# Intelligent Ad Resizing

by

Anthony P. Badali

A thesis submitted in conformity with the requirements
for the degree of Master of Applied Science
Graduate Department of Electrical & Computer Engineering
University of Toronto

Library and Archives
Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

*Your file Votre référence*
ISBN: 978-0-494-59287-8
*Our file Notre référence*
ISBN: 978-0-494-59287-8

# Abstract

Intelligent Ad Resizing

Anthony P. Badali

Master of Applied Science

Graduate Department of Electrical & Computer Engineering

University of Toronto

2009

Currently, online advertisements are created for specific dimensions and must be laboriously modified by advertisers to support different aspect ratios. In addition, publishers are constrained to design web pages to accommodate this limited set of sizes.

As an alternative we present a framework for automatically generating visual banners at arbitrary sizes based on individual prototype ads. This technique can be used to create flexible visual ads that can be resized to accommodate various aspect ratios. In the proposed framework image and text data are stored separately. Resizing involves selecting a sub-region of the original image and updating text parameters (size and position). This problem is posed within an optimization framework that encourages solutions which maintain important structural properties of the original ad. The method can be applied to advertisements containing a wide variety of imagery and provides significantly more flexibility than existing solutions.

# Contents

**1 Problem Definition & Motivation** **1**

1.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Problem Definition . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2.1 Assumptions & Constraints . . . . . . . . . . . . . . . . . . . . 5

1.3 Notation & Conventions . . . . . . . . . . . . . . . . . . . . . . . . 5

1.4 Statement of Contributions . . . . . . . . . . . . . . . . . . . . . . 7

**2 Background & Prior Work** **8**

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2.2 Image Importance Maps and Saliency . . . . . . . . . . . . . . . . . 9

2.2.1 Itti and Koch Visual Attention Model for Saliency . . . . . . . . . 9

2.2.2 Alternative Saliency Models . . . . . . . . . . . . . . . . . . . . 13

2.2.3 Face and Object Detectors . . . . . . . . . . . . . . . . . . . . . 16

2.2.4 Extracting Semantic Information . . . . . . . . . . . . . . . . . . 17

2.3 Image Retargeting . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

2.3.1 Image Manipulation Techniques for Resizing . . . . . . . . . . . . 18

2.4 Automatic Image Resizing . . . . . . . . . . . . . . . . . . . . . . . 20

2.4.1 Cropping-based Techniques . . . . . . . . . . . . . . . . . . . . 20

2.4.2 Seam-based Techniques . . . . . . . . . . . . . . . . . . . . . . 28

2.4.3 Distortion-based Techniques . . . . . . . . . . . . . . . . . . . . 31

**3 Energy-based Image Importance and Text Relevance 34**

3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

3.2 Coarse Level Center-surround Model . . . . . . . . . . . . . . . . . 35

3.2.1 Surround Sampling . . . . . . . . . . . . . . . . . . . . . . 37

3.2.2 Surround Modeling . . . . . . . . . . . . . . . . . . . . . . 39

3.2.3 Energy-based Center-surround Difference . . . . . . . . . . 40

3.2.4 Post-processing . . . . . . . . . . . . . . . . . . . . . . . . 47

3.3 Text Relevance . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48

3.4 Image Importance . . . . . . . . . . . . . . . . . . . . . . . . . . . 49

3.4.1 Weighted Importance Map . . . . . . . . . . . . . . . . . . 50

**4 A Framework for Cropping-based Poster Resizing 52**

4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52

4.2 Optimization Framework . . . . . . . . . . . . . . . . . . . . . . . 53

4.3 Text Layout . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59

4.4 Constrained Optimization . . . . . . . . . . . . . . . . . . . . . . 61

4.5 Arbitrary Image Resizing . . . . . . . . . . . . . . . . . . . . . . . 63

**5 Experimental Results & Analysis 65**

5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65

5.2 Importance & Relevance Maps . . . . . . . . . . . . . . . . . . . . 66

5.2.1 Qualitative Results . . . . . . . . . . . . . . . . . . . . . . 67

5.2.2 Quantitative Results & Analysis . . . . . . . . . . . . . . . 71

5.3 Automatic Poster Resizing . . . . . . . . . . . . . . . . . . . . . . 78

5.3.1 Qualitative Results . . . . . . . . . . . . . . . . . . . . . . 78

5.3.2 Quantitative Results & Analysis . . . . . . . . . . . . . . . 85

**6 Conclusion 91**

6.1 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 91

6.2 Future Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92

6.3 Final Thoughts . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 93

**Bibliography** **94**

# List of Figures

1.1 Advertising Banner Examples . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.2 Example Poster . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.3 Entropy-based Salient Point Detectors: Examples . . . . . . . . . . . . . . 5

2.1 Itti and Koch Saliency Maps . . . . . . . . . . . . . . . . . . . . . . . . . 12

2.2 Entropy-based Salient Point Detector Examples . . . . . . . . . . . . . . . 14

2.3 Saliency Based on Self-information . . . . . . . . . . . . . . . . . . . . . 15

2.4 Gaze-based Cropping: Image with Eye Fixations . . . . . . . . . . . . . . . 26

2.5 Gaze-based Cropping: Content Maps . . . . . . . . . . . . . . . . . . . . . 27

2.6 Horizontal and Vertical Image Seams . . . . . . . . . . . . . . . . . . . . 29

2.7 Seam Carving Examples . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

2.8 Distortion-based Resizing: Examples . . . . . . . . . . . . . . . . . . . . 33

3.1 Relevance Map Block Diagram . . . . . . . . . . . . . . . . . . . . . . . . 35

3.2 Example Images for the Coarse Center-surround Assumption . . . . . . . 37

3.3 Coarse Center-Surround Sampling . . . . . . . . . . . . . . . . . . . . . . 38

3.4 Example Text Relevance Maps . . . . . . . . . . . . . . . . . . . . . . . . 49

3.5 Example Importance Maps . . . . . . . . . . . . . . . . . . . . . . . . . . 51

4.1 Resizing Solution Parameterization . . . . . . . . . . . . . . . . . . . . . 53

4.2 Resizing Solution Text Region . . . . . . . . . . . . . . . . . . . . . . . . 56

4.3 Optimization Window Weighting Functions . . . . . . . . . . . . . . . . . 56

4.4 Center Weighted Content Term . . . . . . . . . . . . . . . . . . . . . . . 58

4.5 Examples of Line Wrapping Methods . . . . . . . . . . . . . . . . . . . . 61

4.6 Arbitrary Resizing Procedure . . . . . . . . . . . . . . . . . . . . . . . . 63

5.1 General Importance Map Comparison . . . . . . . . . . . . . . . . . . . . 68

5.2 Examples of PCRF and CRF Importance Maps . . . . . . . . . . . . . . 70

5.3 Examples of CRF and ICRF Importance Maps . . . . . . . . . . . . . . . 71

5.4 Example Saliency Masks Used for Evaluation . . . . . . . . . . . . . . . 72

5.5 Normalized Importance Sum for Importance Maps . . . . . . . . . . . . . 73

5.6 Precision Curve for Importance Map Algorithms . . . . . . . . . . . . . . 74

5.7 Precision-Recall Curve for Importance Map Algorithms . . . . . . . . . . 75

5.8 Example Saliency Markers Used for Evaluation . . . . . . . . . . . . . . 76

5.9 Importance Map Localization Error . . . . . . . . . . . . . . . . . . . . . 77

5.10 Poster Resizing Examples Set #1 . . . . . . . . . . . . . . . . . . . . . . 79

5.11 Poster Resizing Examples Set #2 . . . . . . . . . . . . . . . . . . . . . . 80

5.12 Poster Resizing Examples Set #3 . . . . . . . . . . . . . . . . . . . . . . 82

5.13 Poster Resizing Failure Examples . . . . . . . . . . . . . . . . . . . . . . 84

5.14 User Study Results: Overall Selection Percentages . . . . . . . . . . . . . 88

5.15 User Study Results: Pair-wise Selection Matrix . . . . . . . . . . . . . . 89

5.16 User Study Results: Selection Percentages by Target Size . . . . . . . . . 90

# Chapter 1

# Problem Definition & Motivation

## 1.1 Introduction

Over the last two decades the world wide web has emerged as an integral part of society. Among other things, the Internet has become the central hub for the distribution of image, audio, and video content across the globe. As a result, numerous online businesses have emerged in an attempt to capitalize on these new markets. Advertising is closely associated with media, and not surprisingly, online ads have been a primary sector of the Internet economy.

Online advertising is primarily coordinated using ad delivery services such as Google Adsense. A publisher who would like to use AdSense on his or her website simply allocate rectangular regions for advertisements and the system displays ads which maximize the expected revenue. To coordinate the use of ads from many different companies on a website the ad regions are constrained to a set of allowed dimensions. Although more convenient for publishers, this requires advertisers to create multiple versions of each ad. Textual advertisements are particularly appealing in this context because they can be easily modified to fit within standard size ad frames. The increasing popularity of text based ads also suggests that the ability to automatically resize ads is highly desirable.

Image based or visual advertisements represent an important advertising segment because they stand out and can convey much more information than a small textual headings. However, these ads cannot be easily displayed in different shapes and sizes which makes them less convenient to work with. To help stimulate visual ads Google Inc. recently released the Ad Builder tool which allows advertisers to create simple visual ads using templates from which a set of up to six (fixed size) banners can be generated. This is performed by rescaling and repositioning ad content according to a set of pre-specified rules, but ads are still constrained to a small set of standard sizes.

Automatically resized ads are very valuable from the perspective of the web publisher. A publisher's primary interest is to implement the core functionality of a web site, which in turn attracts the traffic required to generate ad revenue. With fixed size ads web content must be designed around ad banner dimensions. A standard set of sizes introduces some flexibility, however advertisers can only provide a limited number of sizes due to budgetary and time constraints. The layouts of many popular websites provide evidence that banner sizes still impose restrictions on web designers. For example, websites such as The Toronto Star and The New York Times contain at least one column width chosen to fit standard ad dimensions.

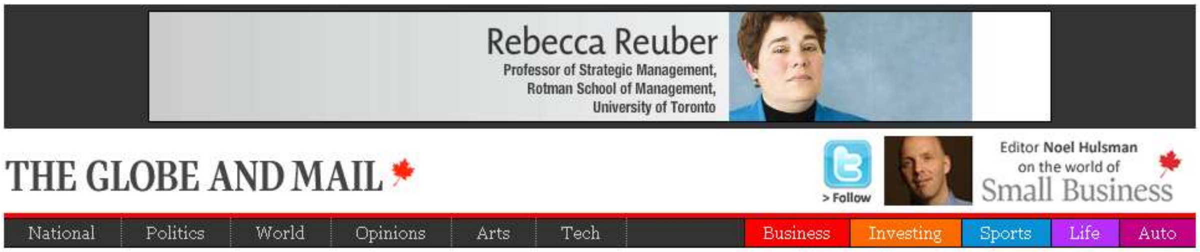


Figure 1.1: A segment of the Globe and Mail website (www.theglobeandmail.com) which demonstrates an advertisement for the Rotman School of Management. In this case the dimensions of the ad do not match those of the allocated region resulting in unused space.

With the development of technologies such as AJAX and DHTML web sites are also becoming increasingly dynamic. In web pages with dynamically changing layout, modifications to ad region sizes can unpredictably impact ad revenues, constraining web designers even further. An ideal ad delivery system would seamlessly accommodate ad regions of arbitrary dimensions without penalizing publishers or requiring costly labor for advertisers. Most importantly, automatic ad resizing technology would allow publishers to monetize space which current banners cannot. An example illustrating unused space on the Globe and Mail website is shown in Figure 1.1.

This work aims to take a step forward in this direction by investigating the idea of automatically generating visual banners at arbitrary sizes from a single prototype ad. This technology would facilitate the use of image-based ads in many more scenarios, increase revenues, and, provide additional flexibility for all parties involved.

## 1.2 Problem Definition

The problem addressed in this work is the automatic resizing of images containing textual headings, which will be referred to as posters or ads. An example of a rendered poster is shown in Figure 1.2.

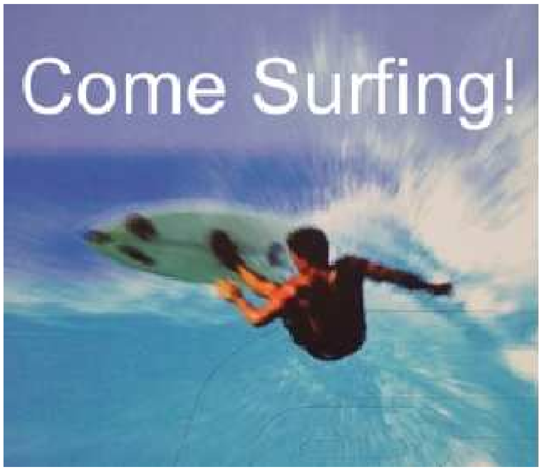


Figure 1.2: An example poster. Image content based on an actual advertisement for www.earthsports.com.

Formally a poster **P** is a set which consists of the following items,

$\mathbf{I}$ - An $m \times n \times 3$ color image which contains the poster's image content.

$s_0$ - The text object size for the original poster.

$\mathbf{c}$ - A heading position vector $(x, y)$ which specifies the pixel coordinates of the text center in image $\mathbf{I}$.

$\mathbf{\Gamma_c}$ - Constant heading section information. This includes the specification of a font type, text message string, and word wrapping technique. These parameters are all required for dynamically computing the poster text region and influence the resizing solutions.

Based on this definition the textual information within a poster is stored separately from the image content and poster images are rendered by generating the text content from $\mathbf{\Gamma_c}$ at size $s$ and combining it with the image content at position $\mathbf{c}$. The most important property of the poster related to resizing is the aspect ratio $\rho$ which is defined as the ratio of columns to rows in $\mathbf{I}$.

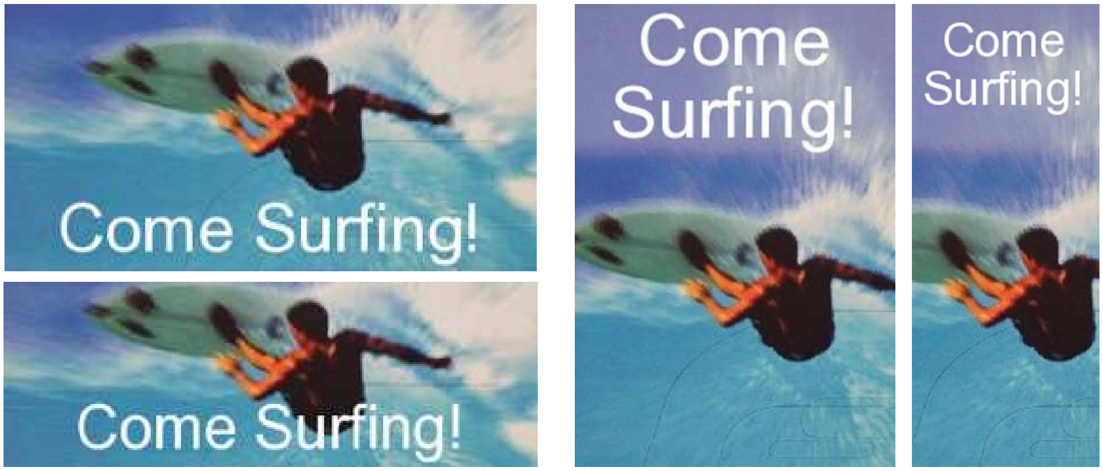
$$\rho = \frac{n}{m} \tag{1.1}$$

An instance of this problem involves an example poster $\mathbf{P}$ and a scaling parameter $\gamma \in [0, 1]$ and produces an output poster $\mathbf{P}'$ for which the horizontal (column) or vertical (row) dimension has been reduced by a factor of $\gamma$.

There are two main sub-problems which must be solved in order to produce the new poster $\mathbf{P}'$ :

1. Selecting a subregion of $\mathbf{I}$ which best represents the original image.
2. Adjusting the text object parameters $s_0$ and $\mathbf{c}$ for the new poster.

This problem is difficult because the parameters related to text and image content influence one another. For example, the best text position will vary depending on the sub-window location. A satisfactory resizing solution must carefully select text and content parameters while respecting these interrelations.



(a) (b)

Figure 1.3: Automatically Resized Posters. Examples of (a) Row reduction and (b) Column reduction.

### 1.2.1 Assumptions & Constraints

Generally speaking, the creation of an advertisement is an artistic endeavorer. Designers select or create an image, compose a textual message, and combine the two. However, the goal of this work is not to develop a computer system which makes artistic decisions, this would be too subjective and nearly impossible to formalize. The aim here is to analyze and extract information from a prototype poster, and then use this knowledge to construct a resized version which maintains important structural properties of the original. This approach imposes restrictions on the configuration of prototype posters and resizing results. However, it is still applicable to a wide range of ads and can be easily modified to accommodate various poster structures.

## 1.3 Notation & Conventions

The following mathematical notation will be used in subsequent sections. Scalars will be denoted with lowercase letters, e.g. $\gamma$. Vectors with lowercase bold letters, e.g $\mathbf{u}$. Matrices, sets, and multi-dimensional images will be capital, and sometimes boldface for clarity, e.g. $S$ or $\mathbf{A}$. Images and matrices may be indexed by lowercase letters, such as $\mathbf{A}(i, j)$, or as $\mathbf{A}(\mathbf{p})$, where $\mathbf{p}$ is a 2D position vector.

The following table lists identifiers for frequently occurring quantities or items.

Common Identifiers

| | |
|---|---|
| $\mathbf{P}$ | A poster or ad |
| $\gamma$ | Rescaling factor $\in (0, 1)$ |
| $\mathbf{A}$ | Energy-based CS difference map |
| $\mathbf{I}$ | Image (color or arbitrary features) |
| $C(\cdot)$ | Resizing solution cost function |
| $S(\cdot)$ | A saliency or importance map |
| $R(\cdot)$ | A text relevance map |
| $\boldsymbol{\Omega}$ | A poster resizing solution |
| $\boldsymbol{\Gamma}$ | A poster's text parameter set |
| $\mathbf{R}$ | A rectangular (cropping) region of an image |

The following table contains a list of the most common acronyms which appear in the subsequent sections.

Common Acronyms

| | |
|---|---|
| AAR | Automatic Ad Resizing |
| AIR | Automatic Image Resizing |
| CRF | Condition Random Field (Saliency Method) |
| CCS | Coarse Center Surround |
| CS | Center Surround |
| IK | Itti & Koch (Saliency Method) |
| ROI | Region of Interest |

## 1.4 Statement of Contributions

The primary objective of this work was to develop a complete algorithm for automatically resizing image-based posters or advertisements. A complete solution to this problem required several issued to be addressed resulting in the following contributions,

1. An optimization framework for determining parameters of resized ads. Key practical issues such as constraining search space and automatic text wrapping are also addressed within this solution.

2. A novel saliency map algorithm for estimating image importance maps using an energy-based model.

3. A set of evaluation metrics which can be used to evaluate importance maps algorithms for resizing problems.

The energy-based importance maps are shown to yield significantly better results than standard algorithms and the resizing technique presented here outperformed conventional methods by over 20% in a user study. Overall, the results demonstrate that within appropriate constraints, automatic poster resizing can be performed with much higher fidelity than conventional image resizing.

# Chapter 2

# Background & Prior Work

## 2.1 Introduction

The problem of automatic poster resizing has not previously been discussed in the common literature, however the more general problem of image resizing is closely related and will be the emphasis of this chapter.

In general, an automatic image resizing (AIR) algorithm takes an input image $I$ with pixel dimensions $m \times n$ and produces an output image $I'$ of dimensions $m' \times n'$. The dimensions of the output image may be specified as input parameters or selected as part of the algorithm based on a criteria such as content retention. Within the domain of automatic image resizing there are three key subproblems which must be addressed,

1. Identification - The construction of maps to identify important or key regions of the image. Common examples include gradient energy and saliency maps [28].

2. Reshaping - Developing or selecting an algorithm for reducing the size (number of pixels) in an image to the target size. Examples include rectangular cropping [6][51] and seam removal [1].

3. Selection - Determining which pixels of the image (or what information) will be

discarded. The technique used here will depend significantly on the reshaping method and importance maps but is usually formulated as an optimization problem [6][43][45][51].

The majority of research in this area has focused on subproblems 2 and 3 whereas, subproblem 1 has generally been addressed by applying existing techniques from the human visual psychology [28] and object recognition [55] communities. The goal of this section is to provide background on these research areas. Section 2.2 introduces popular techniques for addressing the identification problem by computing saliency or importance maps. The remainder of the chapter presents material related to image resizing itself which relates directly to the shaping and selection problems. Section 2.3 provides background into the basic image resizing operation and applications of interest. Recently published algorithms for automatic image resizing are presented in Section 2.4.

## 2.2 Image Importance Maps and Saliency

The most challenging problem related to automatic image resizing is the computation of importance maps to identify the primary objects in an image. This problem has been typically addressed by applying techniques from the visual psychology and object detection communities. Specifically, low-level saliency models of visual attention and face detectors. This section provides an overview of these techniques and discusses their relationship with the semantic concept of image importance.

### 2.2.1 Itti and Koch Visual Attention Model for Saliency

In [26][27][28] a bottom-up model for computing topographical saliency maps is developed and analyzed. This model is intended to mimic rapid and task-independent scene analysis which is performed during the early stages of visual attention. This corresponds to pre-attentive non-voluntary visual attention and does not include any top-down semantic information. It is arguably the most popular model of low-level attention in the literature.

The model is based on the concept of ”center-surround” which is founded on the idea that objects that differ from their immediate surroundings are more likely to be visually salient. Computationally, each location in an image is characterized by a set of local features related to color, intensity, and edge orientation. Center-surround response is then computed by taking differences between features responses at different scales. This measures the degree to which each feature value differs from the average in its local neighborhood.

For each pixel of the input image intensity, color, and orientation visual features are computed. An intensity image $I$ is obtained by linearly combining the image color channels and applying non-linear normalization. To yield a multiscale representation a Gaussian pyramid [20] $I(\sigma)$ is computed for $\sigma \in [0..8]$ (1:1 to 1:256 scale). Color features are computed by transforming the image into a four channel opponent color space and computing Guassian pyramids for each channel. This yields four color maps, $R(\sigma)$, $G(\sigma)$, $B(\sigma)$, and $Y(\sigma)$. Finally local orientation images are computed using Gabor filters [20][21] which are described by

$$\mathrm{G}(x, y, \sigma, \theta) = \exp\left(\frac{-\left(x_\theta^2 + y_\theta^2\right)}{2\sigma}\right) \cos\left(2\pi \cdot x_\theta \lambda_\theta\right) \tag{2.1}$$

where,

$$x_\theta = x\text{cos}\theta + y\text{sin}\theta \quad \text{and} \quad y_\theta = -x\text{sin}\theta + y\text{cos}\theta$$

The intensity images are convolved with Gabor filters at four different orientations and expanded into Gaussian pyramids, $O(\sigma, \theta)$ where $\theta \in \Theta = \{0, \pi/4, \pi/2, 3\pi/2\}$ and $\sigma \in [0..8]$.

A set of center-surround maps is then computed for each feature type. The center-surround difference is a binary operator between a pair of images at difference scales, it is denoted by $\ominus$ and corresponds to resampling the coarser map to the finer scale and performing pixel-wise subtraction. Sets of intensity maps are computed as, $I(c, s) =$

$|I(c) \ominus I(s)|$. Two sets of color double opponency maps are computed as $RG(c,s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|$ and $BY(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|$. Orientation maps are computed as $O(c,s,\theta) = |O(c,\theta) \ominus C(s,\theta)|$. Using $c \ \in \{2,3,4\}$ and $s = c + \delta$, $\delta \in \{3,4\}$ results in 6 intensity maps, 12 color maps, and 24 orientation maps.

The final step in the saliency map creation process is the feature map combination step. This process is somewhat complicated because quantities in different types of feature maps cannot be easily converted to into a common and comparable unit. The proposed method for normalization attempts to enhance maps which contain a small number of relatively large peaks while suppressing maps which contain many (large or small) peaks of comparable magnitudes. The map normalization operation is denoted $N(\cdot)$ and consists of the following steps.

1. Normalize the input map to a fixed range, $[0, R]$.
2. Find the global maximum of the map, $M$, and compute the average of all other local maxima, $\overline{m}$.
3. Multiply all quantities in the map by $(M - \overline{m})^2$

For maps with numerous maxima of compatible magnitudes the quantity $(M - \overline{m})^2$ will be small and act as a global inhibitor. Once normalized, the feature maps are combined by resampling them to a common scale (size) and performing pixel-wise addition. This 'across-scale addition' operation is denoted by "$\oplus$". Using the above, feature maps of each type (intensity, color, and orientation) are combined together into 3 "conspicuity maps".

$$\overline{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(I(c,s)) \tag{2.2}$$

$$\overline{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(RG(c,s)) + N(BY(c,s)) \tag{2.3}$$

$$\overline{O} = \sum_{\theta \in \Theta} N\left( \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(O(c,s,\theta)) \right) \tag{2.4}$$

Finally, a single saliency map is computed by normalizing and summing the feature conspicuity maps

$$S = \frac{1}{3}\left(N(\overline{I}) + N(\overline{C}) + N(\overline{O})\right) \tag{2.5}$$

Although the above procedure is primarily based on heuristics, each step in the computation is biologically inspired. This model has also been shown to perform similarly to the human visual system for certain tasks. Specifically, it can reproduce visual pop-out effects and the maxima of saliency maps have also been correlated with human eye fixations [41]. Moreover, the peaks of the saliency maps are frequently located on or near important objects in images. Examples of saliency maps are shown in Figure 2.1. One

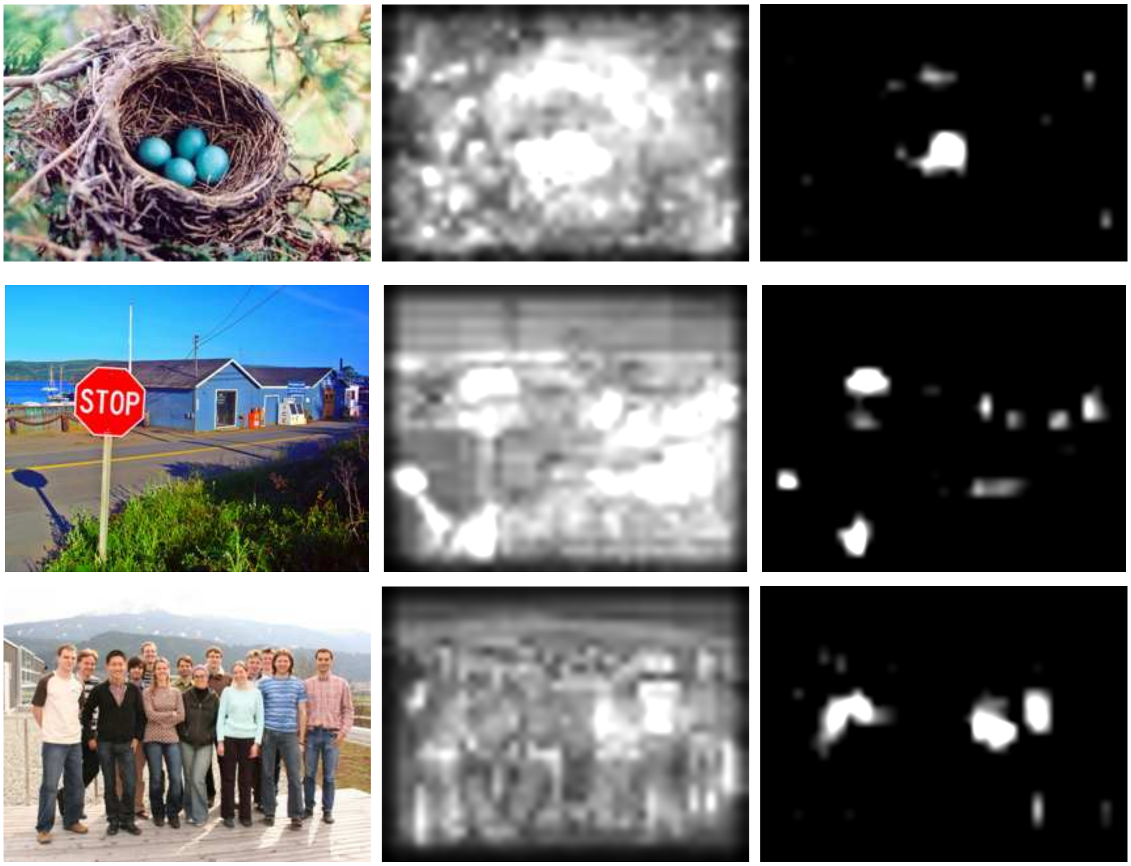


Figure 2.1: Examples of saliency maps created using Itti and Koch's method. (Left) Original Images (Center) Maps creates using local maxima normalization (Right) Maps created using iterative normalization.

undesirable property of saliency maps created using the above procedure is that peaks are not always very strong and can be broad. To resolve this issue an alternative normalization operator based on local competition between neighboring locations was defined [27]. This operator is referred to as 'iterative normalization' and yields sparse saliency maps with large peaks as shown in Figure 2.1. The iterative procedure involves repeated convolution and accumulation operations where at each step convolution is performed using a difference of Gaussian filter. This filter mimics a neural architecture which exhibits a strong local response in competition with broad inhibition from neighboring locations.

### 2.2.2 Alternative Saliency Models

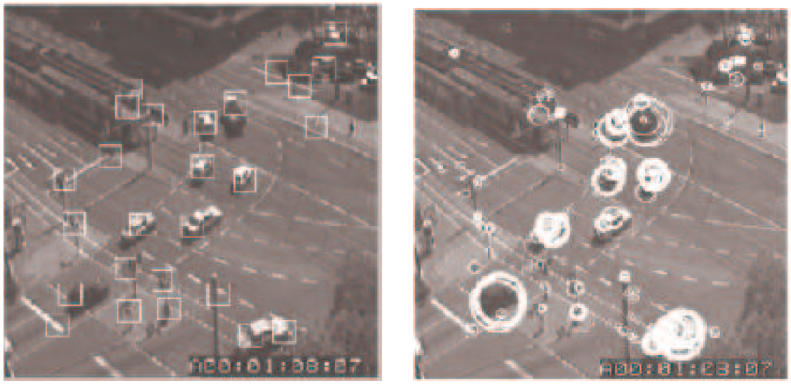
Saliency and visual attention are well studied problems within computer vision and visual psychology. Several other algorithms for computing saliency maps have been proposed within both communities, and a large fraction of these are explicitly or implicitly based on the center-surround paradigm. This concept is very general and the main difference between algorithms within this class is the how exactly it is defined and measured. Methods of this type are also sometimes referred to local contrast-based techniques.

#### Information Theoretic Models

Although highly successful, the Itti and Koch saliency model is primarily based on heuristics, and several researchers have developed information theoretic and probabilistic models to formalize the concept of image saliency.

In [19] Gilles defines saliency in terms of local image complexity, where the entropy [40] of local image intensity is used to measure saliency. The author argues that salient regions are portions of the image which have large variations in intensity. The entropy of the distribution of intensity values over a neighborhood around each pixel is a suitable measure for complexity.

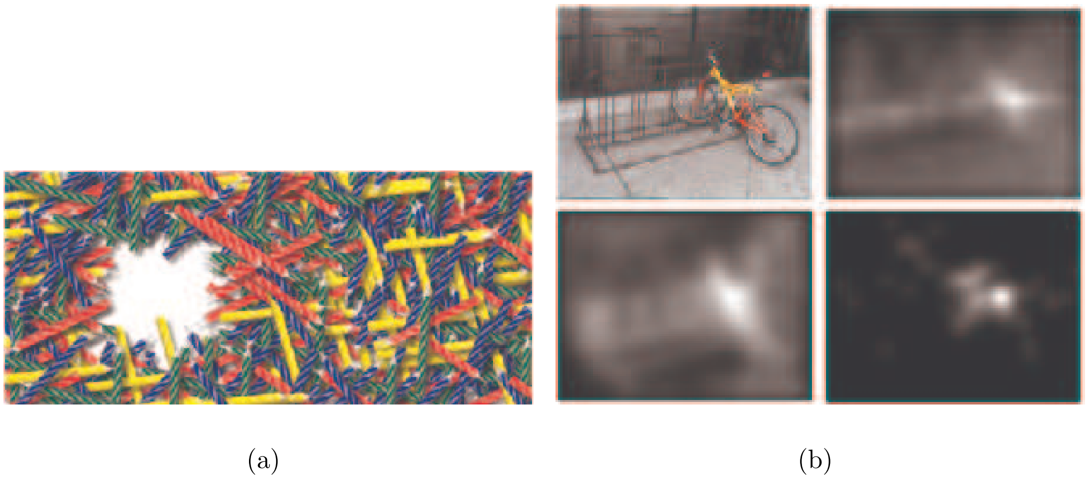
Kadir and Bruce [29] develop a multiscale version of this model by locally selecting scales at which the entropy is peaked. For each peak in scale space a saliency value is computed as the product of the entropy and a derivative (or center-surround) term related to the change the intensity PDF between neighboring scales. Examples of both Gilles and Brady salient point detectors are shown in Figure 2.2.



(a) (b)

Figure 2.2: Examples of salient points computed using entropy-based saliency point detectors from [29]. The size of each marker indicates the scale at which each point was detected. (a) Fixed scale Method [19] (b) Multiscale Approach [29]

In [54] Bruce and Tsotsos argue that self-information is a more appropriate metric than entropy (or local complexity) for modeling bottom-up attention. The self-information I(·) is defined as the negative log probability of an outcome (or observed value) of a random variable X, I($x$) = −log(p(X)). The difference between entropy and self-information is intuitively demonstrated using the example image shown in Figure 2.3(a). Within this image the point of visual fixation is the uniform patch, however this is the area of least entropy (or complexity). This demonstrates an example where the saliency of an image region is a function of local contrast. Within this model the image is decomposed fixed (7x7) RGB patches and the saliency of each patch is inversely proportional to it's probability given the distribution of surrounding intensity values. To account for the complexity of estimating local high-dimensional probability distribution functions the image data is represented in an alternative basis frame computed using independent component analysis [25]. In this reference frame, basis coefficients are statistically independent allowing the probability of a given patch to be computed as the product of probabilities over individual basis coefficients. The probability distribution for each basis coefficient is modeled using a Gaussian kernal density estimate [3]. An example saliency map computing using this method is shown in Figure 2.3(b), results are similar Itti and Koch saliency maps [28].



(a) (b)

Figure 2.3: Images demonstrating the concept of saliency based on self-information from [54] (a) An image demonstrating the benefits of self-information over entropy (b) Example saliency maps. (Top Left) Original Image (Top Right) Itti and Koch Method (Bottom Left) Self-Information Method [54] (Bottom Right) Fixation map computed from human eye tracking data.

### Interest Point Detectors

Interest-point detectors have been very popular within the computer vision community for detecting salient key points in images. They are typically based on local geometric features such as edges or corners, the classic example being the Harris Corner detector [23]. Other examples include symmetry detectors [42] which measure symmetry using the magnitude and orientation of gradients in a local neighborhood. An example of a state-of-the-art interest point detector is the SIFT algorithm [35] which selects the local extrema in scale-space by taking differences between several nearby scales of a Gaussian pyramid. This operation resembles Itti and Koch's method [28] using only intensity features and no normalization operation.

In general, interest point detectors are employed to find a relatively small number of points or regions in an image which are useful for performing object detection and recognition [12]. They are usually designed to maximize two properties known as information content and repeatability [47]. Information content corresponds to the distinctiveness or uniqueness of local descriptors computed at interest points. Repeatability measures of persistence of detected interest points under changes of object imaging conditions such as lighting and viewing angle. Although relevant to object recognition tasks, these criteria do not address the requirements of image resizing importance maps. Within the context of visual attention interest point detectors can be considered local-complexity based techniques and are likely to produce very noisy importance maps. For this reason they have not been employed for image resizing problems.

### 2.2.3 Face and Object Detectors

To address the purely bottom-up nature of saliency models most automatic resizing techniques employ a face detector to explicitly incorporate high-level semantic information into the resizing process. Face detection is an instance of the more general target detection problem for which many highly effective techniques have been developed. For example, state-of-the-art face detectors have recognition rates above 90% [48][55]. Face detectors can be easily integrated with image importance models by setting regions corresponding to faces to have large saliency values. This will generally improve the performance AIR algorithms when applied to images which contain semantically important faces, but these benefits are limited to a small class of imagery.

### 2.2.4 Extracting Semantic Information

In general, image importance is a subjective measure which can vary depending on the individual viewing an image. Under ideal circumstances importance maps would be constructed using high-level semantic information just as humans do. One of the major challenges within this domain is the *semantic gap*, which relates to the challenges of extracting high-level semantic information from low-level computational representations. Broadly speaking, this refers to the problem of going from image pixels to a scene or object description [22]. The semantic gap is manifested at various steps throughout the recognition process; when going from image pixels to labels (e.g segmentation and grouping), and also when classifying labels into objects or scenes. Current object recognition technologies cannot reliably perform these tasks on a general level [12]. Still, bottom-up techniques have been successfully used for many problems such as content-based image retrieval and object recognition [4][52]. These methods often leverage large-training datasets or are designed for a small number of specialized classes (e.g frontal face detectors) to partially overcome the semantic gap.

The most popular image saliency models are also based on low-level image descriptor statistics [17][28][54] and the semantic gap is a serious issue. Furthermore, the subjectiveness of image saliency makes it difficult to create even small ground truth data sets for training. This suggests specialization is the best approach for tackling semantic problems. For image resizing, face detection represent a simple approach that has been used for creating specialized importance maps. However, this does not directly address the construction of saliency maps and it may be fruitful to investigate alternative types of specialization which apply to a wider class of object categories. For example, the additional text region information in **P** can be used to create specialized importance maps for poster resizing. The degree of specialization will be constrained by the applicability of importance maps to a wide range of image content and there will still be limitations when images do not comply with model assumptions. However, given current technologies, this is arguably the easiest and most effective way to cope with the semantic gap.

## 2.3 Image Retargeting

Recently, there has been an increasing amount of research work in the area of automatic image resizing. This work is primarily driven by the important role of digital media, specifically digital images, in society [1]. There are two main applications associated with image resizing. The primary and most researched application, commonly referred to as retargeting, involves adapting images for viewing on small displays [33][49]. This application is motivated by the ubiquity of handheld computing devices such as cellular phones and the evolution of communication infrastructure towards web, image, and video content. The second application is related to efficiently displaying large image collections for human visual searches [51]. This is motivated by the recent popularity of large online photo-sharing communities such as Flickr, Facebook, and MySpace.

### 2.3.1 Image Manipulation Techniques for Resizing

Image resizing is a simple operation which involves changing the pixel dimensions of a digital image. There are two main approaches which can be taken for manual image resizing, cropping and re-sampling. Cropping refers to selecting a contiguous rectangular sub-region of the desired size from the original image. This is very easy task for humans, however poses challenges for computer systems because of the inability of current technologies to extract the semantic information required to select an appropriate image sub-region. The second approach, known as re-sampling, corresponds exactly to the operation of digital signal re-sampling. This is a well-defined process which can be easily applied to images, however it is only applicable when the aspect ratio of the image is not being changed.

### Image Cropping

As stated above, image cropping involves selecting a contiguous region rectangular subregion of an image. The cropping rectangle can be specified by a set of 4 parameters, $\mathbf{R} = \{x_0, y_0, w, h\}$, where $(x_0, y_0)$ is the top-left corner of the cropping window and $w$ and $h$ are its width and height dimensions respectively.

### Downsampling

Image downsampling is a common operation typically used to reduce the information content of an image to decrease storage or communication bandwidth requirements. In the context of image re-targeting, this operation is often required for converting image resolutions to a range suitable for display on hand held devices.

Formally, the image downsampling operation can be performed on an image $\mathbf{I}$ to yield an image $\mathbf{I}'$ where the row and column dimensions have both been reduced by a factor of $\gamma \in (0, 1)$ . The maximum detectable spatial frequency component in the original image is $\pi$ radians/sample [39], however $\mathbf{I}'$ is sampled at a coarser interval so we cannot retain frequencies components above $\pi/\gamma^{-1}$. To ensure re-sampling without aliasing we must first apply an anti-aliasing low-pass filter to $\mathbf{I}$. Popular choices for image re-sampling include Gaussian filters and Lanczos-windowed sinc functions [**?**].

The down sampled image is computed by uniformly sampling the anti-aliased image at intervals of $\gamma^{-1}$ (e.g for $\gamma = 0.5$ we sample every $2nd$ pixel). In general, this will involve estimating color intensity for non-integer pixel coordinates which can be performed using various methods. The simplest approach is nearest-neighbor interpolation which simply rounds non-integer pixel coordinates to the nearest pixel in the image. Better results can be achieved through bicubic interpolation, which estimates pixel intensity as a linear combination of the $4 \times 4$ neighborhood around the point of interest [30].

## 2.4 Automatic Image Resizing

As discussed above there are a few straightforward operations which can be used to resize images. In general, these are sufficient for most manual resizing tasks. However, it is much more challenging to develop systems which can autonomously decide when and how much an image can be resized without reducing its perceived fidelity. This section provides an overview of various automatic image resizing systems which have been developed over the last few years.

### 2.4.1 Cropping-based Techniques

The most popular technique for general image re-targeting is cropping, and the majority of automatic resizing algorithms have used this method for reshaping. This operation is favorable because of it's simplicity and intuitive nature - it is the natural method for changing the aspect ratio of an image. This section presents four different cropping based AIR system which have been recently developed.

#### Image Attention Model for Adapting Images on Small Displays

In [6] Chen et. al. develop a general framework for automated image resizing which is extended for various applications in [15][34][58]. This work is significant because it is one of the earliest methods of it's kind in the literature and highlights many key issues which must be addressed to solve this problem.

A general 'image attention model' based on primitives termed attention objects is presented. Each attention object (AO) corresponds to a portion of the image which would capture a viewer's attention. Usually, this corresponds to a semantic object in the image such as a face or house, but can also apply to arbitrary focal points such as textual headings. Attentional objects have three defining attributes; region of interest (ROI), attention value (AV), and minimum perceptible size (MPS). The ROI is defined as the spatial region in the image which corresponds to the AO. The AV is used to quantify the importance of an AO in terms of visual attention. The MPS represents the minimum allowed spatial size of each AO, it is used as a threshold for the amount of sub-sampling which can be applied to the image without severely reducing the perceptibility of the AO.

One challenging aspect of the image attention model is the subjective nature of the attention objects and their respective parameter values. In the simplest case, models can be manually constructed by the image publishers. For automatic model generation the authors employ visual feature analysis techniques along with a set of heuristic rules. Automated attention modeling is achieved via a combination of a Itti and Koch saliency map [28] (discussed in Section 2.2.1), face detector [48], and text detector. The saliency maps provide a brightness, $B(i, j)$, value for each pixel which corresponds to the relative amount attention a region should receive. This is converted to a attention value by summing saliency within the ROI,

$$AV_{sal} = \sum_{(i,j) \in R} B(i, j) \cdot W_{sal}^{i,j} \tag{2.6}$$

where $W_{sal}^{i,j}$ is a normalized Gaussian template centered in the image. The use of a Gaussian template is very common within the literature and is intended to model the prior believe the salient objects are more likely to be located at the center of the image. Detected faces and text regions are converted into attention values based on the observation that the importance of a detected object is related to its size,

$$AV_{obj} = \sqrt{Area_{object}} \cdot W_{obj}^{pos} \tag{2.7}$$

where $W_{obj}^{pos}$ is a weighting function based on spatial location for faces and a weighting based on aspect ratio for text. The AV for an AO is computed as a linear combination of the individual saliency, face, and text attention values. The MPS value is heuristically set based on the size of the detected ROI and the types of objects detected within it.

An image is automatically retargeted by maximizing an objective function termed the information fidelity (IF). By definition a retargeted image is computed by selecting a subregion **R** of the original image and then downsampling it to a target size T. The information fidelity of a subregion is the sum of AVs for all attention objects within **R** that remain above their MPS when resized to T.

$$IF(R) \quad = \sum_{ROI_i \subset R} AV_i \cdot \mathrm{u}(r_R^2 \cdot \mathrm{size}(ROI_i) - MPS_i) \tag{2.8}$$

where

$$r_R = \min\left(\frac{Width_T}{Width_R}, \frac{Height_T}{Height_R}\right) \tag{2.9}$$

is the ratio of rescaling by down sampling which must be applied for R to reach the target size, and $\mathrm{u}(x)$ is the continuous time unit step function [38]. A re-targeting which maximizes (2.8) can be efficiently found using integer programming techniques.

The visual attention model and search framework above introduce many important concepts which frequently occur in many many AIR algorithms however this particular technique has several limitations which limit it's performance in practice. The model critically depends on the ability to estimate the ROI of each AO in a scene, which is a very challenging problem in the general case. For images where faces or text are not present it will be extremely difficult to accurately estimate ROI size. The IF as an optimization metric is also limited because it does not account for AOs which are partially included in a resizing solution (they contribute a zero value to the IF) which significantly constrains the solution space.

### Greedy Automatic Thumbnail Cropping

In [51] Suh et. a.l. propose a method for automatic resizing images to thumbnail size for quickly browsing through large collections of images. Examples where this problem arises includes online photo sharing communities and image search engines. Similarly to the image attention model in [6] the important regions of an image are detected by employing Itti and Koch saliency maps [28] in addition to a face detector [48]. The iterative normalization operator is applied to the saliency maps to enhance local maxima and suppress all other regions [27]. Example saliency maps created using the iterative normalization operator are shown in Figure 2.1.

One significant difference between the problem statement here and the image attention model [6] is that the resized window size is not an input parameter. Instead, the goal is to find a cropping rectangle, $R_c$, which is as small as possible but still contains the main or important objects in the image.

Under the assumption that the interest points in the saliency maps will be located on or near the important image content the set of candidate cropping rectangles $S(\lambda)$ is defined as

$$S(\lambda) = \left\{ r : \frac{\sum\limits_{(x,y)\in\, r} S_I(x,y)}{\sum\limits_{(x,y)} S_I(x,y)} > \lambda \right\} \tag{2.10}$$

which denotes all possible subwindows that contain more than a fixed fraction $\lambda$ of the total image saliency. An optimal cropping window $R_c$ is then selected as a subwindow $r$ in $S(\lambda)$ with the minimum area.

$$R_c = \operatorname*{argmin}_{r \in R(\lambda)} \{\text{Area}(r)\} \tag{2.11}$$

There are two challenges with the above optimization problem. First, a threshold $\lambda$ must be selected and the most effective value varies across images. Secondly, the number of possible search windows is quadratic in the number of pixels which results in time consuming searches.

The authors propose a greedy algorithm for finding the optimal window and then a method to dynamically adjust $\lambda$. This algorithm is based on the form of the saliency map - that it a sparse map with highly peaked interest points. To prune the search space only the rectangles which contain peaks of the saliency map are considered. The optimal window $R_c$ is computed incrementally by starting with a small window centered in the saliency map and incrementally adjusting the size of the window to include the next salient peak (and a small neighborhood around it). A window $R_c$ which satisfies a fixed threshold can be found by incrementally adding peaks until the saliency sum is greater then $\lambda$.

A suitable value for $\lambda$ is adaptively selected based on the heuristic that it is best to set $\lambda$ just below the point when adding a small amount of saliency (or peak) results in a large increase in the cropping window size. In terms of image content, this may correspond to the point where a secondary object is being included in the crop window. Mathematically, this is the point where the gradient of the function of cropping window size to percentage saliency included is maximized. Images containing faces are handled as a special case. When any faces are detected the image is automatically cropped to include all of them.

Two user studies were performed to evaluate this technique for two tasks related to image retrieval and browsing systems. In a task involving object recognition in small thumbnails this method was shown to improve human recognition rates between 10 and 20 percent compared to standard resizing. Similarly, in a visual search task where participants were asked to search over a set of thumbnail images to find a particular image the average search times were 15 to 20 percent less using automatic cropping.

### Contextually Adaptive Image Cropping

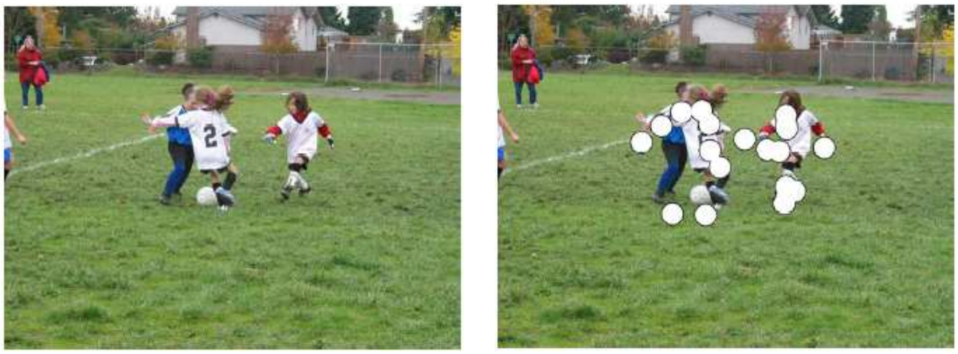
The authors of [8] develop a technique for automatic image cropping where images are first classified so class-dependent cropping heuristics can be used to perform retargeting. For the purpose of classification image descriptors are computed from low-level features related to color, texture, edge intensity, and image composition [46]. Images are then classified into one of three possible categories using ensembles of decision tree classifiers. A decision tree is a model which applies sequence of binary decisions until the set of possible classes is reduced to a single candidate. The well-known classification and regression tree (CART) methodology [3] can be used to construct these trees from a training dataset. In the implementation discussed here a large number of decision trees are constructed image features and the individual classifications are combined to a single decision using a majority voting scheme.

Images can be classified as landscape, close-up, or other. For images in the ‘landscape’ class minimal cropping is performed, the image is simply resized to target dimensions. For images in the ‘close-up’ class a procedure terms relevant regions detection and analysis is performed. This involves computing an Itti and Koch saliency map [28], binarizing the map and discarding areas smaller than a threshold, and finally computing a bounding box which contains all salient regions which remain in the saliency map. The image is then resized to the target dimensions with respect to this region. The images in the ‘other’ class are analyzed with a face detector [55]. If a face is not detected then a variant of the relevant regions detection procedure above is executed to find a cropping window. For images which contain faces a skin detector is used to compute a binary skin color map [5]. The skin color map is combined with the saliency map and a relevant regions detection procedure is executed to find a target region which will be adapted to the desired size.

A small user study was performed to evaluate the above algorithm against the a rescaling (down sampling) approach for re-targeting images for small displays. The results showed that only 7% of the resized images were judged worse using this technique, while 53% of the images were considered better than those retargeted using rescaling. A majority of the 40% judged to be equivalent were in the ’landscape’ class for which no cropping is performed.

### Semi-Automated Cropping using Eye Tracking

In [45], an interactive technique for automatically cropping images is developed. To perform cropping the eye-fixations of a human participant are recorded and used to construct a ‘content map’ which serves the same role as a saliency or importance map in other techniques. Using the content map, a resizing is found by optimizing an objective function. This technique is relevant in the context of automated image targeting because it demonstrates a resizing technique which is applicable when content can be reliably identified.



(a) (b)

Figure 2.4: From [45] (a) Original image and (b) Image with fixation locations shown in white.

For each image eye fixation data was collected for 10 seconds using an eye tracking system and asking viewers to 'find important subject matter in the photo'. An example containing fixation locations is shown in Figure 2.4. Input images are divided into small regions by performing a graph-cut based over segmentation. Each fixation is weighted based on the duration of time spent examining a region, and results are summed to yield scores for each region. Initial foreground and background labels are then assigned to the regions with scores in the upper 10th percentile and lower 50th percentile respectively. Using the initial labels a second graph cut optimization is executed which estimates labels for unknown regions and reduces noise in the original labels. The region boundaries are then smoothed to give importance to regions in the neighborhood of the segmented object and this final image is used as the content map. Example images showing each step of this process for the image in Figure 2.4(a) are shown in Figure 2.5. It is also clear from the example that the initial estimates from fixation points are well localized.

An optimal window of a given aspect ratio is then found by minimizing an objective

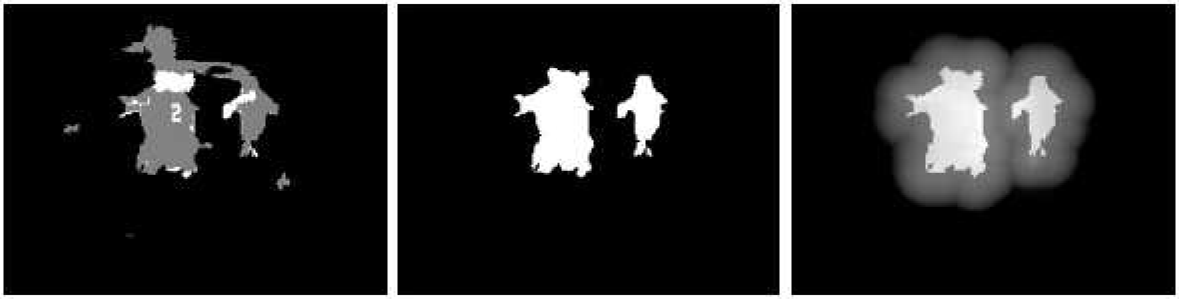


Figure 2.5: From [45] (Left) Initial eye fixation based foreground & background labellings in white and black respectively. (Center) Graph-cut labeling. (Right) Final content map.

function over the cropping window size and position. The cost $C(\Omega)$ of a crop $\Omega$ is computed as a linear combination of five terms.

$$C(\Omega) = \begin{bmatrix} T_{subj}(\Omega) & T^2_{subj}(\Omega) & T_{whole}(\Omega) & T_{cut}(\Omega) & T_{size}(\Omega) \end{bmatrix} \cdot \mathbf{w}^T \tag{2.12}$$

where $\mathbf{w}$ is the weighting vector. The $T_{subj}(\Omega)$ denotes the percentage of 'mass' from the content map omitted by a crop. $T_{whole}(\Omega)$ is the average of all pixels in the content map which pass through the boundary of the crop window. This term is intended to deter the cropping rectangle from passing through high content areas. $T_{whole}(\Omega)$ is the number of graph-cut segmentation boundaries which the current crop passes though normalized by the perimeter of the cropping rectangle. This encourages the crop rectangle to pass through homogeneous areas. Finally, $T_{size}(\Omega)$ is the percentage of the original image covered by the current crop rectangle. An optimal window is found by performing a coarse search over the entire solution space.

A small user study was performed where subjects were asked to choose between two alternative crops. The results showed that images generated using gaze-based cropping were preferred over saliency map based crops and down sampled images 58% and 56% of the time respectively. Conversely, hand cropped images were preferred over gaze-based cropping 73% of the time.

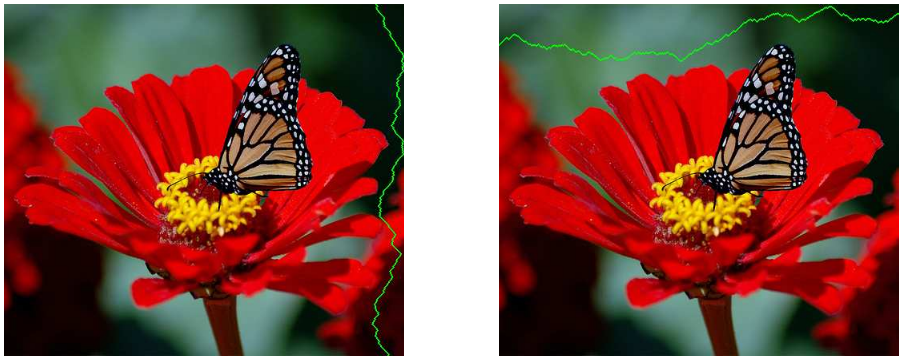
### 2.4.2 Seam-based Techniques

Seam-based approaches represent alternatives to rectangular cropping which can potentially be applied to a broader class of images. Unlike cropping, these techniques are not constrained to select a single contiguous region of the input image and yield significantly different results. They generally perform well for small to medium amounts of resizing. However, for significant resizing (to less than half the original image width or height) image content can be significantly distorted. Just as cropping based techniques, these methods are highly dependent on low-level features for identifying important content.

#### Seam Carving

The first seam-based approach to image resizing was presented in [1]. This approach resizes an image by successively removing seams of pixels from an image until the desired size is reached. A seam is defined as an 8-connection path of pixels which runs across the image. There are two types of seams, vertical seams (Figure 2.6(a)) which run from top to bottom, containing a single pixel from each row and horizontal seams (Figure 2.6(b)) which run from side to side containing a single pixel from each column. Formally, vertical and horizontal seams are ordered lists of n and m pixels respectively where $\mathbf{s}(i)$ is used to denote the image coordinates of pixel $i$ in a given seam. Removing a vertical seam reduces the number of rows in the image and removing a horizontal seam reduces the number of columns. The major problem address by the seam carving algorithm is optimal seam selection. This task is posed as an energy maximization problem where the minimum energy seams are successively removed until the desired size is reached. Various energy functions are studied including Itti & Koch saliency [28] and the Harris corner measure [23], however the authors observe that the simple gradient magnitude energy function defined as

$$e(\mathbf{I}) = \left|\frac{\partial \mathbf{I}}{\partial x}\right| + \left|\frac{\partial \mathbf{I}}{\partial y}\right| \tag{2.13}$$



(a) (b)

Figure 2.6: (a) Horizontal Seam and (b) Vertical Seam (shown in green).

works well in general. The cost of removing a seam is defined as the sum of energies for all pixels in the seam. An optimal seam $s^*$ is a seam with minimum energy defined as,

$$s^* = \min_{\mathbf{s}} \sum_{i=1}^{l} e\left(\mathbf{I}(\mathbf{s}(i))\right) \tag{2.14}$$

where $l$ is equal to the number of pixels in a seam ($n$ for horizontal seams and $m$ for vertical seams).

The number of possible seams is exponential in seam length however because of the 8-connected nature of seams a dynamic programming algorithm can be used to reduce the time complexity of this search to $O(nm)$ time. This algorithm is based on a simple recurrence relationship for the minimum energy of a seam. For example, in the case of vertical seams the energy of a minimum length seam terminating at pixel $(i, j)$, denoted $M(i, j)$, can be expressed as follows,

$$M(i, j) = e(i, j) + min(M(i-1, j-1), M(i-1, j), M(i-1, j+1)) \tag{2.15}$$

By using standard dynamic programming techniques [14] and this recurrence relationship, a minimum energy seam across the whole image can be computed by scanning across the pixels of the (energy) image from top to bottom. Once the minimum energy seam is found it can be deleted to reduce the number of rows or columns in the input image by one. Results of the seam carving procedure for both vertical and horizontal seams are

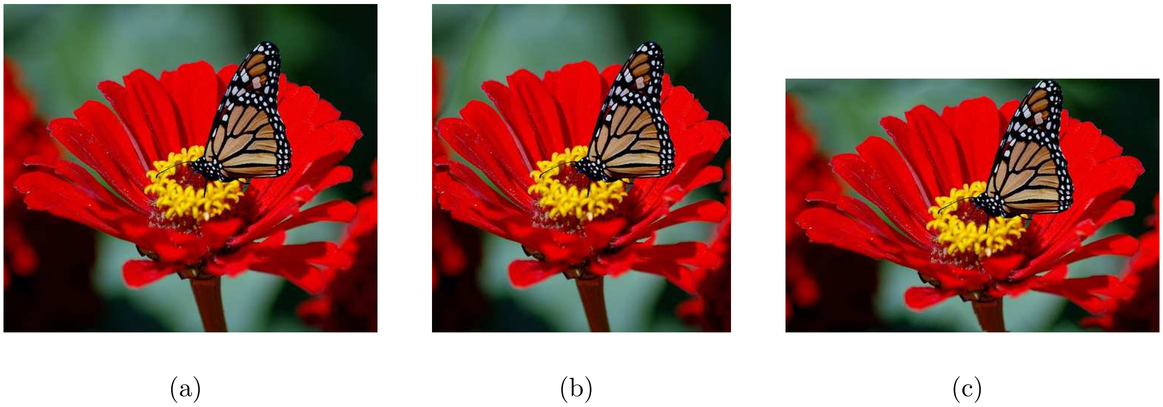


Figure 2.7: (a) Original image. Result of removing 100 seams using (b) vertical and (c) horizontal seam carving.

shown in Figure 2.4.2. Seam carving can also be used to reduce both the number of columns and rows in an image by combining both horizontal and vertical seam removal operations. An optimal seam ordering between vertical and horizontal seams can also be computed using a dynamic programming algorithm. Although, exceptional results can sometimes be achieved with this technique there are several instances where seam carving introduces artifacts in the resized image, especially when images are resized by a large amount.

### Seam Carving Extensions

In [44][50] the seam carving method is extended to video, where a temporal term is added to the energy function so seam removal is smooth through time. Using this new formulation a graph-cut algorithm is developed to find spatiotemporal seams through the video sequence. The authors also develop an new cost function to termed the forward-energy. In [1] the cost of a seam is related to the energy it removes from an image (Eqn. 2.15) , alternatively the forward-energy cost of a seam is a function of the amount of energy which its removal adds to an image through the new edges which are introduced. This alternative cost function works significantly better for video retargeting.

#### Constrained Sampling for Image Retargeting

In [43] Ren et. al. propose a resizing method closely related to seam carving. In this work the removed pixels are no longer constrained to lie along an 8-connected path in the image. Instead the image is segmented into discrete regions and connected 'seams' of regions are selected to have pixels removed for retargeting. The benefit of this method over seam carving is that it allows seams to be locally disjunct which makes it easier to avoid high energy areas. An optimal path through the regions is selected using a graph-cut algorithm [14] where flows are assigned using an importance map based on gradient magnitude and visual saliency. The flow capacities are assigned proportionally to the negative importance of a region.

Although some of the ideas presented in this work are promising, the algorithm formulation depends critically on image segmentation and visual saliency which significantly reduces its robustness.

### 2.4.3 Distortion-based Techniques

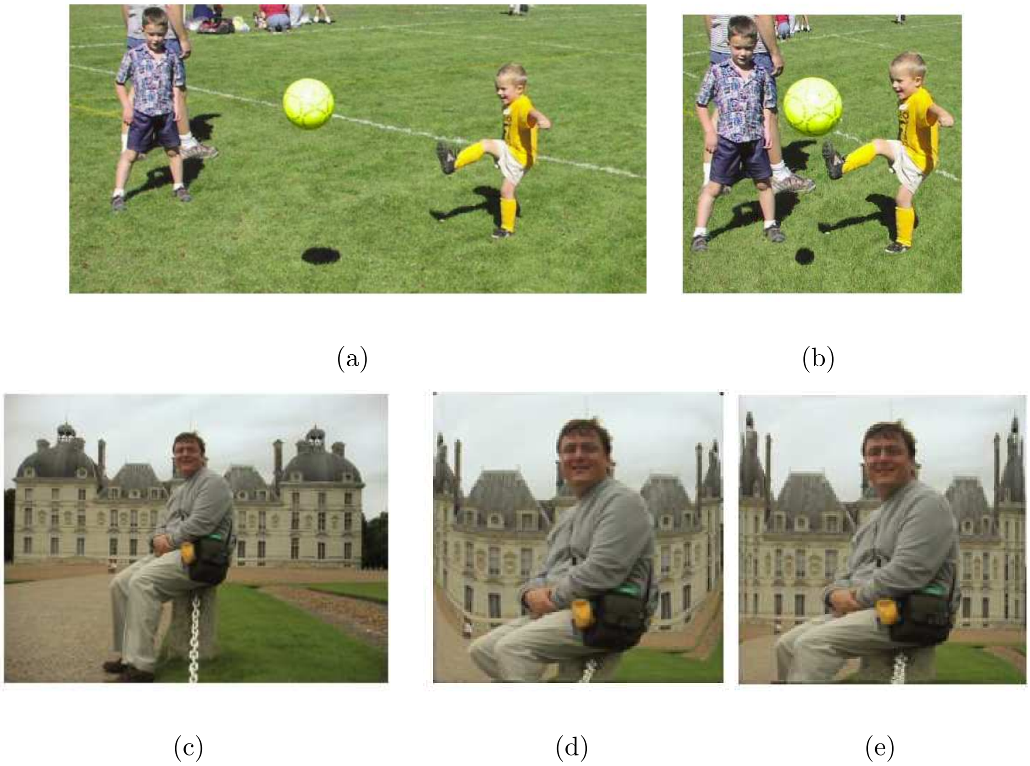
An alternative class of resizing techniques attempt to resize images by distorting selected regions to enhance the foreground content and suppress the background. These methods can yield interesting results in some cases but generally introduce noticeable artifacts into the resized image.

#### Resizing by Cutting and Pasting

In [49] a method for resizing images with multiple disjoint objects is proposed. This technique involves cutting the important objects out of the background image and re-inserting them at a larger scale relative to the original image. The importance value of a region is computed using a saliency map [28] and face detector [48] just as in the image attention model [6]. To extract complete objects the input image is over-segmented using the mean-shift algorithm and then regions are grouped using color similarity and importance. The objects are then removed from the image by applying an inpainting algorithm [24] to fill in the regions where they were located. Finally, a heuristic technique is used to reinsert the objects back into the image at a larger scale by re-sampling them. This method inherently distorts the image content as important objects are re-inserted at an unrealistic scale. An example of an image successfully rescaled by this method is shown in Figure 2.8(b). A major weakness of this algorithm is the image segmentation and grouping step which is likely to fail in a significant fraction of the time.

### Non-linear Fisheye Warping

In [33] the authors present a resizing method based on warping images to emphasize the focus area but still retain contextual information. This is achieved by performing a non-uniform resampling of the image where the sampling density of pixels is reduced proportionally to their distance from the key region of interest. Imagery collected using a fisheye lense exhibits radial distortion which makes objects near the center of images appear larger than those further away from the center point. Thus, various image warps can be developed by applying a fisheye lens model. This method assumes that there is a single relevant ROI, and when rescaling is performed this ROI is uniformly scaled while the remainder of the image is warped. An example with results from two different warping functions is shown in Figure 2.8(d) and 2.8(e). Automatic computation of the image ROI is performed using a modified version of the greedy cropping algorithm in [51], and importance maps are computed using a face detector and a contrast-based saliency map [36]. Besides inherently distorting image content, the main limitations of this technique is the automatic ROI detection.



(a) (b)

(c) (d) (e)

Figure 2.8: Distortion based resizing examples. (a) Original image. (b) Retargeted using cutting and pasting (from [49]). (c) Original image. (d) and (e) Non-linear warping with different warping functions (from [33]).

# Chapter 3

# Energy-based Image Importance and Text Relevance

## 3.1 Introduction

The primary limitation of automatic image resizing methods is the computation of an importance or saliency map for the input image. As discussed in Chapter 2, the most popular saliency models are based on low-level image descriptors and statistics [54][28][17]. These methods have prevailed over high-level approaches because they are computationally convenient to implement and can be easily modeled using the mathematical tools such as probability and statistics. Although these methods are limited by the semantic gap, the tools required for top-down analysis are not yet available. Automatic image resizing is exactly the type of problem which requires semantic analysis and even if robust image understanding tools were available, this problem would still pose significant challenges due to its subjective nature. With this in mind, the saliency model presented here is intended to tighten this gap by incorporating domain specific assumptions about the image into the model. The proposed method is general enough to apply to a wide range of AIR problems and results demonstrate that it is substantially more effective for poster resizing than the standard Itti and Koch method [28].

In the domain of poster resizing it is also desirable able to detect suitable locations on the image for text placement. The additional text information in **P** provides a basis to directly model suitable text regions. Using this information we can construct text relevance maps in a similar fashion as importance maps. Under appropriate assumptions the text region provides insight into the image content and this information can be combined with the existing data to improve both relevance and importance map estimates.

## 3.2 Coarse Level Center-surround Model

The approach presented here is intended to build upon the center-surround (CS) paradigm by constructing a coarse-level model specifically designed for poster and image resizing applications. Figure 3.1 illustrates a block digram outlining the main steps which can be used in any center-surround model. The center-surround difference step generally receives the most attention in the literature, however all steps influence the final results and should be carefully considered.

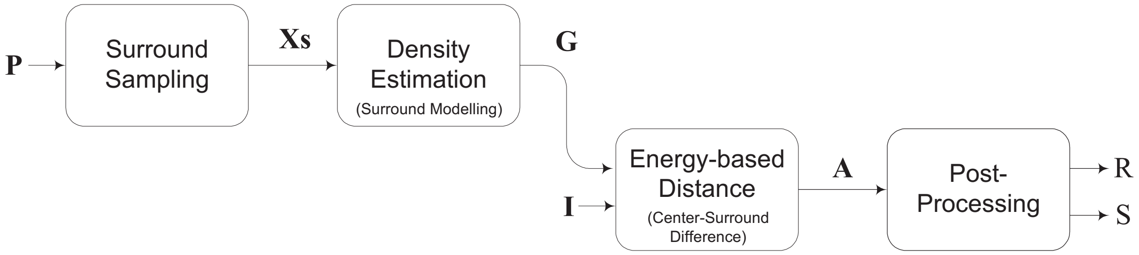


Figure 3.1: Block diagram depicting the major computational steps involved in center-surround relevance map estimation.

The first step is the selection of pixels or features to build a surround model. In most CS techniques the surround sampling step involves selecting pixels in the neighborhood of the saliency value being estimated. However, this process can be specialized for image resizing, resulting in more predicable and usable importance maps.

A subtle aspect of AIR research is that these systems are often evaluated on *"images which can benefit from resizing"* [33][45][49][51]. These images usually contain clearly identifiable foreground objects and a substantial fraction of the image corresponds to background scenery. In essence, these images can be easily reduced in size without removing significant foreground regions. This relates directly to the shortcomings of current saliency maps and identifies an unstated assumption used by nearly all AIR algorithms. Rather than ignoring this, the proposed method uses this assumption to sample in a more effective way.

The next step in map construction is the surround modeling, for existing methods this can range from taking a weighted average of surround samples [28] to estimation of a generalized Gaussian density [16]. In this case the density estimation step is performed by fitting a mixture of Gaussian model to the surround samples.

Given a surround model, center-surround (CS) differences can be used to compute relevance maps. Recent models have used probabilistic or information theoretic metrics to quantify the CS differences. In the case of importance maps, pixels which have a low likelihood under the surround model are given large importance values and for text relevance maps, pixels which are similar (high likelihood) to the text distribution model are given higher relevance values. The method here uses an energy-based model to transform probabilities under the surround model to a scale which is specialized for image resizing.

Finally, there can be a post-processing step used to enhance or refine the maps. This can range from simple smoothing to the normalization procedures used with Itti and Koch saliency maps. In this work, image importance and text relevance maps only differ in the post-processing steps.

The following sub-sections discuss the detailed computation of these maps under the model presented above. To be consistent with the notation for $\mathbf{P}$ the input color image will be denoted by $\mathbf{I}$. The procedure discussed here applies to any pixel features of an image thus **I** can also be thought of as an arbitrary multi-dimensional feature map.

### 3.2.1 Surround Sampling

The surround sampling procedure is designed to exploit the general structure of images used in many automatic resizing and advertising scenarios. This is described here as the concept of *images which can benefit from resizing*. It is a highly subjective idea and can be formalized in numerous ways. The interpretation used here is that these images contain one or more large-scale focal points (objects) bordered by simple or moderately complex background scenery. Examples of such images are shown in Figure 3.2. Although the position and scale of the key objects differs across these images, they have a semantically important regions which stand out from the background context. For example, the man on the camel in Figure 3.2(a) or the people in Figure 3.2(d). For this class of images we can sample from the surround distribution by selecting pixels (or features) from the periphery of the image.

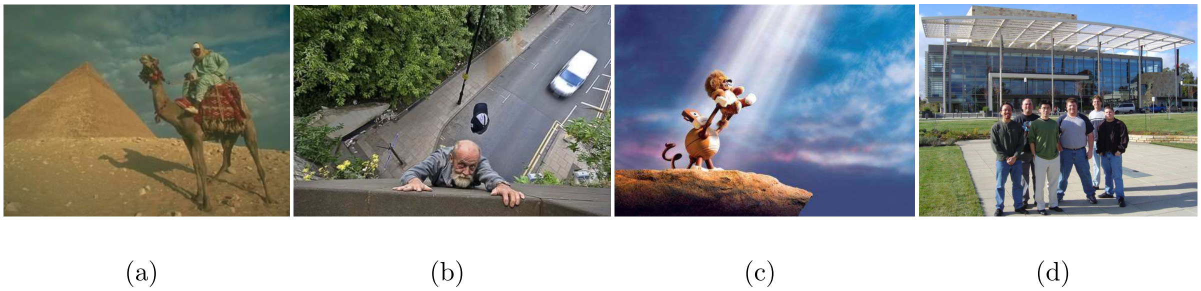
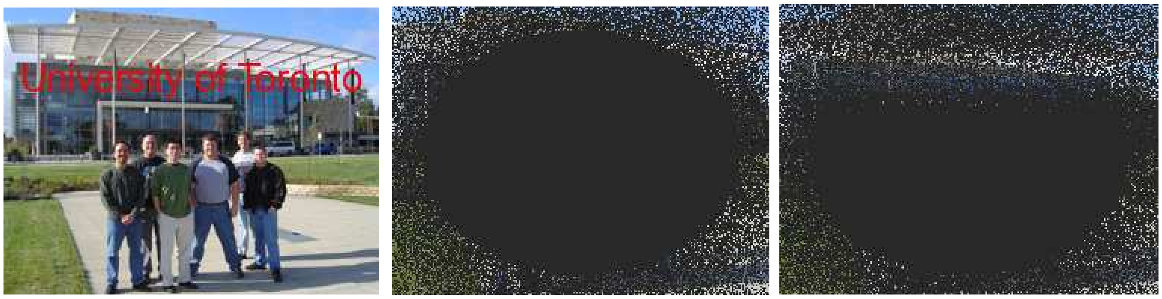

(a) (b) (c) (d)

Figure 3.2: Examples of images which satisfy the coarse center-surround assumption. This model can be applied to a wide-range of images across many different contexts.

In many instances the surround pixels sampled from **I** will be highly redundant which allows us to use a fraction of them to construct the surround model. For images such as Figure 3.2(b) the sampling procedure is likely to contain noisy samples due to important objects close to the image (outer) edges, however results demonstrate that a fixed sampling procedure performs well as long as the surround model complexity is constrained to avoid over-fitting.

In the simplest case, sampling can be performed by uniformly selecting features from rectangular strips along the image edges. This yields good results in practice but the performance can be improved by incorporating an assumption that the likelihood of being from the surround distribution increases proportionally to the distance from the image center. This can be accomplished by sampling edges proportionally to a complementary Gaussian window, $1 - \mathrm{G}(\mathbf{p})$. Figure 3.3(b) illustrates a random sampling of image pixels from the surround distribution and their locations on the input image. The surround model corresponds to the set of image features (pixels) at these locations in $\mathbf{I}$.



(a) (b) (c)

Figure 3.3: Surround sampling procedures. (a) Original Poster (b) Surround sampling (c) Text-surround Sampling. For clarity, samples are shown at their corresponding positions in the original image.

For poster resizing it is advantageous to use the additional information given by the text region position and size. This can be accomplish by selecting pixels in $\mathbf{I}$ from this area to construct a text-region model. The sampling procedure can be implemented using a Gaussian window with the majority of its mass within the text region or by uniformly selecting pixels from the text region bounding boxes.

If we assume text is placed on top of the background content then samples from the surround and text regions can be combined. This provides additional data for modeling which is generally going to improve the resulting importance and relevance maps. We refer to this as sampling from the *text-surround* distribution. An example of this is shown in Figure 3.3(c). In these examples the sampling strips have width equal to 7.5% of the image width at their narrowest point which has been selected using cross-validation. For notational convenience, the set of samples will be denoted $\mathbf{X_s}$, also note that $\mathbf{X_s}$ only contains the feature or pixel values at the sampling locations (without location information). All discussions below apply equally for $\mathbf{X_s}$ created by surround or text-surround sampling.

### 3.2.2 Surround Modeling

To compute center-surround differences each feature in $\mathbf{I}$ must be compared with the set of features in $\mathbf{X_s}$. To make the most of all available information, the range of values in $\mathbf{X_s}$ should be characterized. The surround modeling step involves this characterization and is accomplished using a probability density estimation procedure.

For techniques based on local sampling, unimodal distributions such as generalized Gaussian are effective for modeling [16]. However, the samples in $\mathbf{X_s}$ are global and can be more accurately characterized with a mixture of Gaussian model [3].

#### Mixture of Gaussian Density Estimate

The mixture of Gaussian model (MoG) is a popular technique for clustering and density estimation which has been employed to solve image analysis problems within CBIR [4] and segmentation [2]. Under this model, a sample is generated from one of K different Gaussian distributions each specified by their own mean $\boldsymbol{\mu}_k$ and covariance matrix $\mathbf{\Sigma}_k$. Intuitively, the process of generating a single sample from an MoG involves first selecting one of the Gaussian components, where the $k^{th}$ component has a probability of $\pi_k$ of being selected, and then randomly sampling from the selected Gaussian.

The distribution of $\mathbf{X_s}$ is unpredictable and generally multi-modal which is well suited for this type of density. Formally, the probability of an observed vector $\mathbf{x}$ is expressed as,

$$P(\mathbf{x}) = \sum_{k-1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.1}$$

Model fitting or training is performed by iteratively searching for parameter settings which maximize the log likelihood of the training data ($\mathbf{X}_s$). The main challenge associated with estimating the parameters of an MoG is that the Gaussian component for each observed sample is unknown, commonly referred to as a hidden variable. In this case the expectation maximization (EM) algorithm [3][11] can be used. Although EM is only guaranteed to find a local maxima of the log likelihood it has been shown to yield good results in many applications including this one. For further detail regarding the EM algorithm applied to MoG models the reader may refer to [3].

### 3.2.3 Energy-based Center-surround Difference

The center-surround difference is the core step of the saliency map computation. Using the density estimate computed in surround modeling step a *distance* will be computed between each feature in $\mathbf{I}$ and the MoG model of $\mathbf{X_s}$. In a probabilistic setting this distance is inversely proportional to the probability of each feature under the surround model. Rather than using this probability directly, we can construct an energy-based model, which transforms these probabilities based on specialized constraints related to image resizing problems.

#### Energy-based Relevance Model

The approach taken here is motivated by the work of [7] and [57]. In [7] the authors use partial knowledge of foreground (FG) and background (BG) pixel values and locations to estimate matting coefficients (FG-BG membership) for unknown pixels in an image. In [57] the authors demonstrate that a linear transformation in gradient-space can improve salient region detectors by increasing the influence of uncommon (improbable) feature vectors. Following a similar principle the proposed model implements a transformation which suppresses features vectors which are likely to be part of the surround.

Using the MoG model constructed from $\mathbf{X}_s$, each feature in $\mathbf{I}$ is greedily associated with the Gaussian component with the highest likelihood of generating that sample. For convenience, the Gaussian parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for element $\mathbf{I}(i,j)$ will be denoted by $\mathbf{G}(i,j)$. A relevance map $\mathbf{A}$ is computed by solving for the maximum likelihood state under an energy-based model [3]. In general, the energy contribution of a particular $\mathbf{A}(i,j)$ is a function of $i$, $j$, $\mathbf{A}$, $\mathbf{G}$, and $\mathbf{I}$, expressed as $\eta(i,j,\mathbf{A},\mathbf{G},\mathbf{I})$.

$$P(\mathbf{A} \,|\, \mathbf{G}, \mathbf{I}) = \frac{1}{Z} \prod_{i,j} \exp^{-\eta(i,j,\mathbf{A},\mathbf{G},\mathbf{I})} \tag{3.2}$$

where $Z$ is a normalization factor included to enforce that the probability distribution sums to one. To ensure that this is a valid probability density the energy function $\eta$ is also constrained to be non-negative. Note that the background distribution $\mathbf{G}$ and image feature vectors $\mathbf{I}$ are observed and the relevance values are the only unknown values. The optimal $\mathbf{A}$ is equivalent the map with minimum energy summed over all pixels. This can be shown as follows,

$$\begin{aligned} \log(P(\mathbf{A} \,|\, \mathbf{G}, \mathbf{I})) &= \log Z - \sum_{i,j} \eta(i,j,\mathbf{A},\mathbf{G},\mathbf{I}) && (3.3) \\ \Rightarrow \operatorname*{argmax}_{\mathbf{A}} P(\mathbf{A} \,|\, \mathbf{G}, \mathbf{I}) &= \operatorname*{argmin}_{\mathbf{A}} \sum_{i,j} \eta(i,j,\mathbf{A},\mathbf{G},\mathbf{I}) && (3.4) \end{aligned}$$

where we have made use of the fact that maximizing the log likelihood is equivalent to maximizing the original likelihood [13]. The structure of the model is identical to a conditional random field (CRF) [32], where $\mathbf{A}$ is not a set of discrete labels but a set of real numbers each within the range $[0,1]$. It may also be of interest to note that the previous steps in the importance map computation (sampling and density estimation) can be considered a part of the CRF feature computation.

The intent of the energy-based model is to impose certain domain specific constraints on the resulting relevance maps. To understand the constraints it is helpful to reconsider the basis of the center-surround paradigm. Once again, the goal is to distinguish important regions from background context or surround. The distribution of the background context is observed (with noise) through the surround sampling process. Without any additional information, saliency is measured in terms of dissimilarity with background context. The key assumption here is that the context and important region distributions are disjoint in feature space. However, this does not mean that these distributions are infinitely far away, therefore at some point being distant from surround-space loses meaning as a relative measure of importance. Furthermore, maps estimated from the above assumptions are likely to be noisy, therefore it is important that relative saliency values are appropriately scaled and that individual points do not have unlimited influence on resizing solutions. The energy-based model presented below has been developed with these properties in mind.

Consider the estimation of a single element at position $(s,t)$ in the relevance map, where for readability we will drop subscripts and refer to the key quantities as $\mathbf{A}(s,t) = \alpha$, $\mathbf{I}(s,t) = \mathbf{p}$, and $\mathbf{G}(s,t) = \{\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B\}$. When explicit subscripting is necessary elements of $\mathbf{A}$ will be referred to as $\alpha_{i,j}$. The solution for $\mathbf{A}$ is equivalent to simulating a physical system where each feature of $\mathbf{I}$ can be thought of as a particle in space. The vector $\mathbf{p}$ is used to represent a pixel in $\mathbf{I}$ to emphasize the analogy of a position vector in feature space.

The physical system consists of three competing energies. A directionally-dependent force attracting each particle towards its associated Gaussian (surround) mass with a gravitational force proportional to the distance between $\mathbf{p}$ and $\boldsymbol{\mu}$. A spring force (similar to a prior) encouraging $\mathbf{p}$ to remain at its initial position, and an opposing spring (smoothing) force encouraging particles in the neighborhood of $\mathbf{p}$ to have similar $\alpha$ values. If the new position of each particle is constrained to lie on the line between $\mathbf{p}$ and $\boldsymbol{\mu}_B$ then resulting position can be expressed as,

$$\mathbf{s} = \alpha \cdot \mathbf{p} + (1-\alpha) \cdot \boldsymbol{\mu}_B \tag{3.5}$$

where $\alpha$ corresponds to the weighting coefficient. According to the dynamics of the system the optimal $\alpha$ will lie in the range $[0, 1]$. The energy of a particular setting of $\alpha$ is computed as the sum of energies resulting from the three forces discussed above,

$$\eta\left(i, j, \mathbf{A}, \mathbf{G}, \mathbf{I}\right) = \eta_G(\alpha, \boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B, \mathbf{p}) + \eta_I(\alpha, \boldsymbol{\mu}_B, \mathbf{p}) + \eta_N(\alpha, \mathbf{A}) \tag{3.6}$$

The first term, $\eta_G$, corresponds to a gravitational potential energy from the force attracting the particle towards the background distribution.

$$\begin{aligned}\eta_G(\alpha, \boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B, \mathbf{p}) &= k_g \cdot g \cdot r \\ &= \frac{k_g \cdot \sqrt{(\mathbf{s} - \boldsymbol{\mu}_\mathbf{B})^T \boldsymbol{\Sigma}_B (\mathbf{s} - \boldsymbol{\mu}_\mathbf{B})}}{(\mathbf{p} - \boldsymbol{\mu}_B)^T \boldsymbol{\Sigma}_B^{-1} (\mathbf{p} - \boldsymbol{\mu}_B)}\end{aligned} \tag{3.7}$$

Malhanobis distances [13] are used to take advantage of the information provided by the Gaussian density. The $k_g$ absorbs all constants which would be present in the physical expression (i.e the mass $m$ and gravitational constant $G$). The denominator results from a fixed gravitational force, $g$, which is proportional to (initial) inverse squared distance between objects. Also note that the inverse covariance matrix is used in the denominator to enforce stronger forces in the directions which $\mathbf{G}$ has greater peak width. The second term, $\eta_I$, corresponds to the spring energy attracting the particle to position $\mathbf{p}$.

$$\eta_I(\alpha, \boldsymbol{\mu}_B, \mathbf{p}) = k_s \left(\mathbf{s} - \mathbf{p}\right)^T \left(\mathbf{s} - \mathbf{p}\right) \tag{3.8}$$

This is simply the multi-dimensional expression for spring energy, where $k_s$ corresponds to the spring constant. Finally, the neighborhood smoothing term $\eta_N$ is simply a scalar version of $\eta_I$.

$$\eta_N(\alpha, \mathbf{A}) = k_n \sum_{(i,j) \in N_\alpha} (\alpha - \alpha_{i,j})^2 \tag{3.9}$$

where $\mathbf{N}_\alpha$ corresponds to the 4-connected neighborhood of $\alpha$ and $k_n$ is a spring constant. The smoothing term introduces coupling between the relevance values in $\mathbf{A}$ which complicates estimation. Iterative techniques such as iterated condition modes (ICM) [3] can be employed to search for a solution, however it is preferable to find a more efficient method for solving $\mathbf{A}$ which is not subject to poor local maxima.

### Minimizing the Disconnected Energy Function

To gain insight into the solution for the problem it is beneficial to consider the case when $k_n = 0$ and all the $\alpha_{i,j}$ values decouple. This simplified problem is also interesting because it computationally simpler to solve and yields similar results. In this case we can exactly solve for the minimum energy configuration by computing the derivative of $\eta\left(i, j, \mathbf{A}, \mathbf{G}, \mathbf{I}\right)$ and setting it equal to zero. Setting $k_n = 0$ and subbing Equations (3.5) ,(3.7), and (3.8) into Eqn. (3.6) with algebraic simplification yields,

$$\begin{aligned}\eta_{dis}\left(\alpha, \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \;&=\; k_g g \cdot \alpha \cdot \sqrt{(\mathbf{p}-\mu_{\mathbf{B}})^T \boldsymbol{\Sigma}_B (\mathbf{p}-\mu_{\mathbf{B}})} \;+ \\ &\quad\; k_s \cdot (\alpha-1)^2 (\mathbf{p}-\mu_{\mathbf{B}})^T (\mathbf{p}-\mu_{\mathbf{B}}) \qquad (3.10)\end{aligned}$$

where $\eta_{dis}$ is used to indicate that this is the disconnected version of the energy function. From this form it is straight forward to compute the partial derivative with respect to $\alpha$ and solve $\partial\eta_{dis}/\partial\alpha = 0$. A complication arises because the derivative of the $\eta_G$ term is discontinuous at $\mathbf{s} = \boldsymbol{\mu}$ or equivalently $\alpha = 0$. Fortunately, this zero of the partial derivative is trivially related to the correct answer. Denoting this quantity as $\alpha_e$ we have

$$\frac{\partial\eta_{dis}}{\partial\alpha} = 0 \Rightarrow \alpha_e = \frac{k_s(\mathbf{p}-\boldsymbol{\mu}_B)^T(\mathbf{p}-\boldsymbol{\mu}_B) - \frac{1}{2}k_g g \cdot \sqrt{(\mathbf{p}-\mu_{\mathbf{B}})^T \boldsymbol{\Sigma}_B (\mathbf{p}-\mu_{\mathbf{B}})}}{k_s(\mathbf{p}-\boldsymbol{\mu}_B)^T(\mathbf{p}-\boldsymbol{\mu}_B)} \qquad (3.11)$$

from which the exact (or adjusted) $\alpha$ can be computed simply as

$$\alpha = \max\left(\alpha_e, 0\right) \qquad (3.12)$$

This result can be directly reasoned from analysis of (3.11) and the physical interpretation of the model. The problem occurs when the gravitational force is stronger than the spring force at $\mathbf{s} = \boldsymbol{\mu}$. In this case the minimum energy configuration would place the particle exactly at $\mathbf{s} = \boldsymbol{\mu}$ whereas $\alpha_e$ indicates a negative value because of the aforementioned discontinuity in $\partial\eta_G/\partial\alpha$. The result in Eqn. (3.12) has also been verified via numerical simulation.

### Minimizing the General Energy

Minimization of the complete energy is more difficult because of the coupling, however we can take a similar approach as above and find an approximate solution for for $\mathbf{A}$ which can be expressed as a closed-form expression and is unique. Following algebra as above, the partial derivative of $\eta\,(i, j, \mathbf{A}, \mathbf{G}, \mathbf{I})$ can be computed and set to zero, with non-zero $k_N$ this yields,

$$\frac{\partial \eta}{\partial \alpha} = 0 \Rightarrow$$

$$\alpha_e = \frac{k_s(\mathbf{p} - \boldsymbol{\mu}_B)^T(\mathbf{p} - \boldsymbol{\mu}_B) - \frac{1}{2}k_g g \cdot \sqrt{(\mathbf{p} - \mu_{\mathbf{B}})^T \boldsymbol{\Sigma}_B (\mathbf{p} - \mu_{\mathbf{B}})} + k_n \sum_{(i,j)\in\mathbf{N}_\alpha} \alpha_{i,j}}{k_s(\mathbf{p} - \boldsymbol{\mu}_B)^T(\mathbf{p} - \boldsymbol{\mu}_B) + k_n N_n} \tag{3.13}$$

where $N_n$ is the number of neighbors in $\mathbf{N}_\alpha$ which is equal to 4 in this work. Once again $\alpha_e$ is used to emphasize that the zero of the partial derivative does not always yield the correct solution. The first two terms in the numerator of (3.13) are the same as the problematic terms in (3.11). Using the max$(\cdot)$ operation to avoid a negative energy difference will no longer yield an exact solution because the $\alpha_{i,j}$ values in the third term are also unknown, nonetheless it is a close approximation. The coupling between $\alpha$ terms requires all equations to be solved simultaneously. Further algebraic manipulation and the addition of max$(\cdot)$ to truncate negative energy differences yields a system of linear equations of the form,

$$\alpha \ + \sum_{(i,j)\in\mathbf{N}_\alpha} \psi \cdot \alpha_{i,j} \ = \ \nu \tag{3.14}$$

where,

$$\nu = \max\left(\frac{k_s(\mathbf{p} - \boldsymbol{\mu})^T(\mathbf{p} - \boldsymbol{\mu}_B) - \frac{1}{2}k_g g \cdot \sqrt{(\mathbf{p} - \mu_{\mathbf{B}})^T \boldsymbol{\Sigma}_B (\mathbf{p} - \mu_{\mathbf{B}})}}{k_s(\mathbf{p} - \boldsymbol{\mu}_B)^T(\mathbf{p} - \boldsymbol{\mu}_B) + k_n N_n}, 0\right) \tag{3.15}$$

$$\psi = \frac{-k_n}{k_s(\mathbf{p} - \boldsymbol{\mu}_B)^T(\mathbf{p} - \boldsymbol{\mu}_B) + k_n N_n} \tag{3.16}$$

From inspection of the above expressions, the solutions will only differ significantly from the exact solutions at boundaries between large and small relevance values, where the approximation can cause over-compensation of the smoothing terms which would have been otherwise negated by a very strong gravitational term. This 'over-smoothing' is a minor difference and, for all practical purposes, results in equally good relevance maps. To further motivate the above approximation it is interesting to note that subbing in $k_n = 0$ yields the exact solution for the disconnected case.

There are a total $N_\alpha = n \cdot m$ equations of the form in Equation. (3.14). To express this relationship in standard matrix notation the $\alpha$ and $\nu$ coefficients must be represented as $N_\alpha$ element column vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\nu}$ respectively. The $\psi$ coefficients on the l.h.s of (3.14) can be represented by a $N_\alpha \times N_\alpha$ matrix denoted $\boldsymbol{\Psi}$. The relevance maps can then be solved as the solution to the linear system

$$(\,\mathbf{I}_{N_\alpha} + \boldsymbol{\Psi}) \cdot \boldsymbol{\alpha} = \boldsymbol{\nu} \tag{3.17}$$

where $\mathbf{I}_{N_\alpha}$ is the $N_\alpha \times N_\alpha$ identity matrix. If the 2-dimensional map of $\alpha$ coefficients are mapped to $\boldsymbol{\alpha}$ by sequentially concatenating the columns of $\mathbf{A}$ in order then $\boldsymbol{\Psi}$ will be be a sparse-banded matrix with 4 non-zero elements in each row corresponding to the neighbor terms in (3.14). The solution can be easily computed using standard tools such as Matlab. To reduce computational requirements, the solution $\boldsymbol{\alpha}$ can also be computed at a smaller scale by downsampling $\mathbf{I}$ prior to computing the relevance maps.

The subjective nature of relevance makes setting parameters using quantitative metrics difficult, however in an attempt to maximize performance parameters were selected based on observed results on a small training set. The chosen parameters were $k_s = 1$, $k_g = 12$, $k_n = 25$, with the number of Gaussian components set to $K = 3$ for all images. The solutions vary smoothly with respect to parameter settings and the adaptive properties of the MoG model allow a single parameter set to yield good results across many different images.

Solutions to this model have various desirable properties as mentioned above. As previous discussed, the concept that distance from background (or surround) space as a measure of saliency is much more meaningful in the proximity of this space. Under the energy-based model, distance from surround-space has a diminishing effect on importance values. There is a point where being further from surround-space no longer increases relevance estimates. Furthermore, relevance values have finite influence because of their bounded range.

As a final note, it is insightful to expand on the previous view of this operation as a transformation. In this context the transformed image pixels corresponds to the **s** values in Eqn. 3.5. Points in close proximity of the background distribution will be collapsed or compressed toward the respective mean which can be thought of as a form of background suppression. These values are not explicitly computed in the above formulation but this point of view provides an alternative way of reasoning about the model.

### 3.2.4 Post-processing

Following the center-surround distance computation we have a map **A** which can be directly used as an importance map. However, post-processing steps can often be applied to improve results. The normalization procedure used to compute Itti and Koch saliency maps [28] is an example of this. In [28] maps are also smoothed by combining them at a coarse scale. Similarly, **A** can be filtered with a small Gaussian kernel [53] ($\sigma = 0.005$ of the image dimensions) to introduce additional the smoothness. This qualitatively improves the resulting maps.

Under a text-surround model, the text relevance and image importance maps are closely related. Both maps can be computed directly from **A**, and in the simplest case one is simply the complement of the other. The final two sections of this chapter discuss remaining considerations regarding the construction of text relevance and image importance maps once **A** has been computed. For future reference the proposed technique for computing **A** (from Eqn. 3.17) will be denoted as CRF. Maps created using the disconnected energy function (Eqn. 3.12) will be denoted ICRF (where I stands for independent).

## 3.3 Text Relevance

The construction of text relevance maps follows directly from the computation of **A**. Using the conventions above, the text relevance map is simply equal to,

$$R_{\mathbf{T}}(i,j) = 1 - \mathbf{A}(i,j) \tag{3.18}$$

The main consideration which must be made when constructing **A** is the specific form of image features used. Although, the above technique applies to features of arbitrary dimensionality, experimental results demonstrate that simple color vectors yield good results. The use of such simple features is motivated by two main observations. Firstly, the text regions can be extremely simple, particularly when dealing with advertisements. In this case many components of complex, high-dimensional feature vectors are likely to be roughly constant across the sample-space which can result in nearly singular density estimates. Secondly, the set of samples is limited and highly redundant which makes accurately fitting a high-dimensional density function difficult.

The choice of color space is very important and for predicable and intuitive results a perceptually uniform color space is desirable. Various alternatives have been qualitatively evaluated using a training image dataset and the CIELab [9] and opponent color space presented by Gevers in [18] both yield good results. All computations in this work made use of the opponent color-space.

Example relevance maps constructed from a text-surround model are shown below in Figure 3.4. When inspecting the resulting maps note that large relevance values (yellow-white) correspond to regions which are more suitable for text placement. In these particular examples the use of the text-region samples ensures regions such as the red roof or doll's body do not receive large relevance estimates. Although the body of the man milking the cow receives an erroneously high relevance value, this is still a good result because there are very low relevance values within this region of the image. Also note that the man's face and hands have been correctly assigned low relevance values.

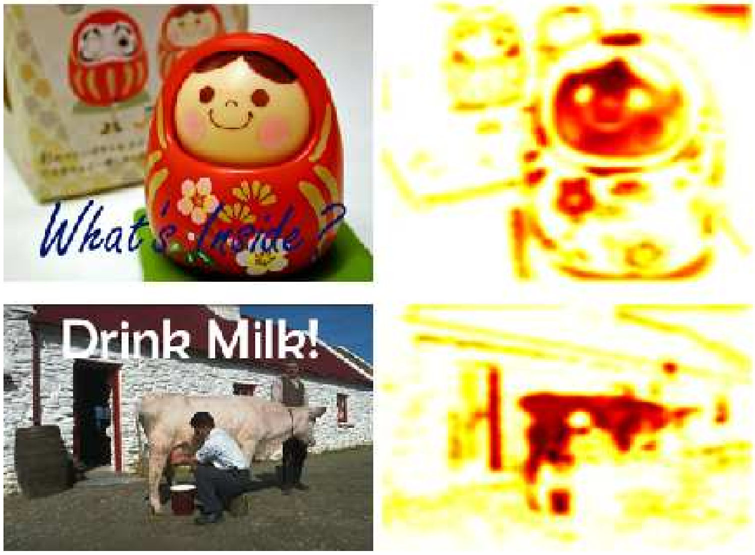


Figure 3.4: Text relevance maps computing using samples from the text-surround distribution. Original posters are shown in the left column with resulting relevance maps in the right column. Relevance values are encoded with a 'hot' color map where white and black correspond to large and small text relevance respectively.

## 3.4 Image Importance

In general, the most appropriate features for computing importance maps will be different than the most appropriate features for text relevance. For text relevance, color features are appropriate however 'interesting' regions of an image typically contain intensity variations and it would be beneficial to use features which also incorporate local complexity. This section demonstrates how we can still use the color-based **A** to compute importance maps and add local-complexity information in the post-processing step. In the subsequent sections image importance maps will be denoted $S$, following the common convention for saliency.

### 3.4.1 Weighted Importance Map

Image importance maps can be constructed in a similar fashion to relevance maps by simply setting $S_{\mathbf{I}}(i,j) = \mathbf{A}(i,j)$. However, it is advantageous to incorporate local complexity information into $S$ by multiplying $\mathbf{A}$ with a local-complexity map. In a sense, $\mathbf{A}$ acts as a color-based weighting function.

The most common indicator of image complexity is the image gradient. The main shortcoming of gradient information is that it is inherently noisy as an importance measure. In most circumstances interesting regions of an image contain non-zero gradient magnitude, however the magnitude itself is a crude indicator of relative importance between regions. The relevance values in $\mathbf{A}$ provide a more accurate measure of relative importance. Taking this into consideration, weighted saliency can be computed using a (binary) gradient indicator map,

$$S_{\mathbf{I}}(i,j) = \mathbf{A}(i,j) \cdot \mathbf{I}_g(i,j) \tag{3.19}$$

where $\mathbf{I}_g$ is the gradient indicator image which is simply computed as,

$$\mathbf{I}_g = \begin{cases} 1\,, & \text{if } \left|\frac{\partial \mathbf{I}}{\partial x}\right|^2 + \left|\frac{\partial \mathbf{I}}{\partial y}\right|^2 > \epsilon \\ 0\,, & otherwise \end{cases} \tag{3.20}$$

where $\epsilon$ is a threshold used to eliminate noise from very small intensity variations. Setting $\epsilon = 0.02$ of the maximum gradient magnitude yields good results. For future reference, the weighted importance maps created using this gradient/boundary map will be referred to as *boundary weighted*. Boundary weighted CRF and ICRF maps will be denoted as BCRF and BICRF maps respectively.

Example CRF and BCRF maps computed from surround models (not text-surround) are shown in Figure 3.5. One important difference between relevance and importance maps is that the latter are generally more sparse. Also note that the weighting procedure usually increases the accuracy of maps, which is very important for image resizing

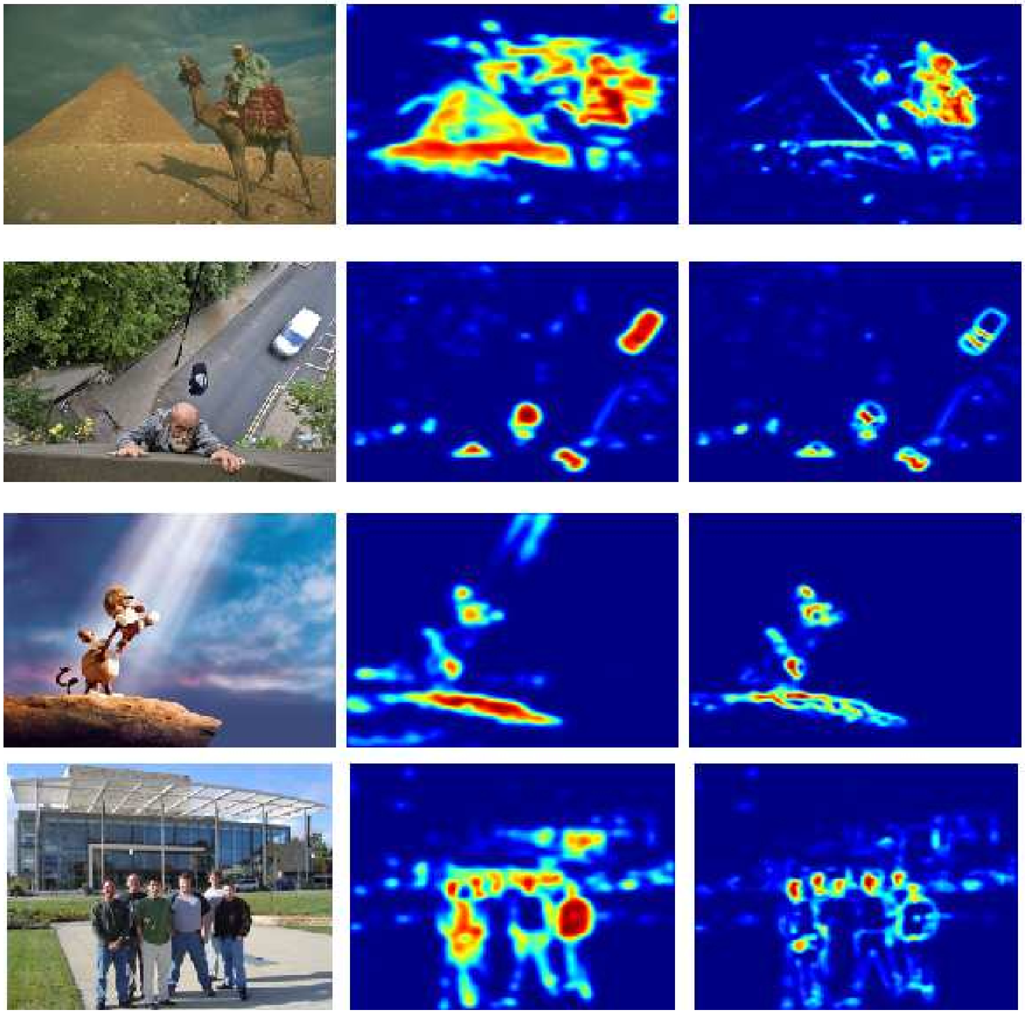


Figure 3.5: Examples of energy-based importance maps. Maps in the center column were computed using CRF model (Eqn. (3.17)). Boundary weighted (BCRF) maps (Eqn. (3.19)) are shown in the left column.

applications. The results also demonstrate the utility of **A** as weighting function for improving noisy saliency maps. As previously stated, **A** behaves like a (color-based) contextual weighting function.

# Chapter 4

# A Framework for Cropping-based Poster Resizing

## 4.1 Introduction

This chapter contains a general framework for resizing a poster **P** to an arbitrary aspect ratio. As previously stated, the purpose of this framework is not to objectify the artistic concept of visual appeal. Instead, the goal is to find a resizing solution which preserves the important properties of the original poster which made it appealing in the first place. This method will be successful when resized posters do not require overall structure to be dramatically changed to accommodate new aspect ratios.

Within this framework image reshaping is performing using rectangular cropping and parameter selection (sub-window location, font size and position) is posed as a minimization problem over an objective function intended to capture the important elements of the poster. The primary reason cropping is used for selection is because it is the most predictable and reliable resizing approach. The limitations of a crop window resizing are generally easy to predict and this significantly simplifies the poster design process.

## 4.2 Optimization Framework

As stated above, a solution to the resizing problem is found by optimization, however before a suitable objective function can be formulated the set of feasible poster resizing solutions must be appropriately parameterized. Recall that an input instance to the problem consists of a rescaling factor $\gamma$ specified for either row or column scaling and poster $\mathbf{P}$ which includes an image $\mathbf{I}$ with spatial dimension $m \times n$. A solution consists of an output poster $\mathbf{P}'$ of dimension $m' \times n'$. The set of feasible solutions can be parameterized in a straightforward way as demonstrated in Figure 4.1. Each solution is specified by a crop window offset vector, $\mathbf{o} = [\, o_x \;\, o_y \,]$, a text heading position vector $\mathbf{c} = [\, c_x \;\, c_y \,]$, and a text size $s$. This particular example corresponds to a column reduction, in which case the search window width is $n' = \gamma \cdot n$ and the height $m' = m$. Also note that in this case the y-offset is fixed $o_y = 0$ for all solutions, resulting in four free parameters. Similarly, when performing row reduction the x-offset is fixed to $o_x = 0$. The parameter

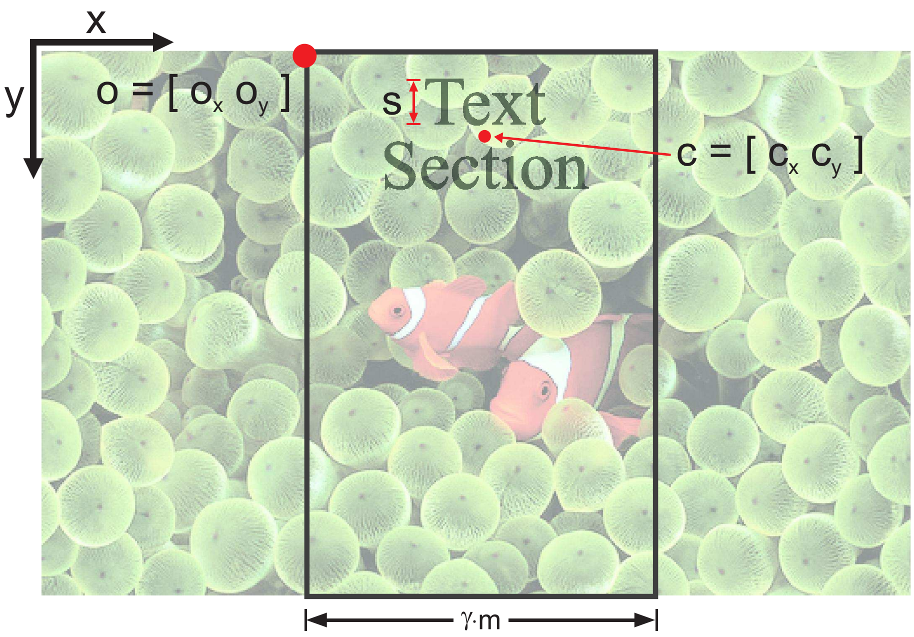


Figure 4.1: Figure demonstrating parameterization of a resizing solution with a image window offset $\mathbf{o}$ , text region center position $\mathbf{c}$, and text (font) size $s$.

$s$ also introduces a complication because it does not completely specify the shape of the text region. It is reasonable that the shape of the heading section should depend on the string of text contained with it, however this adds complexity to the resizing process. For a complete specification of the text region shape the font type, text string, and word wrapping algorithm must also be provided. For notational simplicity the set of all text parameters (including $\mathbf{c}$ and $s$) will be denoted $\mathbf{\Gamma}$. Text wrapping methods will also be discussed below in Section 4.3. As an additional shorthand, $\mathbf{\Omega}$ is used to denote a complete resizing solution, $\mathbf{\Omega} = \{\mathbf{o}, \mathbf{\Gamma}\}$.

Using the above notation a suitable cost function for resizing can now be developed. The form of the objective function is similar to the one used by Santella et. al in [45] (Section 2.4.1) which consists of a weighted sum of cost terms related to the visual properties of the current solution.

$$\begin{aligned} \mathrm{O}\left(\mathbf{\Omega}\right) &= \mathbf{C}(\mathbf{\Omega}) \cdot \mathbf{w}^{\mathrm{T}} \\ &= \begin{bmatrix} \overline{C}_{cont}(\mathbf{\Omega}) & \overline{C}_{cont}(\mathbf{\Omega})^2 & \overline{C}_{text}(\mathbf{\Omega}) & \overline{C}_{text}(\mathbf{\Omega})^2 \end{bmatrix} \cdot \mathbf{w}^{\mathrm{T}} \end{aligned} \tag{4.1}$$

where $\mathbf{w}$ is a four element row vector of constant weights, such as $[\,1.0 \;\; 0.5 \;\; 0.8 \;\; 0.4\,]$. The $\overline{C}_{cont}$ and $\overline{C}_{text}$ are transformed terms used to convert this into a minimization problem. However, it's easier to describe the the objective function in terms of $C_{cont}(\mathbf{\Omega})$ and $C_{text}(\mathbf{\Omega})$ which are defined as follows,

$$\overline{C}_{cont}(\mathbf{\Omega}) = 1 - C_{cont}(\mathbf{\Omega}) \tag{4.2}$$

$$\overline{C}_{text}(\mathbf{\Omega}) = 1 - C_{text}(\mathbf{\Omega}) \tag{4.3}$$

The $C_{cont}(\mathbf{\Omega})$ term measures the the degree to which the important image content is captured in the current window and and $C_{text}(\mathbf{\Omega})$ term measures the appropriateness of the current text parameters. As in [45] the quadratic terms in Eqn. (4.1) are included to discourage solutions with a small content or text saliency value in favor of more balanced solutions in terms of text and content costs.

Both content and text terms are of a similar form with a few importance differences. For descriptive purposes it is easiest to first explain the $C_{text}(\mathbf{\Omega})$ cost function. This term uses a text relevance map $R_{\mathbf{T}}(x, y)$ to measure the suitability of the image content at the current text location.

$$C_{text}(\mathbf{o}, \mathbf{\Gamma}) = \frac{1}{Z_{text}} \sum_{r=1}^{m'} \sum_{c=1}^{n'} W_{\mathbf{T}}(r, c, \mathbf{\Gamma}) \cdot R_{\mathbf{T}}(r - o_x, c - o_y) \tag{4.4}$$

where $W_{\mathbf{T}}$ is a text weighting function and $Z_{text}$ is a normalization term. The weighting function acts as a mask which only selects the area of the text relevance map which corresponds to the current text region. This region is defined by the union of bounding boxes for all lines of text. An example bounding box for a single line of text is shown in Figure 4.2. By denoting the set of pixels contained with these bounding boxes as $B_T(\Gamma)$ the text weighting function is simply defined as,

$$W_{\mathbf{T}}(x, y, \mathbf{\Gamma}) = \begin{cases} 1 & \text{if } (x, y) \in B_T(\Gamma) \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

The weighting function has dimension $m' \times n'$ since it is only defined for the resizing sub-window. As a further enhancement the text weighing function can also be smoothed by convolution with a running average or Gaussian filter [53]. The additional smoothness in the text function reflects uncertainty in the location of the boundary between text and image content regions. An example demonstrating a resizing solution and the corresponding text weighting function smoothed with a running average filter is shown in Figure 4.3. The content term $C_{cont}(\mathbf{\Omega})$ has the same form as the text term but different weighting and importance functions are used. This term measures the mass of image importance located in the current cropped window and is expressed as

$$C_{cont}(\mathbf{o}, \mathbf{\Gamma}) = \frac{1}{Z_{content}} \sum_{r=1}^{m'} \sum_{c=1}^{n'} W_{\mathbf{C}}(r, c, \mathbf{\Gamma}) \cdot S_{\mathbf{I}}(r - o_x, c - o_y) \tag{4.6}$$

where $W_{\mathbf{C}}$ is a weighting function and $Z_{content}$ is a normalization term. In words, the $C_{cont}$ term is a weighted sum of image importance within the selected region of the poster.

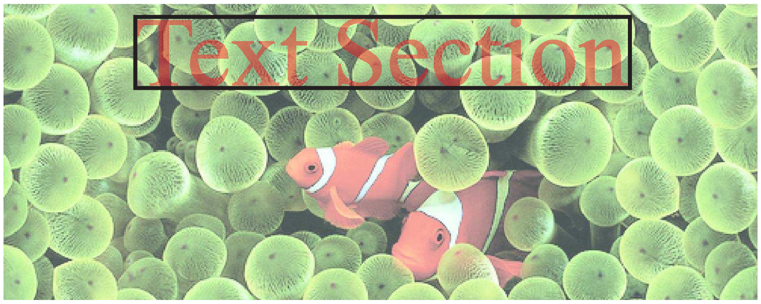
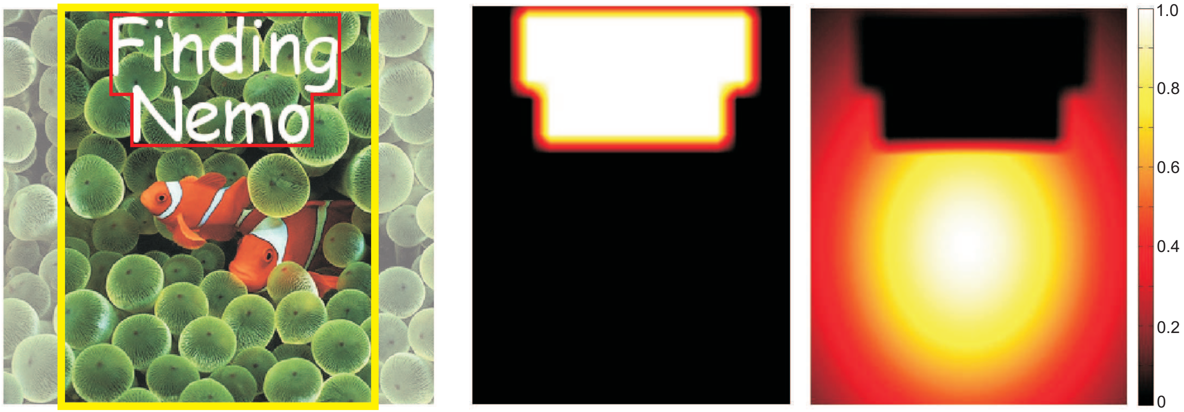


Figure 4.2: Figure demonstrating the text region bounding box of a poster which is shown in black. The set of pixels located within this region is denoted $B_T$.



(a) (b) (c)

Figure 4.3: An example demonstrating the weighting functions used to compute the cost of a resizing solution. (a) A resizing solution with the selected image rectangle in yellow and the text bounding box of $B_T(\mathbf{\Gamma})$ in red. (b) The text weighting function, $W_{\mathbf{T}}(\mathbf{\Gamma})$. (c) The content weighting function, $W_{\mathbf{C}}(\mathbf{\Gamma})$.

The content weighting function is constructed from the complementary text weighting function defined as $W^c_{\mathbf{T}}(x, y, \mathbf{\Gamma}) = 1 - W_{\mathbf{T}}(x, y, \mathbf{\Gamma})$. Also note that $W^c_{\mathbf{T}}$ is also bounded in the range $[0, 1]$ and acts as a mask which gives zero weight to content importance in the text region of the poster. However, this weighting function alone does not generally yield good resizing results. The main problem with $W^c_{\mathbf{T}}$ as a content weighting function is that it does not encourage image content to be centered. Consider the simple case where an importance map contains a single important region which is much smaller than the search window and has zero importance elsewhere. If $W^{c}_{\mathbf{T}}$ is used as the content weighting function then any solution which completely contains the important region will be a minimum cost solution for Equation 4.6. A more appropriate cost function would yield a minimum cost for the single solution where the window is centered on the important region. Alternatively, there are cases were an image contains multiple salient objects where a solution that completely includes one object may be desired over one which partially includes several objects. A simple example of this is shown in Figure 4.4. A second subtle problem with $W^{c}_{\mathbf{T}}$ is that it implies that all portions of the image window which do not contain text should contain salient content. This is a strong assumption which is generally inappropriate as in the case of Figure 1.2. Intuitively, a poster which satisfies this assumption is likely to be too cluttered, especially as an advertisement. As a simple solution to these shortcomings a center weighted content window is computed as follows,

$$\begin{aligned} W_{\mathbf{C}}(x,y,\mathbf{\Gamma}) &= W^{c}_{\mathbf{T}}(x,y,\mathbf{\Gamma}) \cdot \mathrm{G}(x,y|\boldsymbol{\mu}_{TW},\sigma_x,\sigma_y) \\ &= (1 - W_{\mathbf{T}}(x,y,\mathbf{\Gamma})) \cdot Ae^{-\frac{1}{2}\left(\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right)} \end{aligned} \qquad (4.7)$$

where $\mathrm{G}(x,y|\boldsymbol{\mu}_{TW},\boldsymbol{\Sigma}_{TW})$ is a 2-dimensional Gaussian window over image coordinates centered at position $\boldsymbol{\mu}_{TW}$. For resizing images without text the image center is a natural choice for $\boldsymbol{\mu}_{TW}$, however the focal point of a poster should be influenced by the positioning of the heading. This can be accommodated using the image centroid of $W^{c}_{\mathbf{T}}(x,y,\mathbf{\Gamma})$ as the center position. Furthermore, the standard deviation parameters are set to a fixed fraction $k_w$ of the image dimensions (i.e. $\sigma_x = k_w n'$ and $\sigma_y = k_w m'$). This allows the weighting function to be adapted for various image dimensions and aspect ratios. An example content weighting function for $k_w = 1/3$ is shown in Figure 4.3(c). Notice that the use of the centroid to compute the Gaussian center point has shifted focal point away from the center to accommodate for the region covered by the text. The small weights far from the image center also account for uncertainty regarding the image content at the edges of the image. Specifically, these regions can have large text relevance or importance without significantly effecting the value of $\mathrm{O}(\mathbf{\Omega})$.

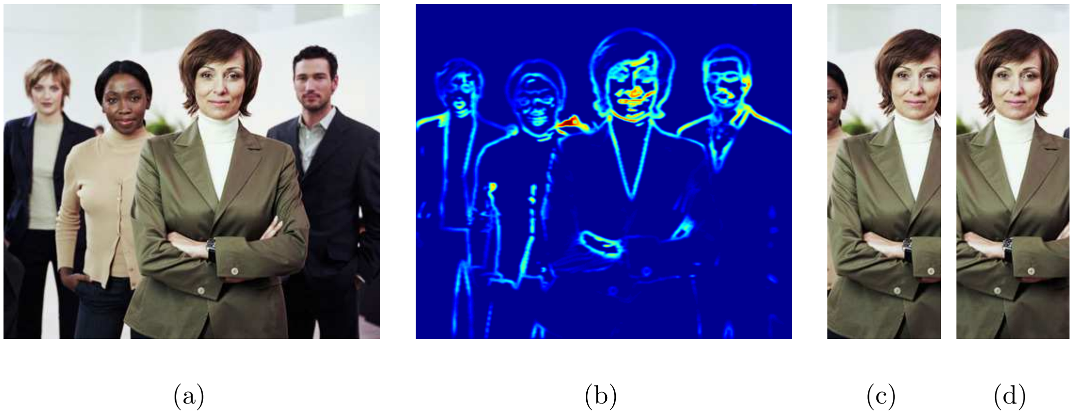


Figure 4.4: Example demonstrating the effects of a center weighting window in $C_{cont}(\mathbf{\Omega})$. (a) Original image (b) Computed importance map from text-surround model. Optimal subwindows for $\gamma = 0.3$ with (c) uniform weighting and (d) center weighting.

The final terms in the objective function which must be discussed are the are the normalization factors $Z_{text}$ and $Z_{content}$. These terms are very important because the text relevance and image importance maps have very different properties and generally do not share common units. The text relevance is often populated with many large values with low responses in small regions corresponding to important objects while the image importance map is likely to be the opposite, primarily low responses with a few large responses in key areas. Thus the ranges of the sums in $C_{cont}(\mathbf{\Omega})$ and $C_{text}(\mathbf{\Omega})$ can be significantly different. To adjust the cost terms to a common scale each is normalized by largest observed value over all the solutions which are considered. Denoting the searched solution space $\mathbf{S}_{\Omega}$ the normalization terms are defined as

$$Z_{text} = max\left\{\mathbf{\Omega} \in \mathbf{S}_{\Omega} : C_{text}(\mathbf{\Omega})\right\} \tag{4.8}$$

$$Z_{content} = max\left\{\mathbf{\Omega} \in \mathbf{S}_{\Omega} : C_{cont}(\mathbf{\Omega})\right\} \tag{4.9}$$

Normalizing relative to the solutions in $\mathbf{S}_{\Omega}$, rather than the set of feasible solutions, provides a form of non-parametric regularization because $\mathbf{S}_{\Omega}$ is constructed based on our prior expectation of the new resizing parameters. By normalizing relative to these cases, extreme solutions (large $C_{cont}$ and small $C_{text}$ or vice versa) are discouraged in favor of balanced solutions. The construction of $\mathbf{S}_{\Omega}$ will be discussed in Section 4.4.

The final resizing objective function is fairly straightforward and easy to implement. The objective function is non-convex and may have several extrema but because the solution space is discrete, an optimal solution can be found using an exhaustive search. As formulated above, a resizing solution has at least 3 free parameters resulting in cubic search complexity. Although this may be considered computationally demanding, many exiting resizing algorithms have been able to perform these searches reasonably quickly [6][45]. In this particular case, only a subset of the set of feasible solutions is actually searched which helps reduce the time required for the optimization procedure. In practice, the majority of computation time is spent on text rendering and computing the weighting $W_{\mathbf{C}}$ and $W_{\mathbf{T}}$. For a given $\mathbf{\Gamma}$ the family of resizing solutions over the window offset vector $\mathbf{o}$ share the same content and text weighting functions which can dramatically reduce the overall computation time.

## 4.3 Text Layout

Thus far text layout has only been discussed at a high-level, abstracting away the word wrapping problem which must solved for automatic poster resizing. Word wrapping (or line breaking) is the operation of adding line breaks to sentences (or paragraphs) of text to make them fit within a fixed width frame. Alternatively, this can be described as "breaking paragraphs into lines" [31]. The text in advertisements is generally manually arranged by the publisher but to facilitate a text size search we must do this automatically. This issue is also important for visual appeal because artifacts such as a single stranded word following a long line of text can be distracting.

There are two standard methods typically used for word wrapping in large bodies of text, minimum length and minimum raggedness [31]. The minimum length method represents the simple greedy algorithm for word wrapping. We sequentially go through the text from start to finish and place the maximum number of words on each line before inserting a line break. Although straightforward to implement, this approach typically leads to unbalanced line lengths (i.e some very long or short lines) as shown in Figure 4.5(b). The standard methods employed by documentation preparation systems such as LaTeX, use some variant of the minimum raggedness criteria. This algorithm selects the line break positions which minimize the absolute sum of squared (or cubed) unused line space at the end of all lines [31].

For headings in posters we present an alternative objective function which promotes neighboring lines to have similar lengths, the corresponding cost function is expressed as,

$$C_{bal}(\mathbf{l}) = \sum_{i=2}^{N_l} (\mathbf{l}(i) - \mathbf{l}(i-1))^2 + (N_l - 1) \cdot l_{max}^2 \tag{4.10}$$

where $N_l$ is the number of lines, $\mathbf{l}$ is the vector of (occupied) line lengths, and $l_{max}$ is the maximum length of a line. The second term in the above expression enforces the constraint that the minimum number of lines are used. The line lengths are computed as,

$$\mathbf{l}(i) = \left( \sum_{w \in \mathbf{W}(i)} \text{length}(w) \right) + \left( N_{\mathbf{W}(i)} - 1 \right) \cdot l_{blank} \tag{4.11}$$

where $\mathbf{W}(i)$ is the ordered set of words on line $i$, $N_{\mathbf{W}(i)}$ is the cardinality of $\mathbf{W}(i)$, and $l_{blank}$ is the length of a blank space. For small headings this cost function often yields the same result as minimum raggedness, however in cases with many similar solutions the balanced criteria favors those with smoothly varying line widths. Examples demonstrating these differences are shown in Figure 4.5. Also note that although the number of possible word wrapping solutions is exponential in the number of lines, an optimal solution for $C_{bal}$ and $C_{rag}$ can be found efficiently using a dynamic programming algorithm[14][31]. For short headings, solutions can also be found quickly via recursive backtracking by pruning the search tree once the remaining text fits on a single line.

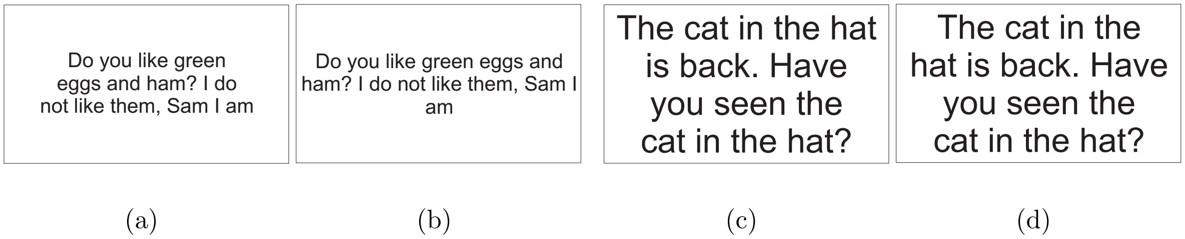


Figure 4.5: Examples demonstrating alternative approaches for automatic line wrapping. (a) Balanced Wrapping (b) Greedy Wrapping (c) Balanced Wrapping (d) Minimum Raggedness

Overall, there is still some subjectivity regarding the best line breaking method for a given poster. In addition, the best choice is going to be dependent on the poster's image content. In general, performing wrapping with $C_{bal}$ or $C_{rag}$ yields satisfactory results, however $\mathbf{S}_{\Omega}$ can also be expanded to include solutions with multiple wrapping methods if text layout is important to the poster publisher.

## 4.4 Constrained Optimization

Technically, a poster can be resized by searching the complete set of feasible parameter vectors for an optimal solution. However, $\mathrm{O}\left(\boldsymbol{\Omega}\right)$ does not maintain some of the high-level structure which may have made the original poster visually appealing. For example, if the original poster contained centered text it is usually desirable for the resized poster to contain centered text. In general, modeling these preferences is a challenging task because it requires semantic analysis of the poster content. This work takes an alternative and much simpler approach to address these issues. Specifically, the set of potential resizing solutions is constrained to maintain an analogous parameterization to the original poster. This method introduces limitations but also makes the resizing framework more predicable and makes posters easier to design.

The searched solution space will be referred to as the set of *acceptable solutions* and denoted $\mathbf{S}_\Omega$ as above. Formally, the solution to a resizing problem is computed as,

$$\boldsymbol{\Omega}_{opt} = \underset{\boldsymbol{\Omega}\,\in\,\mathbf{S}_\Omega}{\operatorname{argmin}}\ \mathrm{O}(\boldsymbol{\Omega}) \tag{4.12}$$

Recall that the free parameters of a resizing solution are the window offset $\mathbf{o}$, text center $\mathbf{c}$, and text size $s$. The window offset vector is not constrained because ideally the image importance map identifies the key regions of the image and promotes proper centering of the content. Conversely, the text relevance generally identifies large regions which are suitable for text placement. This requires the text parameters to be constrained. For example, there may be a large area of the image which can accommodate text but from the publisher's perspective a large font size may be unappealing. Therefore, the new text size is constrained to be within a fixed range $[s_{min}, s_{max}]$ where $s_{max}$ is always less than or equal to $s_0$. The text size range should be a function of the original size $s_0$ and the resizing factor $\gamma$. For example, linear expressions the form

$$s_{min/max} = (1 - k(1-\gamma)) \cdot s_0 \tag{4.13}$$

can be used for selecting search limits. Upper and lower limits should be computed by using different proportionality constants ($k$). For best results, these constants should also be set differently for row and column scaling.

The text position vector $\mathbf{c}$ is also sensitive to over-adjustment. For example, if text is centered horizontally small changes of the x-coordinate violate symmetry and reduce the poster's visual appeal. Similarly, adjusting the y-coordinates so that the text is very close to the upper or lower borders of the image can have a negative impact on a poster's appearance. We can handle these peculiarities by constraining the text position based on our knowledge of the original poster. The text search neighborhood is centered at the same (normalized) position in the resized frame as it was in the original poster. The neighborhood is then selected to be some fraction $\rho$ of the resized image dimensions along the search direction. As above, $\rho$ is set differently for row and column scaling, denoted $\rho_x$ and $\rho_y$ respectively. In the case of horizontally centered text ($c_x = 0.5$) this can be handled fairly easily. The set $\mathbf{S}_\Omega$ only includes solutions with centered text ($\rho_x = 0$) and only changes in the y-coordinate are considered. In other cases the search can be performed in a small neighborhood along both directions. Since most posters fall into the former category, it is the primary emphasis in this work.

## 4.5 Arbitrary Image Resizing

This section describes a simple procedure which can be used to perform general image resizing (where both dimensions are reduced) by combining uniform downsampling with the proposed aspect ratio change procedure. In this step, the poster is uniformly resized such that one dimension of the rescaled image matches the target dimension. Once this is performed, the final resized ad can be computed by changing the aspect ratio. This procedure is schematically illustrated in Figure 4.6.

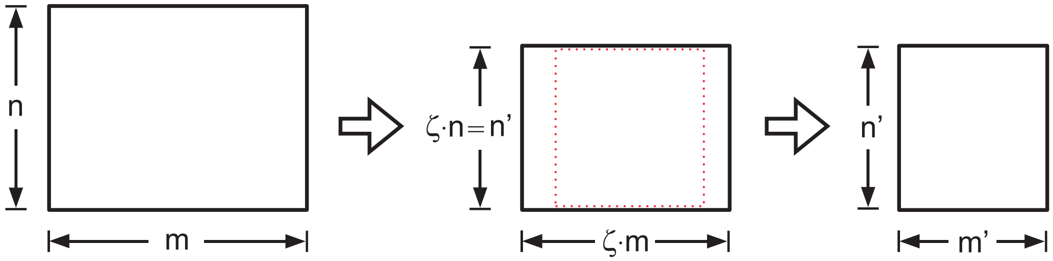


Figure 4.6: Illustration of preprocessing step which can be used for arbitrary ad resizing. Rectangular regions represent outer edges of posters.

Specifically, a poster $\mathbf{P}$ containing an $m \times n$ image $\mathbf{I}$ may be resized to arbitrary image dimensions $m' \times n'$ as follows,

1. Compute the uniform rescaling factor as $\zeta = max\left(\frac{m'}{m}, \frac{n'}{n}\right)$.

2. Perform uniform downsampling on $\mathbf{I}$ by a factor of $\zeta$ as described in Section 2.3.1. Denote the resulting image as $\mathbf{I}_r$. Compute the new poster text size $s = \zeta \cdot s_0$. Construct a new poster $\mathbf{P}_r$ using $\mathbf{I}_r$ and $s$.

3. Compute a rescaling factor as $\gamma = min\left(\frac{m'}{m\cdot\zeta}, \frac{n'}{n\cdot\zeta}\right)$.
4. Apply the proposed poster resizing algorithm to $\mathbf{P}_r$ along the dimension which does not match the target size with rescaling factor $\gamma$.

In the case that the aspect ratio of the resized image $\rho_r = n'/m'$ is the same as the original aspect ratio then the poster is only uniformly resized. This operation would not be acceptable for image retargeting (e.g for small displays), but it is suitable for automatically resizing ads to accommodate different website layouts. For example, Google Ad Builder applies uniform resampling to resize advertisements. In this sense, the proposed framework represents an enhanced ad resizing system which facilitates aspect ratio changes in addition to uniform rescaling.

# Chapter 5

# Experimental Results & Analysis

## 5.1 Introduction

The primary goal of this work has been to develop a system which can automatically resize online advertisements. The solution to this problem has two major components. The computation of importance and text relevance maps to characterize the structure of the original poster, and then an optimization framework which employs these maps to determine the parameters of a resized poster. One of the main challenges with both of these problems is their inherently subjective nature. This makes evaluating results and comparing different algorithms a challenge on its own.

This chapter contains results and analysis intended to assess the relative performance of the proposed methods against existing techniques. Both qualitative and quantitative measures are employed to achieve this goal. In the case of importance maps, comparative examples are presented as a qualitative measure. In addition to this, several new metrics for evaluating importance maps in the context of image resizing are proposed. With respect to poster resizing, examples are presented for subjective evaluation in conjunction with the results of a user study which provides a quantitative measure of performance. In all cases, the proposed approach is shown to decisively outperform alternative methods.

## 5.2 Importance & Relevance Maps

In the context of image resizing the importance map computation amounts to identifying the focal point or key objects in an image. Similarly, relevance maps identify suitable locations for text placement. Together, these two maps contain the core information which guides the resizing process. The maps are critical to the success of the resizing algorithm, thus it is crucial to select the best possible technique for map estimation. The goal of this section is to compare the proposed coarse center-surround (CCS) method to the standard Itti and Koch model typically used for image resizing.

Several different variations of importance maps can be computed from the model in Chapter 3. For clarity, each will be restated here with an appropriate name. Different maps can be created based on choice of energy function (disconnected or complete), sampling method (surround or text-surround), and optional boundary weighting procedure. As previously defined, the energy function solutions are referred to as,

> CRF – The solution to the connected energy function, where **A** is directly used as the importance map.
>
> ICRF – The solution to the disconnected energy function, again used directly as the importance map.

As before, boundary weighted importance maps can are denoted Boundary CRF (BCRF) or Boundary ICRF (BICRF). Unless otherwise stated, all maps will be computed using surround sampling. When importance maps are computed using text-surround distributions which use additional text region data this will be explicitly denoted by the prefix ‘P’ or the complete word ’Poster’. For example, PCRF, PBICRF, and PBCRF importance maps. For the purpose of poster resizing these last two methods are of primary interest, however the other formulations are presented to provide insight into solutions of energy-based model and effects of text region sampling. Finally, all maps shown in the subsequent sections are computed using the same parameter set, corresponding to the values presented in Chapter 3.

Within the proposed saliency model, text relevance and unweighted image importance (CRF or ICRF) are closely related. Thus, to avoid redundancy all results are demonstrated in the context of importance maps.

### 5.2.1 Qualitative Results

The most straight-forward approach for demonstrating the performance of an importance (or saliency) map algorithm is through a case study of a test image set. This is necessary for a complete evaluation of such an algorithm and examples help provide intuition and insight into the results each technique yields.

The examples shown here, in conjunction with the top and bottom images in Figure 3.5, form the test image set which will be used throughout the importance map evaluation sections. The selected images are intended to be typical examples of advertisements and were selected *without* prior knowledge of the resulting importance maps.

Examples are provided in small groups followed by observations and discussion. For comparison, the Itti and Koch (IK) saliency map for each image is also shown. The IK saliency maps have been computed using the Saliency Toolbox [56] which is developed by the authors' research group. For clarity, images will be identified by number according to their row in the image set from top to bottom (i.e the top row corresponds to image one).

The first example set, shown in Figure 5.1 shows importance maps computed using the PCRF and PBCRF techniques. These are the most relevant in the context of poster resizing. This image set demonstrates many general properties of the respective importance maps. Firstly, since all techniques (including IK) are based on center-surround they often have maxima in similar image locations, as in images one to three. Generally speaking, the CRF and PBCRF maps are more strongly localized than the IK maps.

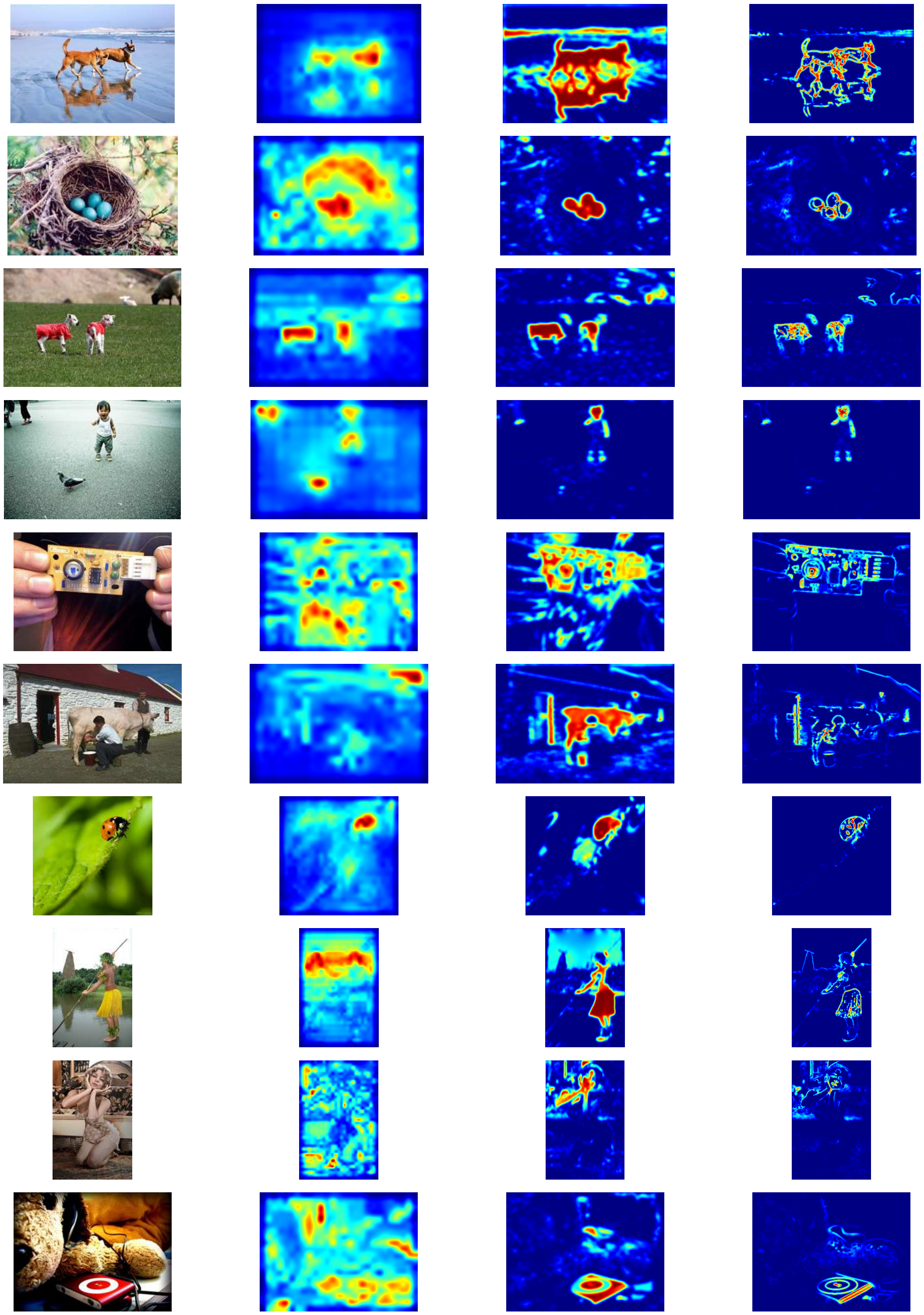


Figure 5.1: Poster Importance Map Examples. Each row represents a set of importance maps for a given poster. From Left to Right: Poster Image , Itti & Koch Saliency Map, Poster CRF , Poster BCRF

As shown, the proposed techniques can be applied to images of various sizes and aspect ratios without adjusting any system parameters. The boundary weighting operation reduces noise in the maps and focuses importance at complex image regions and object boundaries. In some cases the boundary weighting procedure results in sparse maps as shown in images six and nine. Note that the sparsity does not generally degrade resizing performance, it actually improves results in many cases.

The results in Figure 5.1 suggest that the proposed model yields superior importance maps for poster resizing however it is insightful to evaluate the resulting maps when the text region information is not utilized. This additional information does increase the accuracy and robustness of the procedure, but computing maps without text region information (surround sampling) also yields very good results. For example, the CRF and BCRF maps for images one through five are nearly identical to poster versions shown in Figure 5.1, similarly the BCRF maps for images six and seven are very similar to the Poster BCRF maps shown. Often the text region sampling reduces erroneous maxima in the importance maps but does not drastically change results. Figure 5.2 shows examples where these differences are significant. In particular the bottom two images demonstrate cases where the CRF saliency method performs poorly while the PCRF method does well. When the surround-model fails this additional robustness appreciably improves resizing results. Finally, note that the IK method does not result in accurate maps in any of these cases. The results from the surround distribution indicate that the proposed saliency model, even without text-region information, improves importance map estimates.

The next set of saliency maps illustrate differences between the ICRF and CRF techniques. Recall that these maps are computed by minimizing the disconnected and connected energy functions respectively. One of the main complications with the CRF importance map model is the complexity of computing $\boldsymbol{\alpha}$ because this involves solving a very large system of equations. As an alternative, the BICRF importance maps are computationally simpler and the algorithm is highly parallizable. The downside of this is

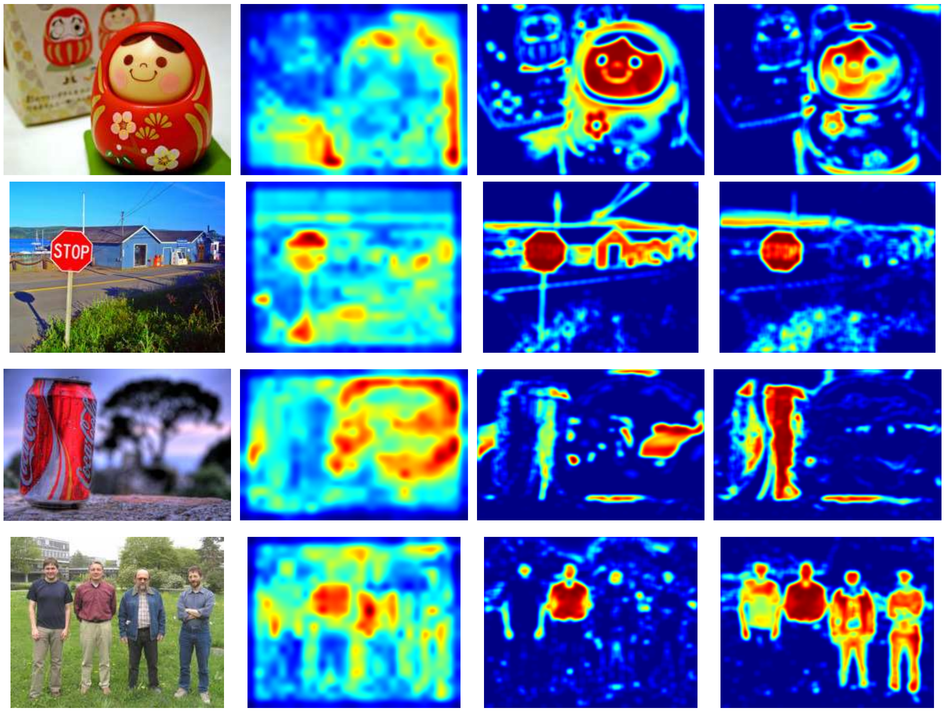


Figure 5.2: Examples demonstrating difference between importance maps generated using surround distributions (CRF) and text-surround distributions (PCRF). From left to right. Original , Itti & Koch, CRF , Poster CRF

the loss of neighborhood smoothing terms which generally reduce noise in the importance map estimates. Regardless, the BICRF technique still produces high fidelity importance maps. Examples demonstrating these differences are shown in Figure 5.3. The first three images are typical examples where the CRF maps are less noisy. The final image provides an example where the smoothing has a negative effect by reducing saliency in an important region. However, this is rarely a problem as the positions of significant local maxima are maintained.

Overall, the examples illustrate high quality importance maps which can be generated using the coarse center-surround model with energy-based differences. Results also show the effectiveness of using text region information to improve results.

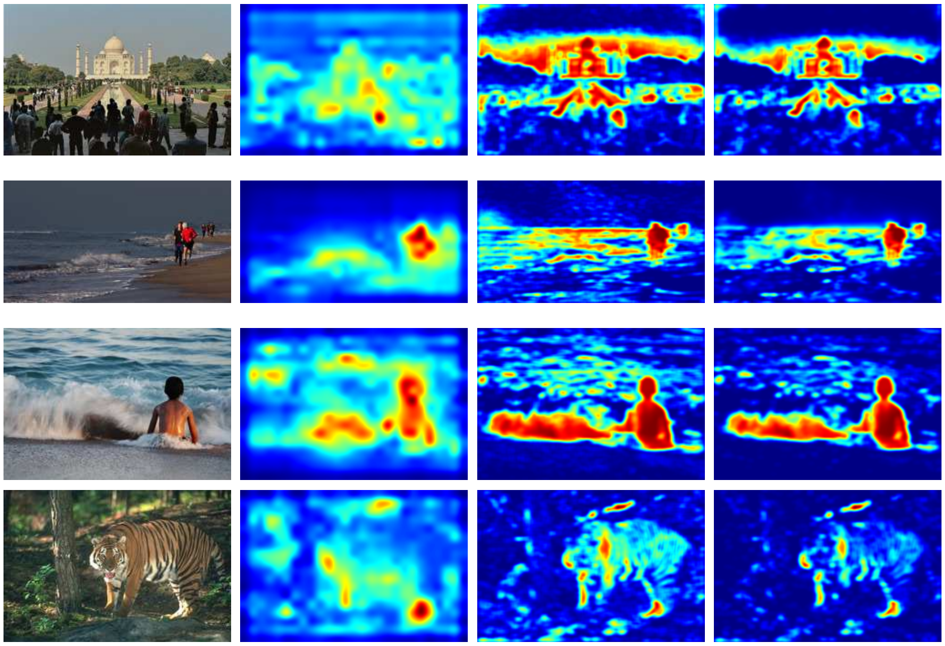


Figure 5.3: Examples demonstrating difference between the connected (CRF) and disconnected (ICRF) energy functions. From left to right. Original , Itti & Koch, ICRF , CRF

### 5.2.2 Quantitative Results & Analysis

Due to their subjective nature saliency maps are difficult to objectively evaluate with a quantitative metric. When metrics are used they are generally application specific. For example, saliency maps are often evaluated against fixation maps collected from human eye-tracking data [17][41]. One of the major challenges with developing importance maps for high-level tasks such as image resizing is the absence of a common set of metrics for comparing different techniques.

This section presents a set of quantitative measures which have been developed to compare different importance maps specifically for image resizing. The general approach here is motivated by image segmentation testing methods. Specifically, ground truth datasets are created by manually identifying the most important or ‘key’ regions in an image. This is done by constructing binary masks which identify the focal point(s) of an image. In many respects, the shapes of these masks are similar to segmentation data but with background or secondary regions discarded. Example masks are shown in Figure 5.4 with boundaries of the most salient regions drawn in green. Mask regions are constructed by taking high-level information into consideration. In particular, they are chosen to include as much of the most salient objects as possible without including secondary regions which could yield incorrect resizing results. For example, if we have an image of a man playing a guitar, the mask may include his upper body and guitar but not his shoes (unless something is interesting about them). A saliency map with all its mass concentrated at his shoes would be considered a very poor estimate of importance, but large importance around his face or the guitar would be reasonable. In the following paragraphs these binary maps will be referred to as importance masks, and denoted **M**.

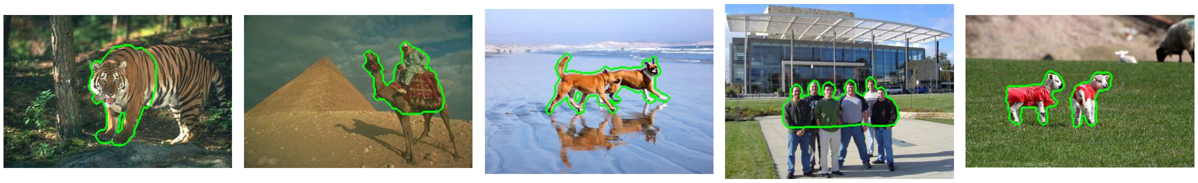


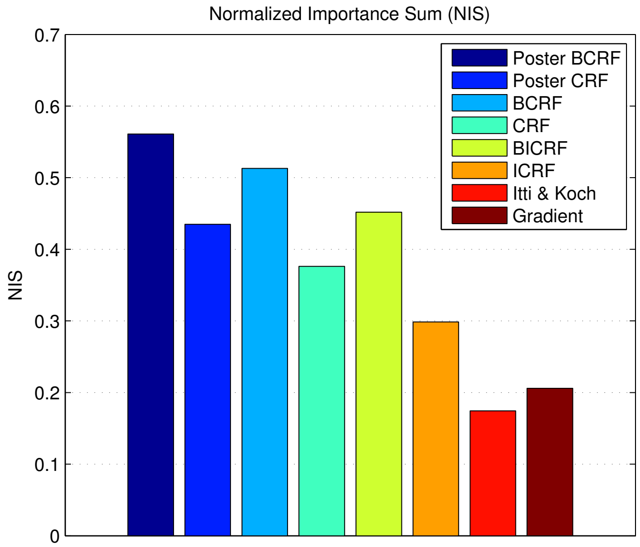
Figure 5.4: Saliency masks from the test set with the borders of the most salient region shown in green.

A group of importance masks was created for the test image set from Section 5.2.1, the mask construction was performed with feedback from three individuals who all agreed on the final masks. The availability of importance masks introduces several potential methods for evaluation. It is important to develop metrics which correspond well to qualitative observations, and different measures usually have different strengths and limitations in this regard.

The first proposed metric, is highly intuitive and straight-forward to compute. We define the ‘normalized importance sum’ for an importance map $\mathbf{S}$ with mask $\mathbf{M}$ as,

$$NIS = \frac{\displaystyle\sum_{\mathbf{p}\in\mathbf{M}} \mathbf{S}(\mathbf{p})}{\displaystyle\sum_{\mathbf{p}} \mathbf{S}(\mathbf{p})} \tag{5.1}$$

This corresponds to the percentage of image importance located with the importance mask and is used as a measure of fidelity. Although simple, the computed values strongly agree with qualitative observations and resizing performance. The mean NIS results computed for the test image set are shown in graphical and tabular form in Figure 5.5. These



| Algorithm | NIS |
|---|---|
| Gradient Magnitude | 0.21 |
| Itti and Koch | 0.17 |
| ICRF | 0.3 |
| Boundary ICRF | 0.45 |
| CRF | 0.38 |
| Boundary CRF | 0.51 |
| Poster CRF | 0.43 |
| Poster BCRF | 0.56 |

Figure 5.5: Mean NIS for different importance map algorithms.

NIS results demonstrate similar trends as the qualitative evaluations. The connected energy model outperforms the disconnected model, the ‘poster’ maps, which use text region information, outperform their surround only counterparts, and the boundary weighting procedure improves the ‘accuracy’ of resulting maps in general.

By viewing saliency maps under a discriminative framework we can formulate a related metric which provides more insight into the distribution of importance values within the map. A similar approach was first proposed in [16], where the authors use a precision-recall (PR) curve [10][37] to evaluate the ability of their saliency model to predict human eye fixation data. Within this context, the importance maps can be used to construct 'important region classifiers', where all importance values above a threshold are classified as 'important'. Under this model common metrics such as precision and recall can be employed to evaluate different techniques.

Precision is computed as the fraction of classifications within the importance mask region. A precision curve illustrates a classifier's accuracy over the complete range of decision thresholds. Thus, each point on this curve represents the precision for a given threshold. This curve can be estimated by normalizing the importance map values to the range $[0, 1]$ and then computing (sampling) precision values for various thresholds within this range. The precision curve for the test set is shown below in Figure 5.6, overall the results are similar to those computed using the NIS metric. However, note that there is no point on the NIS curve which is equivalent to the precision curve because the latter does not actually accumulate importance values.

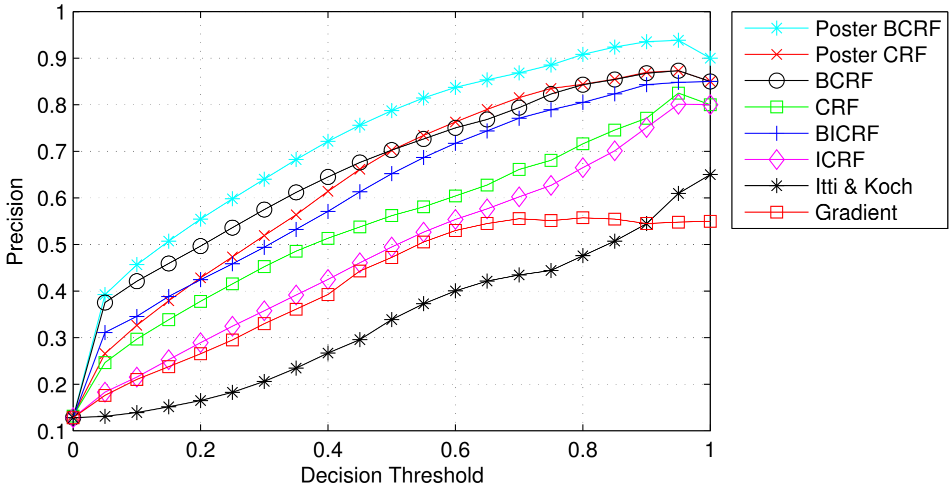


Figure 5.6: Precision curve computed for different importance map algorithms using the test image set.

Next, the recall corresponds to the fraction of the importance mask which is successful classified as important for a given decision threshold. In the context of image resizing performance the recall is much less significant than precision. Although somewhat desirable, it is not crucial that an importance map has strong responses for all pixels within the mask region. Experimental results suggest that erroneous importance values outside the mask region are the primary sources of problems because they cause resulting posters to be poorly centered. Sparse maps with importance allocated at object boundaries may have low recall values but are well localized and produce centered resizing results. Regardless, given two methods with similar precision, the map with a higher recall value will be more informative. There is also a trade-off between precision and recall and this provides insight into different importance map algorithms. A PR curve can be used to express this relationship. This curve is estimated by computing point pairs of the form $(R(\tau), P(\tau))$ for each threshold, $\tau$. A precision-recall curve computed from the importance test image set is shown in Figure 5.7. In the current work we have observed that the area under the PR curve is not completely consistent with observed resizing performance. As previously mentioned, this is attributed to the fact that precision is much more important than recall within this domain. Conversely, this curve still provides insight into the properties of different algorithms and has been included here for completeness.

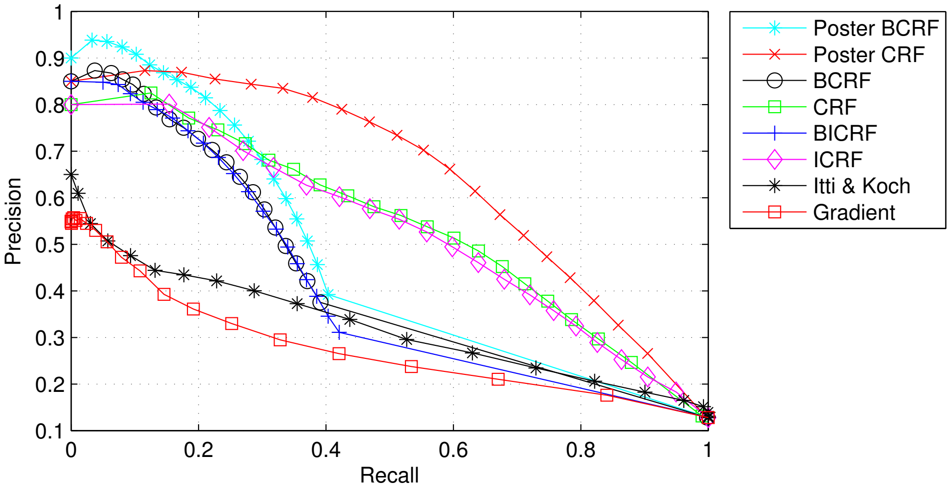


Figure 5.7: Precision-Recall curve computed for different importance map algorithms using the test image set.

The major limitation of the metrics based on saliency masks is that the spatial distribution of importance within the mask region is not considered. This can be addressed by using more complicated multi-level masks, but this introduces additional subjectivity into their construction. As an alternative, we can identify key focal points in the image and then compare these positions to the centroid of the importance map. The euclidean distance between these two points will be referred to as the localization error (LE). This measure works particularly well for images with a single focal point, but even when images contain multiple important region we can take the LE to be the distance between the centroid and nearest focal point.

For images typically used in AIR systems, the perceived focal points are almost 'obvious' and do not vary significantly between individuals. This is particularly true here as multiple points can be selected for each image to satisfy different opinions. The current test point set was constructed by three people and there was very little disagreement on point placement. Example images illustrating focal point positions are shown in Figures 5.8. The mean localization error for images in the test image set is shown in graphical and tabular form in Figure 5.9. To ensure proper averaging, image positions are represented in normalized coordinates. The table also shows the mean absolute localization errors along the individual x and y directions. Overall, the localization results closely mirror the NIS results. The LE was also computed only considering pixels within **M** with interesting results. In this case, we observe similar performance across all techniques. This identifies an important limitation of the center-surround assumption.

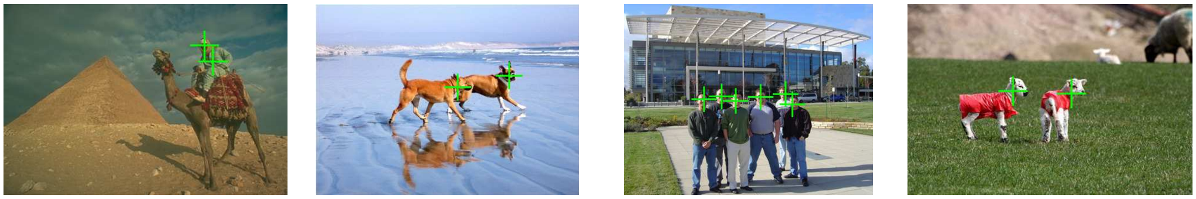
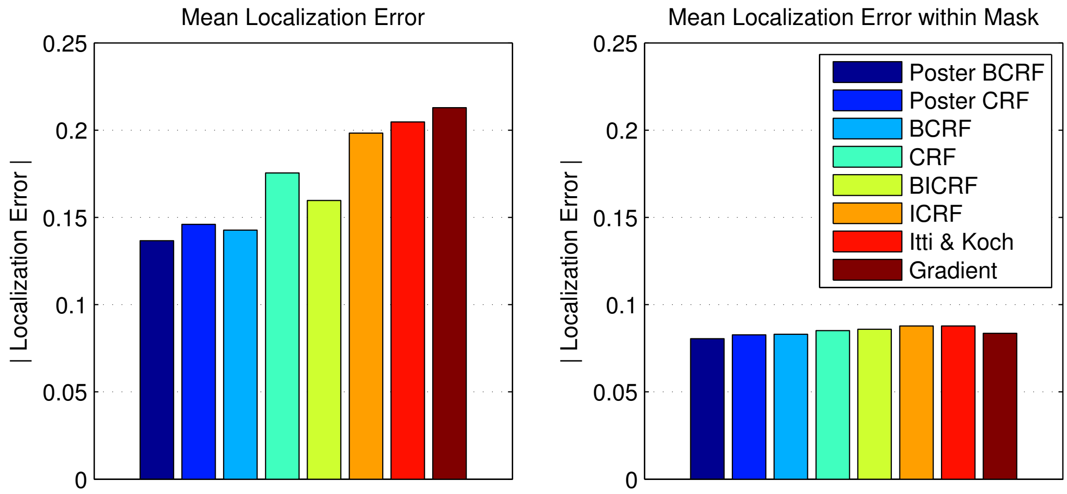


Figure 5.8: Example focal point positions shown for images in the test image set.



| Algorithm | \|Loc. Error\| | $\|X_{Error}\|$ | $\|Y_{Error}\|$ | \|Loc. Error\|$_{MASK}$ |
|---|---|---|---|---|
| Poster BCRF | 0.137 | 0.109 | 0.065 | 0.080 |
| Poster CRF | 0.146 | 0.11 | 0.082 | 0.083 |
| Boundary CRF | 0.143 | 0.103 | 0.075 | 0.083 |
| CRF | 0.176 | 0.123 | 0.096 | 0.085 |
| Boundary ICRF | 0.16 | 0.117 | 0.083 | 0.086 |
| ICRF | 0.198 | 0.142 | 0.106 | 0.088 |
| Itti and Koch | 0.205 | 0.151 | 0.103 | 0.088 |
| Gradient Magnitude | 0.213 | 0.167 | 0.103 | 0.084 |

Figure 5.9: Mean Localization error for different importance map algorithms.

Although, it can be effective for finding important objects at a large scale relative to the image dimensions, it is weakly connected to the high-level problem of locally identifying important regions on an object of interest. Hence, the mid and fine scale center-surround used in IK saliency maps does not improve LE on the objects of interest. These results also suggest that an alternative model (not based on center-surround) would be more appropriate for locally estimating saliency within important regions.

In summary, we introduced a new approach for evaluating importance map algorithms in the context of AIR systems. This new structure facilitates the computation of many intuitive and common metrics, including those described above and potentially many more. Based on the proposed measures, the CCS energy-based saliency maps outperform conventional Itti and Koch saliency maps.

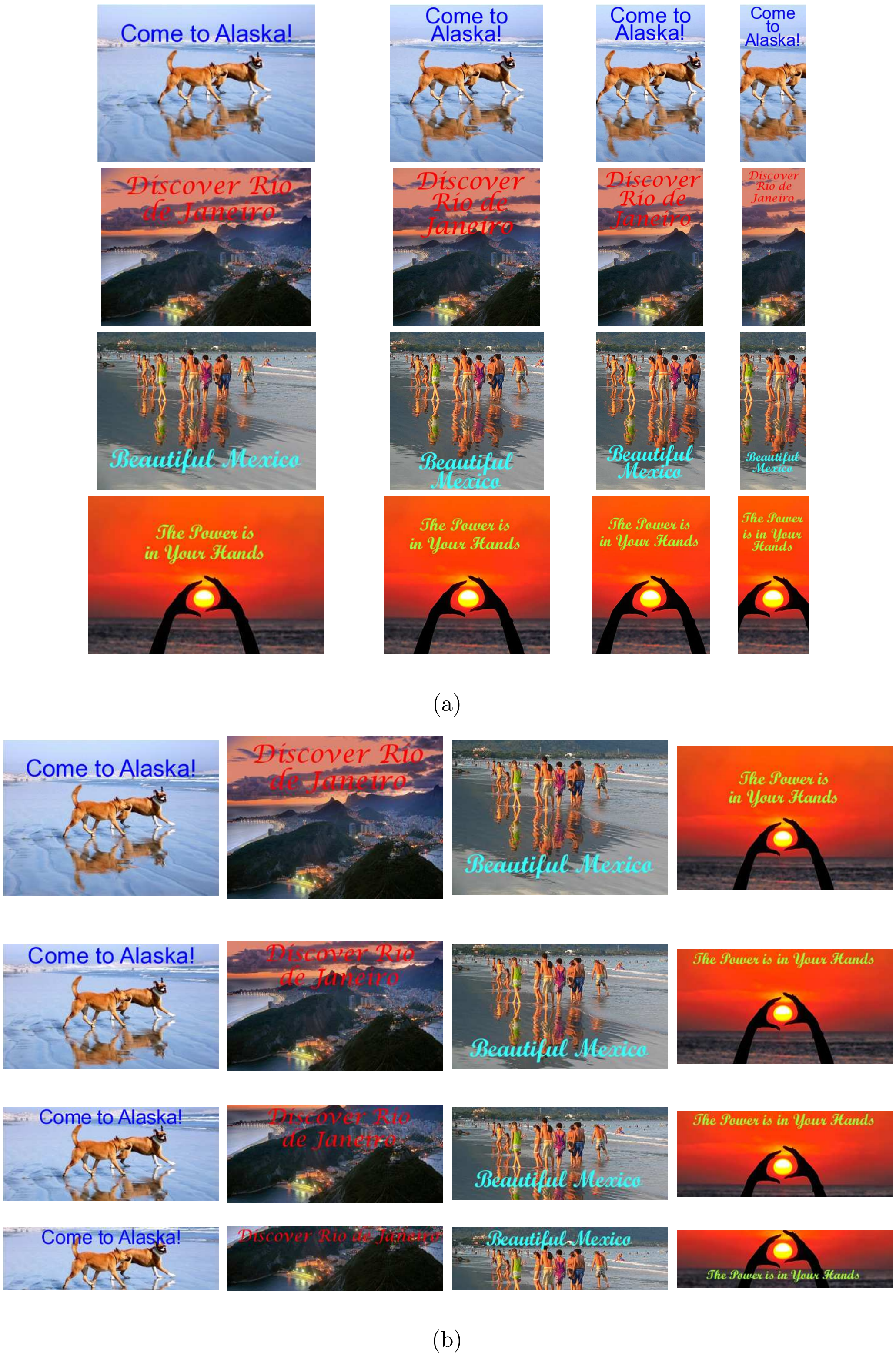
## 5.3 Automatic Poster Resizing

This section is intended to provide an evaluation of the proposed image resizing framework in an analogous fashion to the above discussion regarding importance maps. This is accomplished through a qualitative analysis of automatic resizing results in addition to the results of a user study comparing various alternative resizing algorithms.

### 5.3.1 Qualitative Results

The following section illustrates several automatic ad resizing results including successful and unsuccessful resizing attempts. The resized posters were all computed using Poster BCRF importance maps and Poster CRF text relevance maps. Resizing solutions were found by maximizing the objective function in Equation (4.1) with $\mathbf{w} = [\,1.0 \quad 0.3 \quad 0.7 \quad 0.1\,]$ and the balanced criteria was used to perform line wrapping.

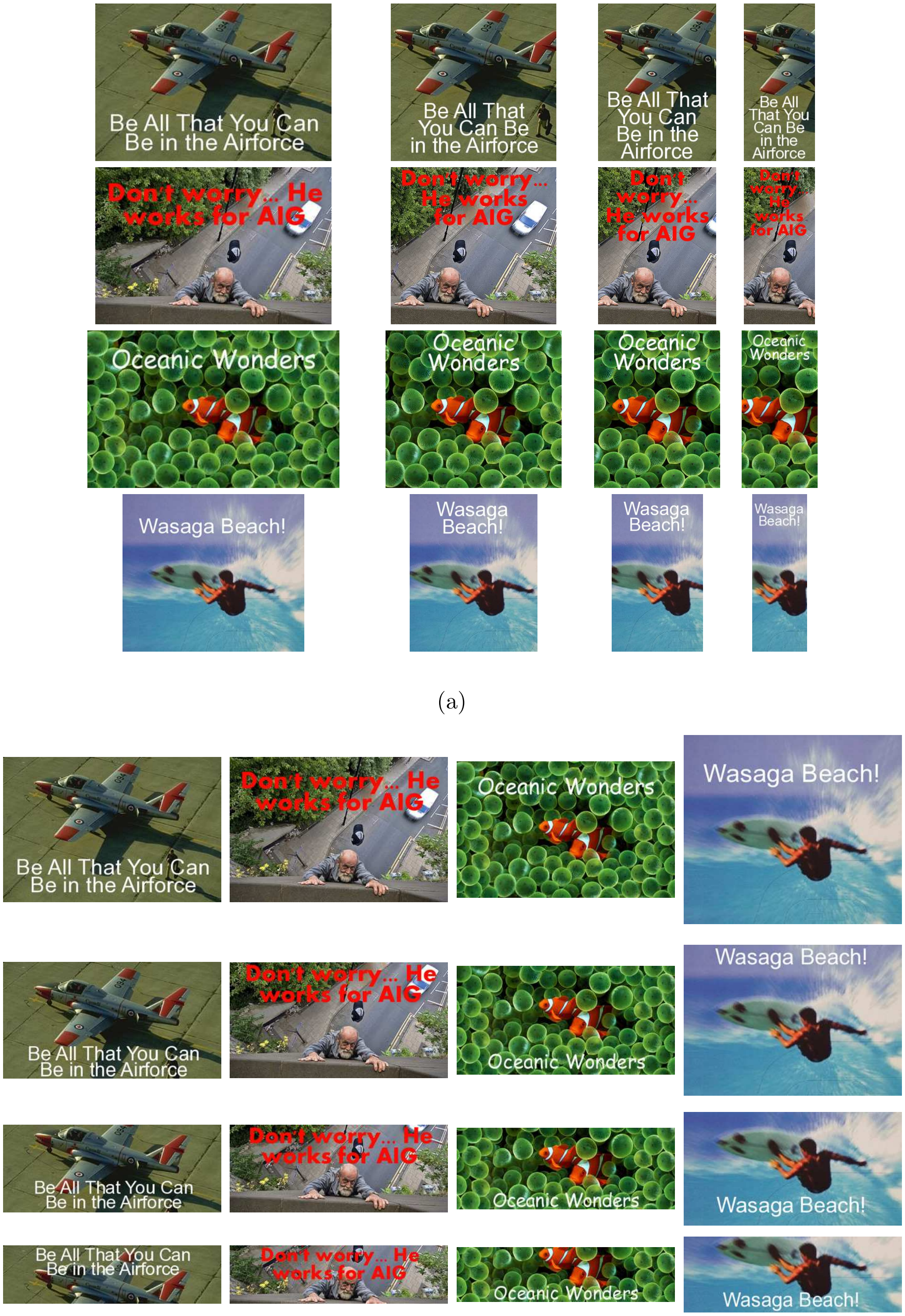
A series of resizing results along both horizontal and vertical dimensions are shown in Figures 5.10 and Figures 5.11. The same convention as in Section 5.2.1 will be used to refer to each example poster, specifically images are referred to by row number (top to bottom) or column number (left to right). In these examples posters were resized by predefined factors ranging from $\gamma = 0.8$ to $\gamma = 0.3$. These images have aspect ratios $\rho$ less than 1 which allows them to be resized more aggressively in the column dimension, thus row reduction is only performed up to $\gamma = 0.4$. The results show how this method can accommodate a wide variety of imagery including both natural and man-made scenery. These examples also illustrate several challenges associated with low-level techniques for poster (or image resizing). Consider the images in Figure 5.11(a). The second poster



(a)

(b)

Figure 5.10: Poster Resizing Examples. (a) Column reduction. (From left to right) Original, Resized with $\gamma = 0.7$ , 0.5, and 0.3 respectively. (b) Row Reduction. (From top to bottom) Original, Resized with $\gamma = 0.8$, 0.6, and 0.4 respectively.


Be All That You Can Be in the Airforce
Don't worry... He works for AIG
Oceanic Wonders
Wasaga Beach!

(a)

(b)

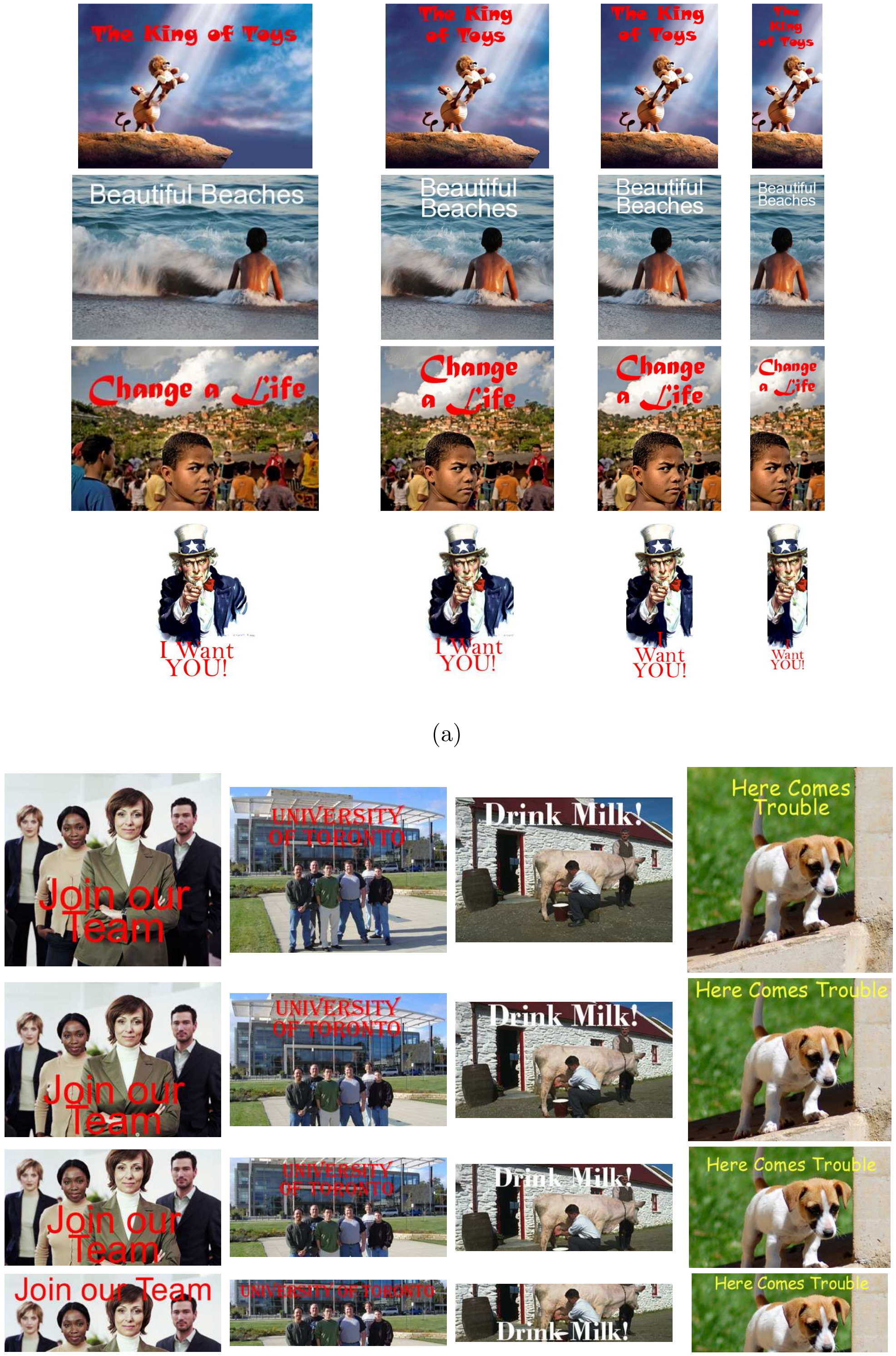
Figure 5.11: Poster Resizing Examples. (a) Column reduction. (From left to right) Original, Resized with $\gamma = 0.7$ , 0.5, and 0.3 respectively. (b) Row Reduction. (From top to bottom) Original, Resized with $\gamma = 0.8$, 0.6, and 0.4 respectively.

(row 2) demonstrates a semantic problem because it contains a secondary object (i.e the car) which could be of high-level importance. The midrange resized poster ($\gamma = 0.4$) includes a small portion of this car at the cost of the man not being centered within the frame. Whether the chosen solution is better than a centered version is totally subjective. Similarly, the highly resized poster of image four in Figure 5.11(a) may be considered poorly centered, however this depends on the relative importance of the surf board in the advertisement.

The most problematic semantic shortcoming is demonstrated in the resizing results for poster three. The highly resized poster has cutoff the front of the leftmost fish. Unlike the previous issues, this is generally undesirable from an objective point-of-view. However, the sub window must cut through ‘important’ regions simply because they cannot fit in the image window. A low-level model lacks the ability to identify the ‘front’ of the fish. Simple heuristics such as penalizing importance at sub-window edges can correct the problem for this poster, however these modifications tend to degrade performance when applied to other posters. Despite the above limitations, most examples here are visually appealing and would be satisfactory for automatic ad resizing applications.

Beyond importance maps, the major factor which influences the performance of automatic ad resizing is the spatial organization of key objects in the image. Specifically, the cropping approach requires that there is a suitable region on the poster of the desired size which can accommodate text. This can become a significant issue as $\gamma$ decreases below 0.5, particularly in the case of row reduction because text cannot be decreased in height without significantly affecting readability. Conversely, the ability to wrap text in conjunction with column reduction adds flexibility and allows smaller $\gamma$ for horizontal resizing. Examples related to difficult text placement when $\gamma$ is small are shown in the first and second images in Figure 5.11(b). Notice that both of these posters have relatively long headings which require a very small font size to fit on a single line.

In many cases, images are well suited for resizing along a preferred direction, this



(a)

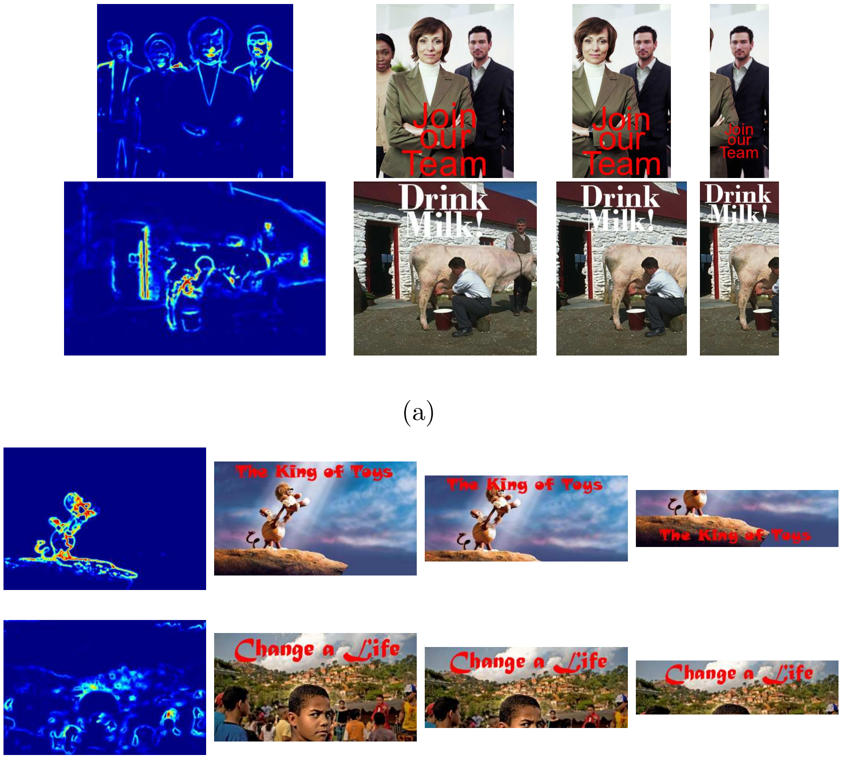
(b)

Figure 5.12: Poster Resizing Examples. (a) Column reduction. (From left to right) Original, Resized with $\gamma = 0.7$ , 0.5, and 0.3 respectively. (b) Row Reduction. (From top to bottom) Original, Resized with $\gamma = 0.8$, 0.6, and 0.4 respectively.

direction can usually be inferred from the image content itself. For example, the images in Figure 5.12(a) are well suited for column reduction because the objects of interest can fit well into thin vertical windows but are difficult to crop into horizontal windows ($\gamma < 0.5$) without losing important content. Similarly, the first three images in Figure 5.12(b) are appropriate for row reduction because important regions (e.g faces) are oriented side-by-side in the images. Furthermore, the importance masks do not need to be as accurately localized on the faces for successful row reduction, while column reduction is subject to many of the semantic challenges discussed above. Finally, image four (column 4) is an example where image content imposes a lower limit on the amount of row reduction by cropping, regardless of importance map fidelity.

The second significant issue which limits the ability to sharply resize an image across in both direction is related to limitations of importance maps algorithms. Although the Poster CRF technique generally produced better results than IK saliency, there is still room for improvement. In particular, importance may be located on or near the important region of an image but in many cases it does not provide accurate information about which parts of a sub-region are the most important. This is a very challenging problem because of its high-level, context dependent nature and as previously suggested the center-surround paradigm may not be appropriate for this task. Figure 5.13(a) demonstrates a pair of posters where the importance maps are reasonable from a low-level perspective, but not ideal based on semantic interpretation of the image. Specifically, the importance is focused on the ‘wrong’ person in the first image and erroneously allocated to the door frame in the second. Figure 5.13(b) demonstrates two additional failure modes of poster resizing. The first image is an example where the image content limits the ability to resize through row reduction and the second example demonstrates a general failure of the importance map algorithm. On a positive note, in all these cases, posters can successfully be resized reasonably well along at least one direction.

The above discussion has omitted the role of text relevance in the resizing process,



(a)

(b)

Figure 5.13: Examples of unsuccessful resizing attempts (a) Column reduction (b) Row reduction

however this component is crucial for text size selection and also improves the resizing robustness in terms of image content selection. In effect, the text constrains the set of resizing solutions to those which satisfy a certain structure. The most significant example of these effects is demonstrated by the poster with heading 'Change a Life' which is resized by column and row reduction in Figures 5.11(a) and 5.13(b) respectively. The importance map estimate is very poor in this case, however the text placement enforces resizing solutions to have a similar structure, specifically text is placed on the sky region. For row reduction, this results in good text parameter selection and better resizing results than the importance map alone would suggest.

In some instances the text component has a negative influence, this is the case for column reduction of this poster because it encourages resizing window to include the sky region which causes important content to be cutoff. Nonetheless this image is not well suited for significant resizing in the vertical direction and posters such as these require a different model for 'good resizings'. For example, at some point perhaps the target poster should have centered text drawn on a background region of the poster – because there is no space to accommodate the key image content. The primary goal here is to demonstrate the proposed method provides a good basis for designing automatic poster resizing systems and these enhancements have been omitted to keep the implementation as simple as possible.

### 5.3.2 Quantitative Results & Analysis

The subjective nature of the automatic resizing problem makes it extremely difficult to define a non-biased metric which can be used to rank different algorithms. However, automatic ad resizing is still a fairly undeveloped technology so evaluation often becomes simply deciding if a given method 'succeeded' where another has 'failed', where success can be measure fairly objectively based on common preferences that people may have (e.g. Is important content missing or cutoff? Is the text directly on top of the foreground object?). Under these circumstances a user study is an effective method to measure the relative performance of different algorithms.

A user study was performed in a forced-choice format [33][45] where each participant was shown a series of image pairs and asked to select the one which was better in terms of image content, text position, and text size. The user study consisted of 12 images which were based on or resembled typical advertisements. Although we do not know of any other published methods for automatic ad resizing, different algorithms were constructed by extending popular approaches for standard image resizing. All together, posters were resized using six different techniques, 3 based on the proposed framework here and 3 alternative approaches. These are as follows,

1. Center (CEN) – Image rescaling by cropping the image around its center point. Text

size was linearly scaled proportionally to the rescaling factor. The text position is fixed to the same normalized coordinate in the resized image.

2. Saliency Sum (SSUM) – Image rescaling by cropping the region with the maximum total importance computed using Itti and Koch saliency maps. Text parameter adjustment was performed as in 1.

3. Seam Carving (SC) – Image rescaling by pure seam carving with a gradient energy function. Text parameter adjustment was performed as in 1

4. AAR-IK – The proposed automatic ad resizing framework with energy-based text relevance maps and Itti and Koch saliency maps for importance.

5. AAR-Grad – The proposed automatic resizing framework with energy-based text relevance and squared gradient maps for image importance.

6. AAR – The proposed automatic ad resizing framework with energy-based text relevance maps and weighted maps for image importance.

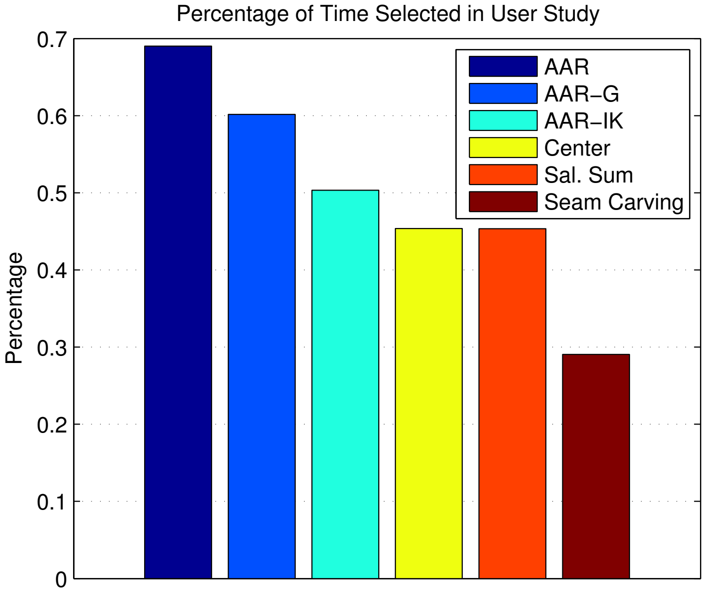
Resizing factors $\gamma$ were selected by randomly generating values within three ranges; large ($\gamma_L \in [0.6, 0.75]$), medium ($\gamma_M \in [0.4, 0.55]$), and small ($\gamma_S \in [0.2, 0.35]$ for column reduction and $\gamma_S \in [0.25, 0.35]$ for row reduction). A separate set of row and column rescaling factors was generated for each poster. All together this yielded 72 resized posters for each algorithm for a total of 432 posters in the study.

Under the forced-choice format the users are shown randomly selected image pairs and asked to choose the 'better' result. The users do not have any knowledge of which resizing algorithm has been used to create a given poster. In cases where the results are nearly identical (which typically happens when methods 4 to 6 are compared) users are told to break ties arbitrarily. This format was chosen because it avoids problems associated with subjective rating scales which participants often interpret differently.

In addition to comparing different resizing approaches, the user study provides an opportunity to compare various importance map algorithms. To take advantage of this three different AAR framework-based methods were included, each of which only differs by the importance map, $S_{\mathbf{I}}(x, y)$, used in the computation. An important goal of the user study was to demonstrate that the proposed approach for generating importance maps, in particular coarse-center surround with an energy based-model, is superior to IK saliency maps within this domain. To avoid any objections based on the computational complexity of solving the large systems of equations required for computing the Poster BCRF importance maps, the computationally simpler Poster BICRF maps were used. In the majority of cases, differences between these types of saliency maps do not substantially change the global minimum of the resizing objective function.

The posters presented in the user study are a subset of those shown in Section 5.3.1, however note that the user study resizings were independently computed using different importance maps and rescaling factors. To ensure a fair comparison, posters were chosen for resizing without any prior observation of any resizing results. Conversely, the results demonstrated in the qualitative analysis section still provide a good indication of the user study results. For example, poor resizing shown in Figure 5.13 are similar to those in the user study image set (where they are arguably the worst resizing results as well).

The user study was published online and participants were individually invited as a measure of quality control. There were a total of 26 participants who together rated over 2600 image pairs. Each individual provided a rating for between 50 to 150 image pairs at their own discretion. A large majority of participants did not have technical background in related areas, which avoids biases which can be introduced by individuals with knowledge of how each technique works. The overall results of the user study are shown in Figure 5.15. This figure shows the percentage of the time a given resizing method was selected when it was shown, denoted selection percentage (SP). The AAR techniques decisively outperformed the alternative methods, and they are the only algorithms se-



| Algorithm | Selection % |
|---|---|
| AAR | 0.69 |
| AAR-G | 0.6 |
| AAR-IK | 0.5 |
| Center | 0.45 |
| Saliency Sum | 0.45 |
| Seam Carving | 0.29 |

Figure 5.14: User Study Results: Overall selection percentage for all resizing algorithms.

lected at least 50% of the time. This suggests that the text relevance component alone improves resizing performance. The AAR technique with Poster CRF importance maps is the top performer by a significant margin, especially considering the subjectivity of evaluating resizing results. Also note that the results here include AAR methods compared against one another. Resizing results between these methods are often similar, and in these instances users will generally choose between them equally. This biases the SP of each toward 0.5. Considering comparisons between only AAR, CEN, SSUM, and SC the selection percentages are 0.73, 0.51, 0.47, and 0.29 respectively.

For an indication of relative performance between specific pairs of resizing algorithms a pair-wise selection matrix is shown in Figure 5.15. Each row of this table represents the selection percentage for a given algorithm compared to each of the others. For example, row 1, column 5 demonstrates that the AAR technique is selected 70% of the time when compared against SSUM resizing results. The upper-right $3 \times 3$ corner demonstrates the performance of all three AAR framework implementations compared to the alternatives. In particular, all AAR methods are selected the majority of the time against the alternatives, providing further evidence that the text relevance component

| | | Algorithm Compared Against | | | | | |
|---|---|---|---|---|---|---|---|
| | | AAR | AAR-G | AAR-IK | CEN | SSUM | SC |
| Algorithm Selection % | AAR | - | 0.58 | 0.69 | 0.67 | 0.70 | 0.80 |
| | AAR-G | 0.42 | - | 0.60 | 0.64 | 0.60 | 0.73 |
| | AAR-IK | 0.31 | 0.40 | - | 0.57 | 0.56 | 0.67 |
| | CEN | 0.33 | 0.36 | 0.43 | - | 0.51 | 0.68 |
| | SSUM | 0.30 | 0.40 | 0.44 | 0.49 | - | 0.65 |
| | SC | 0.20 | 0.27 | 0.33 | 0.32 | 0.35 | - |

Figure 5.15: Pair-wise selection matrix for user study results. This matrix shows the selection percentages between particular pairs of ad resizing techniques. The score in row $i$, column $j$ denotes the fraction of time algorithm $i$ was selected when shown against algorithm $j$.

improves resizing results.

The final table, shown in Figure 5.16, displays selection percentage results across each resizing size range for both column and row reduction. The results here are of a similar nature to the previous tables however they provide some additional insight to the resizing process. Specifically, although the AAR techniques perform well across all sizes the overall column reduction performance is better than the row reduction performance. From a high-level perspective this can be attributed to the benefits of text size searches when line wrapping is possible. This is more significant for column reduction because there is often extra vertical space to display headings on multiple lines. When text is simply resized proportionally to its original size this extra space cannot be used to display headings at larger sizes. This also explains why the best performance is achieved for $\gamma_M$, because within this range there is the most flexibility to adjust text parameters to utilize empty space. In the $\gamma_L$ range the text size is not substantially decreased by the simple methods and in the $\gamma_S$ range the image content has a much larger influence

| Algorithm | Resizing Factor | | | | | |
|---|---|---|---|---|---|---|
| | Row Reduction | | | Col. Reduction | | |
| | $\gamma_L$ | $\gamma_M$ | $\gamma_S$ | $\gamma_L$ | $\gamma_M$ | $\gamma_S$ |
| AAR | 0.60 | 0.63 | 0.76 | 0.66 | 0.80 | 0.73 |
| AAR-G | 0.52 | 0.63 | 0.67 | 0.53 | 0.63 | 0.66 |
| AAR-IK | 0.55 | 0.42 | 0.37 | 0.63 | 0.55 | 0.48 |
| CEN | 0.51 | 0.50 | 0.45 | 0.45 | 0.35 | 0.44 |
| SAL. SUM | 0.35 | 0.47 | 0.48 | 0.50 | 0.44 | 0.48 |
| SC | 0.46 | 0.32 | 0.27 | 0.24 | 0.24 | 0.22 |

Figure 5.16: User study selection percentages computed for each individual resizing range and direction where $\gamma_L \in [0.6, 0.75]$, $\gamma_M \in [0.4, 0.55]$, and $\gamma_S \in [0.2, 0.35]$.

on user preference than text placement. The result by size also reveal that the SP for $\gamma_L$ is lower than SP for $\gamma_S$ for AAR and AAR-G. This can be attributed to the subjectivity of selecting the 'better' poster at large sizes where all methods perform well, for $\gamma_S$ the results are strongly influenced by the quality of image importance and text relevance maps.

The above chapter provided a detailed analysis of the automatic image resizing framework and energy-based importance map algorithms. Through qualitative and quantitative analysis the performance of proposed techniques was shown to be superior to popular alternatives. The user study results demonstrate that the proposed ad resizing framework can substantially improve resizing performance as compared to approaches used in current AIR systems. The results also support the use of the energy-based importance map model over the IK model to compute saliency maps for automatic poster resizing.

# Chapter 6

# Conclusion

## 6.1 Summary

This work has demonstrated a complete system for automatic poster resizing which is applicable to a wide range of imagery. This includes a framework for computing resizing solutions through an objective function which captures important structural aspects of a ‘good’ poster. In addition to this, a novel image importance model has been developed which effectively makes use of text region information to improve the robustness of the resizing procedure.

Experimental analysis of resizing results and importance maps provide evidence that the techniques proposed here are superior to existing methods used within this domain. The analysis also suggested that the proposed importance maps will improve results in standard image resizing systems. Although the range of applicable resizing varies across individual posters, results indicate that moderate resizing ($\gamma > 0.6$) can be achieved with a very high success rate. More aggressive resizing ($\gamma \leq 0.4$) can also be successfully applied along at least one direction in a significant majority of cases. The additional text-region information and text placement constraints clearly help overcome high-level challenges associated with purely bottom-up image resizing tasks. In addition to qualitative observation these conclusion are also supported by a user study where the AAR technique was selected over 70% of the time compared to alternative approaches.

## 6.2 Future Work

The automatic ad resizing system has numerous components which present many opportunities for future enhancements. Staying within the proposed framework this section briefly outlines a few ideas for future research.

With respect to the resizing optimization, there is potential for improvement in the solution space construction. Incorporating a model for selecting a search space would likely improve overall results and add more generality to this process. Another area for future research involves the computational requirement of the poster resizing search which can pose challenges in large-scale applications. One approach for alleviating this could be the construction of a model which approximates the set of all solutions as a curve in parameter space.

The proposed algorithm for constructing importance maps provides another interesting area for future work. Each step of this process can likely be enhanced or tuned to improve performance. In particular, techniques to handle noise in the surround-sampling process or explicitly incorporate robustness to density estimate are both likely to improve the resulting importance maps.

Finally, the above work did not make use of any user feedback in the resizing process, however a supervised model has potential to dramatically improve results. The performance gains and added robustness would likely compensate for the additional work required for poster construction.

This work has investigated the problem of automatically resizing posters in the context of online-ad delivery systems such as Google AdSense. The results suggest that within appropriate constraints, significantly more advanced automatic ad generation methods can be developed than those currently used. In particular, hybrid approaches using both repositioning and re-sampling (as with Google Ad Builder) and automatic poster resizing methods have the potential to yield very powerful tools for visual ad delivery. This would vastly increase the applicability of these systems and go a long way in increasing the flexibility of image-based advertisements.

# Bibliography

[1] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3):10, 2007.

[2] Serge Belongie, Chad Carson, Hayit Greenspan, and Jitendra Malik. Color- and texture-based image segmentation using em and its application to content-based image retrieval. In *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, page 675, Washington, DC, USA, 1998. IEEE Computer Society.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[4] Chad Carson, Megan Thomas, Serge Belongie, Joseph Hellerstein, and Jitendra Malik. Blobworld: a system for region-based image indexing and retrieval. Technical report, Berkeley, Berkeley, CA, USA, 1999.

[5] D. Chai and K.N. Ngan. Face segmentation using skin-color map in videophone applications. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(4):551–564, Jun 1999.

[6] L.Q. Chen, X. Xie, X. abd Fan, W.Y. Ma, H.J. Zhang, and H.Q. Zhou. A visual attention model for adapting images on small displays. *ACM Multimedia Systems Journal*, 2003.

[7] Yung Y. Chuang, Brian Curless, David H. Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *Proceedings of IEEE CVPR 2001*, volume 2, pages 264–271. IEEE Computer Society, December 2001.

[8] G. Ciocca, C. Cusano, F. Gasparini, and R. Schettini. Self-adaptive image cropping for small displays. In *Consumer Electronics, 2007. ICCE 2007. Digest of Technical Papers. International Conference on*, pages 1–2, Las Vegas, NV,, January 2007.

[9] C. Connolly and T. Fleiss. A study of efficiency and accuracy in the transformation from rgb to cielab color space. *Image Processing, IEEE Transactions on*, 6(7):1046–1048, Jul 1997.

[10] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 233–240, New York, NY, USA, 2006. ACM.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[12] Sven Dickinson. *The Evolution of Object Categorization and the Challenge of Image Abstraction. In S. Dickinson, A. Leonardis, B. Schiele, and M. Tarr, eds., Object Categorization: Computer and Human Vision Perspectives.* Cambridge University Press, 2009.

[13] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, November 2000.

[14] Jeff Edmonds. *How to Think About Algorithms*. Cambridge University Press, New York, NY, USA, 2008.

[15] Xin Fan, Xing Xie, Wei-Ying Ma, Hong-Jiang Zhang, and He-Qin Zhou. Visual attention based image browsing on mobile devices. In *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, volume 1, pages 53–6, July 2003.

[16] Dashan Gao, Vijay Mahadevan, and Nuno Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. In *NIPS*. MIT Press, 2007.

[17] Dashan Gao and Nuno Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *NIPS*, 2004.

[18] Theo Gevers. Color feature detection: An overview. In R. Lukac and K.N. Plataniotis, editors, *Image Processing: Methods and Applications*. CRC Press, 2006.

[19] S. Gilles. *Robust description and matching of images*. PhD thesis, University of Oxford, 1998.

[20] H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C. H. Anderson. Overcomplete steerable pyramid filters and rotation invariance. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 222–228, 1994.

[21] S.E. Grigorescu, N. Petkov, and P. Kruizinga. Comparison of texture features based on gabor filters. *Image Processing, IEEE Transactions on*, 11(10):1160–1167, Oct 2002.

[22] Jonathon S. Hare, Paul H. Lewis, Peter G. B. Enser, and Christine J. Sandom. Mind the gap: Another look at the problem of the semantic gap in image retrieval. In Edward Y. Chang, Alan Hanjalic, and Nicu Sebe, editors, *Multimedia Content Analysis, Management and Retrieval 2006*, volume SPIE Vol. 6073, pages 607309–1. SPIE and IS&T, 2006.

[23] C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[24] Paul Harrison. A non-hierarchical procedure for re-synthesis of complex textures. In V. Skala, editor, *WSCG 2001 Conference Proceedings*, 2001.

[25] Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[26] L. Itti and C. Koch. Computational modelling of visual attention. *Nat Rev Neuroscience*, 2(3):194–203, March 2001.

[27] Laurent Itti and Christof Koch. A comparison of feature combination strategies for saliency-based visual attention systems. In *Proceedings of SPIE Human Vision and Electronic Imaging IV (HVEI'99)*, volume 3644, pages 473–482, San Jose, CA, January 1999. SPIE.

[28] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[29] Timor Kadir and Michael Brady. Saliency, scale and image description. *Int. J. Comput. Vision*, 45(2):83–105, 2001.

[30] R. Keys. Cubic convolution interpolation for digital image processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(6):1153–1160, Dec 1981.

[31] Donald E. Knuth and Michael F. Plass. Breaking paragraphs into lines. *Softw., Pract. Exper.*, 11(11):1119–1184, 1981.

[32] J. Lafferty, A. Mccallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling data. In *International Confernence on Machine Learning*, 2001.

[33] Feng Liu and Michael Gleicher. Automatic image retargeting with fisheye-view warping. In *UIST '05: Proceedings of the 18th annual ACM symposium on User interface software and technology*, pages 153–162, New York, NY, USA, 2005. ACM.

[34] Hao Liu, Xing Xie, Wei-Ying Ma, and Hong-Jiang Zhang. Automatic browsing of large pictures on mobile devices. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 148–155, New York, NY, USA, 2003. ACM.

[35] David G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV, Corfu*, pages 1150–1157, 1999.

[36] Yu-Fei Ma and Hong-Jiang Zhang. Contrast-based image attention analysis by using fuzzy growing. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 374–381, New York, NY, USA, 2003. ACM.

[37] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. Performance measures for information extraction. In *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[38] A. V. Oppenheim and A. S. Willsky. *Signals and systems.* Prentice Hall, New York, NY, second edition, 2006.

[39] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *Discrete-time signal processing (2nd ed.).* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1999.

[40] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* Mc-Graw Hill, 2002.

[41] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107 – 123, 2002.

[42] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Detection of interest points using symmetry. In *Computer Vision, 1990. Proceedings, Third International Conference on*, pages 62–65, Osaka, Japan, December 1990.

[43] Tongwei Ren, Yanwen Guo, Gangshan Wu, and Fuyan Zhang. Constrained sampling for image retargeting. *Multimedia and Expo, 2008 IEEE International Conference on*, pages 1397–1400, 23 2008-April 26 2008.

[44] Michael Rubinstein, Ariel Shamir, and Shai Avidan. Improved seam carving for video retargeting. *ACM Transactions on Graphics, (Proceedings SIGGRAPH 2008)*, 27(3), 2008.

[45] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 771–780, New York, NY, USA, 2006. ACM.

[46] Raimondo Schettini, Carla Brambilla, Claudio Cusano, and Gianluigi Ciocca. Automatic classification of digital photographs based on decision forests. *IJPRAI*, 18(5):819–845, 2004.

[47] Cordelia Schmid, Roger Mohr, and Christian Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[48] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces andcars. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 746–751, Hilton Head Island, SC, USA, 2000.

[49] Vidya Setlur, Saeko Takagi, Ramesh Raskar, Michael Gleicher, and Bruce Gooch. Automatic image retargeting. In *MUM '05: Proceedings of the 4th international*

*conference on Mobile and ubiquitous multimedia*, pages 59–68, New York, NY, USA, 2005. ACM.

[50] Ariel Shamir and Shai Avidan. Seam carving for media retargeting. *Communications of the ACM*, 52(1):77–85, 2009.

[51] Bongwon Suh, Haibin Ling, Benjamin B. Bederson, and David W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *UIST '03: Proceedings of the 16th annual ACM symposium on User interface software and technology*, pages 95–104, New York, NY, USA, 2003. ACM.

[52] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

[53] Emanuele Trucco and Alessandro Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1998.

[54] John K. Tsotsos and Neil D. B. Bruce. Saliency based on information maximization. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 155–162, MIT Press, 2006. MIT Press.

[55] Paul Viola and Michael Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[56] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395 – 1407, 2006. Brain and Attention, Brain and Attention.

[57] J. Van De Weijer, Th. Gevers, A.D. Bagdanov, I. A Edge, Feature Detection, and I. D Statistical. Boosting color saliency in image feature detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28:150–156, 2005.

[58] Xing Xie, Xin Fan, Wei-Ying Ma, and He-Qin Zhou. Adapting images on proxies for small form factor devices. *Proceedings of the International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia.*, 1:428–432 Vol.1, Dec. 2003.