sciences*, Cambridge University Press, v. 13, p. 423–469, 1990.

TSOTSOS, J. K. *A Computational Perspective on Visual Attention*. Cambridge, MA, USA: MIT Press, 2011.

TURK, M.; PENTLAND, A. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, MIT Press, v. 3, n. 1, p. 71–86, 1991.

TUYTELAARS, T.; MIKOLAJCZYK, K. Local Invariant Feature Detectors: A Survey. *Foundations and Trends in Computer Graphics and Vision*, Now Publishers, Inc., v. 3, n. 3, p. 177–280, 2008.

VIEIRA NETO, H.; NEHMZOW, U. Visual Novelty Detection with Automatic Scale Selection. *Robotics and Autonomous Systems*, Elsevier, v. 55, n. 9, p. 693–701, 2007.

VIKRAM, T. N.; TSCHEREPANOW, M.; WREDE, B. A Saliency Map Based on Sampling an Image Into Random Rectangular Regions of Interest. *Pattern Recognition*, Elsevier, v. 45, n. 9, p. 3114–3124, 2012.

WANG, J.; JIANG, H.; YUAN, Z.; CHENG, M.-M.; HU, X.; ZHENG, N. Salient Object Detection: A Discriminative Regional Feature Integration Approach. *International Journal of Computer Vision*, v. 123, n. 2, p. 251–268, 12 2017.

WEI, Y.; WEN, F.; ZHU, W.; SUN, J. Geodesic Saliency Using Background Priors. In: *Proceedings of the European Conference on Computer Vision*. Florence, Italy: Springer, 2012. p. 29–42.

WOLD, S. Pattern Recognition by Means of Disjoint Principal Components Models. *Pattern Recognition*, Elsevier, v. 8, n. 3, p. 127–139, 1976.

YANG, C.; ZHANG, L.; LU, H.; RUAN, X.; YANG, M.-H. Saliency Detection via Graph-Based Manifold Ranking. In: *Proceedings of the IEEE Computer Vision and Pattern Recognition*. Portland, OR, USA: IEEE, 2013. p. 3166–3173.

YANG, J.; ZHANG, D.; FRANGI, A. F.; YANG, J.-y. Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 26, n. 1, p. 131–137, 2004.

YARBUS, A. L. *Eye Movements and Vision*. New York, NY, USA: Plenum Press, 1967.

YILDIRIM, G. *Are All Pixels Equally Important? Towards Multi-Level Salient Object Detection*. PhD Thesis — École Polytechnique Fédérale de Lausanne, 2015.

ZHAI, Y.; SHAH, M. Visual Attention Detection in Video Sequences Using Spatiotemporal Cues. In: *Proceedings of the ACM Conference on Multimedia*. New York, NY, USA: ACM, 2006. p. 815–824.

ZHANG, D.; FU, H.; HAN, J.; BORJI, A.; LI, X. A Review of Co-Saliency Detection Algorithms: Fundamentals, Applications, and Challenges. *ACM Transactions on Intelligent Systems and Technology*, ACM, v. 9, n. 4, p. 38, 2018.

ZONTAK, M.; IRANI, M. Internal Statistics of a Single Natural Image. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Colorado Springs, CO, USA: IEEE, 2011. p. 977–984.