Thesis for the degree of Licentiate of Engineering

Enhancement of Salient Image Regions for Visual Object Detection

Keren Fu

Signal Processing Group
Department of Signals and Systems
Chalmers University of Technology

Göteborg 2014

Fu, Keren
Enhancement of Salient Image Regions for Visual Object Detection.

Department of Signals and Systems
Technical Report No. R015/2014
ISSN 1403-266X

Signal Processing Group
Department of Signals and Systems
Chalmers University of Technology
SE-412 96 Göteborg, Sweden
Telephone: + 46 (0)73-5510941

This thesis has been prepared using LaTeX.

Printed by Chalmers Reproservice,
Göteborg, Sweden, November 2014.

To my parents

Abstract

Salient object/region detection aims at finding interesting regions in images and videos, since such regions contain important information and easily attract human attention. The detected regions can be further used for more complicated computer vision applications such as object detection and recognition, image compression, content-based image editing, and image retrieval. One of the fundamental challenge in salient object detection is to uniformly emphasize desired objects and meanwhile suppress irrelevant background. Existing heuristic color contrast-based methods tend to obtain false detection in complex scenarios and attenuate the inner part of large salient objects. In order to achieve uniform object enhancement and background suppression, several new techniques including color feature integration, graph-based geodesic saliency propagation, hierarchical segmentation based on graph spectrum decomposition are developed in this thesis to assist saliency computation. Paper 1 proposes a superpixel-based salient object detection method which takes advantages of color contrast and distribution. It develops complementary abilities among hypotheses and generates high quality saliency maps. Paper 2 proposes a novel geodesic propagation method for salient region enhancement. It leverages an initial coarse saliency map that highlight potential salient regions, and then conducts geodesic propagation. Local connectivity of objects is retained after the proposed propagation. Papers 3 and 4 use graph-based spectral decomposition for hierarchical segmentation, which enhances saliency detection. As most previous work on salient region detection is done for still images, paper 5 extends graph-based saliency detection methods to video processing. It combines static appearance and motion cues to construct graph. A spatial-temporal smoothing operation is proposed on a structured graph derived from consecutive frames to maintain visual coherence in both inter- and intra- frames. All these proposed methods are validated on benchmark datasets and achieve comparable/better performance to the state-of-the-art methods.

Keywords: salient region, visual attention, color contrast and distribution, geodesic distance, propagation, graph spectral decomposition, figure-ground segmentation, video processing.

List of Publication

This thesis is based on the following appended publications

Paper 1
Keren Fu, Chen Gong, Jie Yang, Yue Zhou, Irene Yu-Hua Gu. “Superpixel based Color Contrast and Color Distribution Driven Salient Object Detection”. *Signal processing: Image Communication*, 28(10): 1448-1463, 2013.

(Part of this paper was presented in)
Keren Fu, Chen Gong, Jie Yang, Yue Zhou. “Salient Object Detection via Color Contrast and Color Distribution”. In 11th *Asian Conference on Computer Vision (ACCV)*, Daejeon, South Korea, November 5-9, 2012.

Paper 2
Keren Fu, Chen Gong, Irene Yu-Hua Gu, Jie Yang. “Geodesic Saliency Propagation for Image Salient Region Detection”. In 20th IEEE *International Conference on Image Processing (ICIP)*, 2013.

Paper 3
Keren Fu, Chen Gong, Irene Yu-Hua Gu, Jie Yang, Xiangjian He. “Spectral Salient Object Detection”. In IEEE *International Conference on Multimedia & Expo (ICME)*, 2014.

Paper 4
Keren Fu, Chen Gong, Yixiao Yun, Yijun Li, Irene Yu-Hua Gu, Jie Yang, Jingyi Yu. “Adaptive Multi-Level Region Merging for Salient Object Detection”. In *British Machine Vision Conference (BMVC)*, 2014.

Paper 5
Keren Fu, Irene Yu-Hua Gu, Yixiao Yun, Chen Gong, Jie Yang. “Graph Construction for Salient Object Detection in Videos”. In 22nd *International Conference on Pattern Recognition (ICPR)*, 2014.

Other publications by the author, omitted in the thesis

Keren Fu, Chen Gong, Yu Qiao, Jie Yang, Irene Yu-Hua Gu. "One-Class Support Vector Machine-Assisted Robust Tracking". *Journal of Electronic Imaging*, 22(2), 023002, 2013.

Keren Fu, Kai Xie, Chen Gong, Irene Yu-Hua Gu, Jie Yang. "Effective Small Dim Target Detection by Local Connectedness Constraint". In IEEE *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

Yixiao Yun, **Keren Fu**, Irene Yu-Hua Gu, Jie Yang. "Visual Object Tracking with Online Learning on Riemannian Manifolds by One-Class Support Vector Machines". In 21th IEEE *International Conference on Image Processing (ICIP)*, 2014.

Chen Gong, **Keren Fu**, Enmei Tu, Jie Yang, Xiangjian He. "Robust Object Tracking using Linear Neighborhood Propagation". *Journal of Electronic Imaging*, 22(1), 013015, 2013.

Chen Gong, **Keren Fu**, Artur Loza, Qiang Wu, Jia Liu, Jie Yang. "PageRank Tracker: From Ranking to Tracking". *IEEE Trans on Systems, Man, and Cybernetics, Part B*, 44(6): 882-893, 2014.

Lei Zhou, **Keren Fu**, Yijun Li, Yu Qiao, Xiangjian He, Jie Yang. "Bayesian Salient Object Detection based on Saliency Driven Clustering". *Signal processing: Image Communication*, 29(3): 434-447, 2014.

Contents

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisors for their time and patience. I wish to thank Prof. Irene Gu at Chalmers and Prof. Jie Yang at Shanghai Jiao Tong University in China for their support and constructive instruction. Special thanks to Prof. Irene Gu, who has help me a lot with my project and research articles during my stay in Sweden. Her enthusiasm and well-knit style on research has impressed me a lot and taught me how to do research scientifically.

I would like to thank my colleague Yixiao Yun for valuable discussion and suggestions. Last but not least, I would like to thank the current and former members of the signal processing group at Chalmers.

Keren Fu
Göteborg, November 2014

Acronyms

PCA:	Principle Component Analysis
EM:	Expectation Maximization
ML:	Maximized Likelihood
GMM:	Gaussian Mixture Models
CRF:	Conditional Random Field
MRF:	Markov Random Field
MSRA:	Microsoft Research Asia
Ncut:	Normalized cut
ROI:	Region of Interest
MHOF:	Mean Histogram of Optical Flow

Part I

Introductory chapters

Chapter 1

Introduction

1.1 Background

Images appearing on websites, mobile devices, as well as TVs and computer screens enrich our daily life. However, processing such large amount of visual information in images in short time is a difficult task. Information in images differs in importance. Some are crucial while others are negligible. An automatic and selective mechanism that answers which information is necessary to pick up from an image for further analysis can be useful. A feasible way is by the selective mechanism of human visual attention: according to studies of neurobiology and cognitive psychology [1, 2], human brains are capable of selecting a certain subset of visual information for further processing. Modeling human visual attention on images is referred to as *saliency detection*, which aims at detecting salient image parts that can easily attract human attention. Although attention processes of human rely on bottom-up influences and top-down influences [3], saliency detection focused in this thesis only considers bottom-up factors (usually the influences from low-level features). This type of saliency detection is stimulus-driven as well as widely studied in the past decade. Saliency detection results indicating potential regions of interest (ROI) provide some guidance to further analysis. This has been used in many applications, e.g. object detection and recognition [4, 5], image compression [6], video summarization [7], content-based image editing [8–11] and image retrieval [12]. In the last decade, saliency detection has become a research field in computer vision attracting much attention.

Saliency detection methods can be categorized into either *eye fixation modeling* or *salient region detection*. Most early saliency models belong to the former, aiming at predicting where human look in the scene. Their basis dates back to the “Feature Integration Theory” [1] stating what kinds of visual features are important and how they are combined to direct human attention. Koch and Ullman [2] first proposed a feed-forward model to combine these features and introduced the concept of a saliency map, i.e. a topographic map that represents conspicuousness of scene locations [3]. A winner-take-all neural network was introduced in [2] to select the most salient locations and employs an inhibition of return mechanism to simulate eye shift. The first complete implementation of [2] was proposed by Itti et al [13]. As one of the earliest pioneer work, Itti et al [13] proposed a "center-surround" operation as local feature contrast in the color, intensity, and orientation on a pyramid of an image. Such center-surround operation characterizes the stimuli to visual neural cells and is realized using DOG (Difference of Gaussians). Although eye fixation prediction is the origin of saliency detection and has gained a lot of progress since then, these methods have a typical drawback which impacts their performance in real applications. This drawback is that they tend to generate selectively sparse saliency maps. When using them to detect big salient objects, such models highlight only edges, corners of objects and attenuate their inner parts, due to their favors of high frequency components in the image.

To benefit complex computer vision tasks, such as content-based image editing which prefer enhancement of entire objects, a new sub-field called *salient region detection* has emerged in the light of compensating the drawback of previous eye fixation prediction models. The goal of salient region detection is to detect and segment salient objects in natural scenes. Compared to fixation prediction, salient region detection computes global contrast, or center-surround contrast on image regions to prevent enhancement of only image edges. Hierarchical image segmentation is also considered as a multi-scale operation (rather than commonly used image pyramid in eye fixation prediction) to generate edge-aware saliency maps.

This thesis addresses salient region/object detection in images. Despite many methods on salient object detection have been proposed lately, improving the performance in complex scenarios remains challenging. One of the fundamental challenges in salient object detection is to emphasize desired objects uniformly and suppress irrelevant background. Existing heuristic color contrast-based methods tend to obtain false detection in complex scenarios and attenuate the inner part of large salient objects. In this thesis, several novel techniques are introduced to remedy these problems and improve the performance.

1.2 Overview of Previous Salient Region Detection Methods

This section reviews several state-of-the-art methods on salient object detection.

- **Heuristic Color Contrast-based Methods**

Methods of this category attempt to model saliency using local or global color statistics. It is based on the assumption that salient objects are unique in color and present high color contrast to the rest parts of an image. Zhai et al [14] introduce image histograms which only model luminance channel to calculate pixel-level saliency. Pixel-level spatial saliency is measured as the luminance contrast between image pixels. Achanta et al [15] provide a saliency approximation by subtracting the average color from low-pass filtered result of an image. This operation of [15] is equivalent to combining center-surround differences of all bandwidth to detect objects of different sizes. Goferman et al [11] combine local and global features to estimate the patch saliency in multi-scales. To consider both local and global factors, they compute saliency of a certain patch as its contrast to the K nearest patches in the image. Under this framework, inner parts of an object are often attenuated due to the edge preference. Cheng et al [16] extend the method in [14] and incorporate color histograms. A regional contrast saliency measure is also proposed in [16] as the contrast to other regions. Perazzi et al [17] propose saliency filter, which formulates complete contrast and saliency estimation using high dimensional Gaussian filters. Margolin et al [18] define patch distinctness as L1-norm in PCA coordinates and combine patch distinctness with regional color distinctness. Shi et al [19] compute pixel-wise image saliency by aggregating complementary appearance contrast measures with spatial priors. Most of the above contrast-based methods are straightforward and simple to compute. Their performances are less satisfactory on images with complex background.

- **Learning-based Methods**

This category of methods estimates object saliency through machine learning. In this case, training samples are needed. The fundamental is to learn the weight of features during saliency computation. Jiang et al [20] perform pre-segmentation for an input image and extract a bunch of discriminative features from each segmented region. Then a random forest regressor is adopted to map multiple features to a region saliency score. Liu et al [21] segment salient objects by aggregating pixel saliency cues in a Conditional Random Field. Their saliency cues include center-surround histogram contrast, saliency maps from the spectral residual method [22], and color spatial distribution. The linear weight for those cues are learned under the Maximized Likelihood (ML) criteria by tree-reweighted belief propagation. Learning-based methods can achieve good performance in complex scenarios attributed to the learning. However, high computation is needed for this type of methods due to feature extraction and learning, as comparing to the color contrast-based methods.

- **Hierarchical/Multi-scale Segmentation for Saliency Detection**

The basics of this kind of methods is to generate good segmentation, usually in hierarchy or multi-scale to facilitate saliency computation. Lu et al [23] exploit the concavity context in a scene. They observed salient objects are surrounded by concave edges, so their method boils down to detecting concave arcs from multi-scale segmentation. The detected arcs then contribute to a figure-ground segregation phase. Finally, one of the two segregated regions is selected as salient according to their surroundness. Yan et al [24] propose a hierarchical saliency detection method that merges regions according to user-defined scales (e.g., 3 size scales in their case) to eliminate small-size distracters. In a certain hierarchy, a region would be absorbed by its neighbor region if it is smaller than the defined size. Each region is evaluated using local contrast and location prior. Cheng et al [25] measure saliency by hierarchical soft abstraction. They form a 4-layer hierarchical structure (respectively are pixel layer, histogram layer, GMM layer and clustering layer) with an index table to associate cross-layer relations efficiently. Saliency estimation using color contrast and distribution are conducted on the coarse layers and then propagated to the pixel layer. Jiang et al [26] find potential salient regions by maximizing a submodular objective function. The problem is solved efficiently by finding a closed-form harmonic solution on the constructed graph for an input image. The saliency of a selected region is modeled in terms of appearance and spatial location. In summary, these methods, benefiting from some optimized segmentation phase, can easily make object emphasized uniformly and boost final performance.

- **Graph-based Methods**

The basics of these methods is to represent images using graphs, where natures of salient objects, like high color contrast, compact color distribution are modeled. Gopalakrishnan et al [27] perform random walks on graphs to find salient object. The global pop-out and compactness properties of salient objects are modeled in random walks by the equilibrium access time performed on a complete and k-regular graph. Wei et al [28] propose to treat boundary parts of an image as the background. The patch saliency is defined as the shortest geodesic distance for a graph to image boundaries. As a salient object is often isolated from the background, the geodesic distance between image boundaries and object parts is relatively large, leading to an object being popped out. Yang et al [29] utilize similar boundary priors as [28] but propagate saliency via graph-based manifold ranking from four image borders separately. Four saliency maps generated are then multiplied to achieve the final one. Their method is shown better than state-of-the-art methods (including [29]) for salient object detection. Both [28] and

[29] measure the connectivity of image parts to image boundaries. Salient objects are detected as regions disconnected from the image boundaries. In [30], saliency detection is formulated by using absorbing Markov chain on an image graph model. The virtual boundary nodes are chosen as the absorbing nodes in a Markov chain. The absorbed time is used as a metric for measuring saliency. Generally, graph-based methods may achieve high performance among state-of-the-art salient region detection methods.

- **Other Methods**

Other notable work includes: Shen et al [31] solve saliency detection issue as a low rank matrix recovery problem, where salient objects are represented by a sparse matrix (noise) while background are indicated by a low rank matrix. However, this sparse and low rank assumption may not be satisfied in complex scenes, leading to unsatisfactory results. A Bayesian framework is adopted in [32]. First, saliency points are applied to get a coarse location of the saliency region. Based on the rough region, a prior map is computed for the Bayesian model to achieve the final saliency map. Mai et al [33] propose a data-driven approach for aggregating saliency maps generated by other saliency detection methods using Conditional Random Field (CRF). The weight for aggregation is learned in a data-driven way from k nearest neighbors of the input image retrieved using the Gist feature from a predefined training dataset.

1.3 Addressed Problems in this Thesis

This thesis focuses on salient region/object detection in natural images. The aim is to generate high quality saliency maps that enhance holistic salient objects uniformly meanwhile suppress irrelevant background. Several new methods including color feature integration, graph-based geodesic saliency propagation, hierarchical segmentation through graph spectral decomposition are proposed for saliency detection. Salient object detection in videos is also addressed in this thesis, which is under-explored in its field compared to salient object detection in still images.

1.4 Motivations

Recall that our goal is to enhance salient objects uniformly meanwhile suppress irrelevant background. To achieve this goal, our main motivations for saliency detection include:

- Use multiple hypotheses

There exists several hypotheses for detecting salient objects. However, individual hypothesis can not work well in all cases. To make the performance robust, multi-hypothesis fusion is deserved. It may help pop up salient objects in challenging background. Discovering complementary performances between different hypotheses is also useful.

- Generate good segmentation

  When performing region-level saliency estimation, detection results may highly depend on segmentation. Imagine each salient object is segmented into a single region. It could be analyzed in a holistic way and more accurate object saliency could be achieved. Although this assumption is quite ideal, attempts of generating good segmentation still worth a try.

- Exploit graph representation

  Graph representation, due to its good modeling ability and mature theories, has been used in many fields. Hence, it would be interesting to exploit and use graph representation for salient object detection. In this sense, many graph theories could be considered.

- Apply salient region detection to videos

  While salient object detection in still images has gained much attention, its application to videos remains under-explored.

1.5 Outline of this Thesis

The thesis is divided into two parts. The first part briefly describes the background and the proposed work. The second part includes publications resulted from this thesis work. The first part of the thesis is organized as follows: Chapter 2 gives an overview of related state-of-the-arts. Chapter 3 reviews the hypotheses and theories that are highly related to the proposed work. Chapter 4 summarizes the proposed methods, followed by Chapter 5 on the conclusion.

Chapter 2

Review of Related Work

This chapter briefly reviews the methods on saliency detection that are closely related to this thesis, together with fundamentals that are employed to address the detection problem. Section 2.1 and 2.2 introduce two previous ideas on color contrast [16] and color distribution [21], upon which we build our methods. Section 2.3 introduces basics related to graphs, and consists of 3 subsections that respectively involve geodesic distance [34], normalized graph cut [35, 36], and conditional random field [37].

2.1 Global Regional Contrast

[16] is one of the earliest literature which proposes basics on saliency computation of image regions. An input image is segmented into regions aligned with intensity edges first and then a regional saliency map is computed. Since saliency can be defined as uniqueness, which may be characterized by high feature contrast to the rest parts of the image, saliency is measured by the global contrast between the target region with respect to all other regions in the image. Suppose an input image is pre-segmented into N regions $\{r_i\}_{i=1}^{N}$. The global regional contrast saliency is computed as:

$$S(r_i) = \sum_{j=1}^{N} w_{ij} D_r(r_i, r_j) \tag{2.1}$$

$D_r(r_i, r_j)$ is related to the appearance contrast between two regions, e.g. χ^2 distance between color histograms of r_i and r_j. w_{ij} is the weight defined as:

$$w_{ij} = \exp(-D_s(r_i, r_j)/\sigma_s^2)|r_j| \tag{2.2}$$

where $D_s(r_i, r_j)$ is the spatial distance between region r_i and r_j, and σ_s controls the strength of spatial weighting. Large values of σ_s reduce the effect of spatial weighting, so that contrast to farther regions would contribute more to the saliency of the current region. By letting $\sigma_s \to \infty$, equal weight for all regions is resulted. $|r_j|$ is the size of region r_j. Noting that since saliency values are measurements within an image showing relative importance, finally they are linearly normalized to the interval [0,1] to obtain a saliency map. The above idea on regional color contrast motivates the proposed superpixel-based color contrast calculation in this thesis work.

2.2 Spatial Distribution of Colors

Besides characterizing the uniqueness using contrast, color spatial distribution is another saliency indicator. The more widely a color is distributed in an image, the less possible it belongs to a salient object. This is because objects are usually compact regions surrounded by the background. The extent of how widely a color distributes could be measured by color spatial variance. [21] proposes to compute color spatial variances using Gaussian Mixture Models (GMMs). First, all colors in the image are represented by Gaussian Mixture Models (GMMs) $\{w_c, \mu_c, \sum_c\}_{c=1}^{C}$ by using EM (Expectation Maximization) algorithm, where $\{w_c, \mu_c, \sum_c\}$ is the weight, the mean color, and the covariance matrix of the cth component. Each pixel is assigned to a color component with the probability:

$$p(c|I_x) = \frac{w_c \mathcal{N}(I_x|\mu_c, \sum_c)}{\sum_c w_c \mathcal{N}(I_x|\mu_c, \sum_c)} \tag{2.3}$$

Suppose x_h is the x-coordinate (horizontal coordinate) of the pixel x. The spatial variance for x-dimension of color component c is computed as:

$$\sigma_h^2(c) = \frac{1}{|P|_c} \sum_x p(c|I_x) \cdot |x_h - M_h(c)|^2 \tag{2.4}$$

where $M_h(c) = \frac{1}{|P|_c} \sum_x p(c|I_x) \cdot x_h$, and $|P|_c = \sum_x p(c|I_x)$ is a normalization factor. The vertical variance $\sigma_v^2(c)$ is defined similarly. The spatial variance of a component c is combined as: $\sigma^2(c) = \sigma_h^2(c) + \sigma_v^2(c)$. $\sigma^2(c)$ is normalized by:

$$\sigma^2(c) \leftarrow \frac{\sigma^2(c) - \min_c \sigma^2(c)}{\max_c \sigma^2(c) - \min_c \sigma^2(c)} \tag{2.5}$$

Finally, the saliency $S(I_x)$ of a specific pixel I_x regarding to color spatial distribution is defined as the weighted sum:

$$S(I_x) = \sum_c p(c|I_x) \cdot (1 - \sigma^2(c)) \tag{2.6}$$

Equation (2.6) implies that pixels with small color distribution variances have high saliency values. The above idea motivates the use of superpixel-based color distribution in this thesis work.

2.3 Graph Theories for Salient Region Detection

2.3.1 Geodesic Distance on 2D images

Geodesic distance is defined as the shortest path between any pair of points on a nonlinear surface/space. Here geodesic distance and its definition on 2D images [34] is reviewed since it is more relevant to the thesis work. For 2D images, geodesic distance can be used to describe connectivity of two image locations. It has been shown useful for segmentation, edge-preserving filtering, denoising, stitching, and colorization [34]. Let $I(x)$ be an image: $\Psi \rightarrow \mathbb{R}^d$ ($d = 3$ for a color image), whose support $\Psi \in \mathbb{R}^2$ is assumed to be continuous. Given two points $a, b \in \mathbb{R}^2$, the geodesic distance between them is defined as:

$$d_{geo}(a,b) = \inf_{\Gamma \in \mathcal{P}_{a,b}} \int_0^{l(\Gamma)} \sqrt{1 + \gamma^2 (\nabla I(s) \cdot \Gamma'(s))^2} ds \tag{2.7}$$

where $\mathcal{P}_{a,b}$ is the set of all possible differentiable paths between a and b. The spatial derivative $\Gamma'(s) = \partial\Gamma(s)/\partial(s)$ is the unit vector tangent to the direction of the path with arc length s, and $\nabla I(s)$ is the corresponding gradient vector. The dot-product $\nabla I(s) \cdot \Gamma'(s)$ measures the image gradient along the tangent. The geodesic factor γ weighs the contribution between the gradient and the spatial distance. When $\gamma = 0$, $d_{geo}(a,b)$ degenerates to Euclidean distance.

Given a binary mask $M(x) \in \{0,1\}$ associated to a "seed" region Ω : $\Omega = \{x, M(x) = 1\}$. The geodesic distance transform D assigned to each pixel at x is its minimum geodesic distance from Ω:

$$D(x; M, \nabla I) = \min_{\{x' | M(x')=1\}} d_{geo}(x, x') \tag{2.8}$$

In (2.7), 2D position s is assumed to be continuous. From (2.7), the discrete approximation over image lattice can be derived:

$$d_{geo}(a,b) = \min_{\Gamma \in \mathcal{P}_{a,b}} \sum_{p_k \in \Gamma} \sqrt{1 + \gamma^2 (\nabla I(p_k, p_{k+1}))^2} D_s(p_k, p_{k+1}) \tag{2.9}$$

where p_k is the kth point on the discrete path Γ. $\nabla I(p_k, p_{k+1})$ is the gradient magnitude between p_k, p_{k+1}. $D_s(p_k, p_{k+1})$ is the spatial distance between

p_k, p_{k+1}. When considering only connectivity and ignoring spatial distance, the following variation can be obtained:

$$d_{geo}(a,b) = \min_{\Gamma \in \mathcal{P}_{a,b}} \sum_{p_k \in \Gamma} \nabla I(p_k, p_{k+1}) D_s(p_k, p_{k+1}) \tag{2.10}$$

The geodesic distance in (2.10) can be computed over a graph, where each point p_k of the lattice is a vertex and $\nabla I(p_k, p_{k+1}) D_s(p_k, p_{k+1})$ between two adjacent lattice points is treated as an edge. Calculating (2.10) is equivalent to finding the shortest path over graph and Floyd's/Dijkstra's algorithms can be applied. Since objects usually present connectivity properties, i.e. connected regions are likely to be perceived as one object, in this thesis, superpixel-based geodesic distance is used and a geodesic propagation method for salient region enhancement is proposed.

2.3.2 Normalized Graph Cut (Ncut)

Graph cut is a method to find a partition of a graph such that edges between different groups have very low weight and edges within a group have high weight. This is similar to the aim of clustering: points within a same cluster are similar to each other while points in different clusters are dissimilar from each other. Since the Ncut and spectral clustering are tightly related, below the Ncut is first reviewed and its relation to spectral clustering is further described.

From the partition purpose, a simplest example to start with is the mincut example. Given a similarity graph (a graph whose edges measure similarity between vertices), let $\mathbf{W}$ be its adjacency matrix, $\mathbf{D}$ be its degree matrix (a diagonal matrix with diagonal entry $d_i = \sum_j w_{ij}$), and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ be its Laplacian matrix. The cut cost is defined: $cut(A,B) := \sum_{i \in A, j \in B} w_{ij}$, and let $\bar{A}$ be the complement of A. For a given number k of subsets, the mincut approach chooses a partition $A_1, ..., A_k$ which minimizes:

$$cut(A_1, ..., A_k) = \sum_{i=1}^{k} cut(A_i, \bar{A}_i) \tag{2.11}$$

In particular for $k = 2$, mincut is a relatively easy problem and can be solved efficiently. However, in practice the mincut often does not lead to satisfactory partitions, because the solution of mincut often separates one individual vertex from the rest of the graph. Obviously this is not satisfying to achieve in clustering because clusters should be reasonably *large* groups of points. A common objective function to encode this is by normalizing the cut values using cluster sizes, leading to the normalized cut (Ncut) that minimizes:

$$Ncut(A_1, ..., A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \bar{A}_i)}{assoc(A_i, V)} \tag{2.12}$$

where $assoc(A_i, V) := \sum_{i\in A, j \in V} w_{ij}$ is a measure of set size, i.e. the larger $|A_i|$ is, the higher $assoc(A_i, V)$ will be. The normalized cut is utilized in this thesis to assist saliency computation. By defining a hard indicating vector for each A_i and relaxing the hard constraints (for details please refer to [36]), the continuous indicating vectors for multi-cluster Ncut can be derived from the first k eigenvectors of $\mathbf{D}^{-1}\mathbf{L}$, or the first k generalized eigenvectors of:

$$(\mathbf{D} - \mathbf{W})\mathbf{y} = \lambda \mathbf{D}\mathbf{y}. \tag{2.13}$$

The solution of 2-way Ncut ($k = 2$) is given by its second smallest eigenvector.

Since the continuous indicating vectors for multi-cluster Ncut are derived from the first k generalized eigenvectors of system $(\mathbf{D} - \mathbf{W})\mathbf{y} = \lambda\mathbf{D}\mathbf{y}$, k-means clustering can be applied to these eigenvectors to obtain labels corresponding to clusters, leading to the spectral clustering (Algorithm 1). In this thesis, we propose to use Ncut/spectral decomposition to partition an image into different visual concepts (i.e. clusters). Details can be found in Papers 3 and 4.

Algorithm 1 Spectral Clustering

```
Require: Constructed similarity graph described by $\mathbf{W}$, number $k$ of clusters;
Ensure: Clusters $A_1, ..., A_k$ with $A_i = \{j|y_j \in C_i\}$;
Compute the degree matrix $\mathbf{D}$ and graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$;
Compute the first $k$ generalized eigenvectors $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k$ of the generalized eigenproblem $\mathbf{Lu} = \lambda\mathbf{Du}$.
Let $\mathbf{U} \in R^{n\times k}$ be the matrix containing the vectors $\mathbf{u}_1, ..., \mathbf{u}_k$ as columns.
For $i = 1, 2, ..., n$, let $(\mathbf{y}_i \in R^k)$ be the vector corresponding to the i-th row of $\mathbf{U}$.
Cluster the points $(\mathbf{y}_i)_{i=1,...,n}$ in $R^k$ with the k-means algorithm into clusters $C_1, ..., C_k$ and output clusters $A_1, ..., A_k$ with $A_i = \{j|y_j \in C_i\}$.
```

2.3.3 Conditional Random Field (CRF)

Conditional random fields offer advantages over Markov random fields on relaxing strong dependencies on the observation sequence. It has been proposed to segment and label data sequences [37]. Below its application to image segmentation is described, since it is relevant to the thesis work. The definition for CRF is given according to [37]: suppose X is a set of random variables over data sequences to be labeled, and Y is a set of random variables over corresponding label sequences. $G = (V, E)$ define a graph constructed from the data sequences such that $Y = (Y_v)_{v \in V}$ is indexed by the vertices of G. (X, Y) formulates a conditional random field in case, when conditioned on X, each Y_v obeys the Markov property with respect to the graph: $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means w and v are neighbors in G. This equation infers that probability of Y_v is conditioned on both data sequence X and labels of neighboring vertices Y_w.

In the case of foreground-background segmentation, suppose X is an observation image, and Y corresponds to pixel labels to be estimated. When giving a prior map P to constrain the result of Y, the conditional probability of Y can be written as:

$$p(Y|P, X) = p(Y_1, Y_2, ..., Y_n|P, X) = \frac{1}{Z} \exp(-E(Y|P, X)). \tag{2.14}$$

where Z is the partition function used for normalization. The energy function $E(Y|P, X)$ can be defined as a data term (or called unary term) and a pair-wise smoothness term:

$$E(Y|P, X) = \underbrace{\sum_v |Y_v - P_v|^m}_{\text{data term}} + \lambda_s \underbrace{\sum_{v,w|w \sim v} |Y_v - Y_w|^m A_{vw}}_{\text{pair-wise term}}. \tag{2.15}$$

where λ_s is the weight for smoothing. v and w indicate neighbor pixels in X. Maximize (2.14) equals to minimize the energy function (2.15). A_{vw} captures the affinity between Y_v, Y_w. A common requirement is that adjacent pixels with the same color tend to have the same label. Hence the A_{vw} can be constructed according to color similarity of image pixels, e.g. $A_{vw} = \exp\{-\beta||X_v - X_w||_2\}$ [21]. In this thesis, we use CRF to enhance object saliency as well as conduct spatial-temporal smoothing in video saliency detection. We choose $m = 2$ as it has a relaxed form (soft $Y_v \in [0, 1]$) with a closed-form solution for efficient computation. Noting that soft $Y_v \in [0, 1]$ is intrinsically applicable to saliency maps.

Chapter 3

Summary of Thesis Work

This chapter gives a summary of this thesis work on salient object detection. Papers 1-4 aim at detecting and enhancing salient regions in still images. Paper 5 aims at salient region detection for videos.

3.1 Paper 1: Saliency Detection by using Color Contrast and Distribution

Basic Ideas: Previous work [16] only considers color contrast for salient region detection, and the performance is rather limited. This proposed method aims at enhancing the detection performance by incorporating color contrast with color distribution (motivated by [21]). The basic idea is that color contrast and color distribution are computed independently and integrated through a small number of spatially compact superpixels. This enables efficient and effective computation of these two kinds of features. A superpixel is rendered with high saliency if it satisfies the following observations:

- has strong color contrast to its surroundings. (*contrast*)
- is located near the image center as theme objects tend to be placed near the image center by photographers. (*distribution*)
- has compact color spatial distribution. (*distribution*)

Assuming each salient object satisfies the above three conditions simultaneously, resulting measurements are combined by multiplication to obtain the final saliency.

Main contributions: Color contrast and color distribution are computed according to three observations on superpixels, and combined to achieve complementary performance. Contrast-based saliency of a superpixel is computed by its color contrast to all superpixels in an image. The resulting contrast-based saliency is then weighted by a 2D distribution prior with center bias, and globally smoothed so that superpixels with similar colors yield similar saliency. The motivation is to enhance objects uniformly and resolve ambiguity of saliency maps. Distribution-based saliency of a superpixel is inversely proportional to its color spatial distribution variance. The variance of a superpixel R_i (i is the index) is determined by the variance of spatial position of superpixels that share similar colors with R_i in an image. This is different from GMM-based computation in [21].

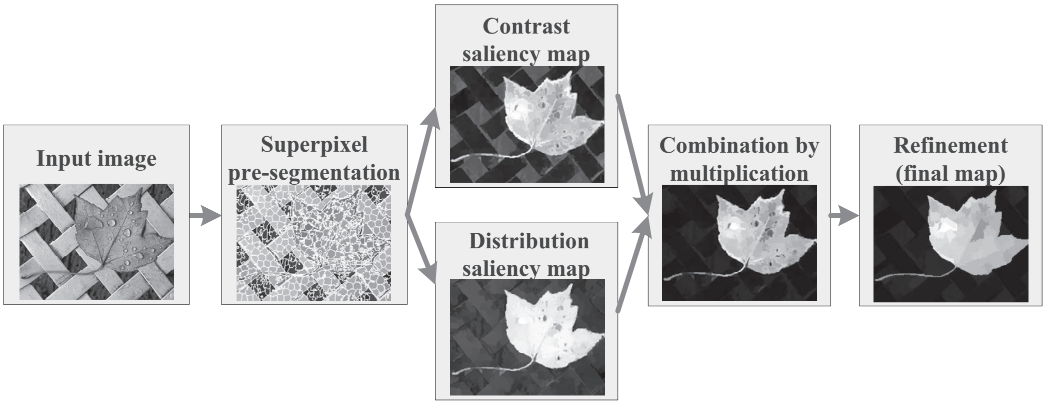
Input image
Superpixel pre-segmentation
Contrast saliency map
Distribution saliency map
Combination by multiplication
Refinement (final map)

Figure 3.1: The block diagram of the proposed method.

Big picture: The proposed method is shown in the block diagram of Figure 3.1. After superpixel pre-segmentation, color contrast and color distribution maps are computed separately on superpixels, and are then combined in superpixel-wise multiplication. Refinement is implemented by coarse segmentation, and the average saliency in each segmented region is computed to obtain the final saliency map.

Results: Tests and comparisons are performed on a public dataset MSRA-1000 (1000 images) [15] and compared with 8 existing methods. Some results are shown in Figure 3.2 and Figure 3.3. In Figure 3.2 visual comparisons, the proposed method is shown to perform better on background suppressing and uniform object enhancement. The previous color contrast-based method [16] (HC, RC) is less satisfactory on these images. Performance evaluation in Figure 3.3 shows that our method consistently achieves higher precision under the same recall.

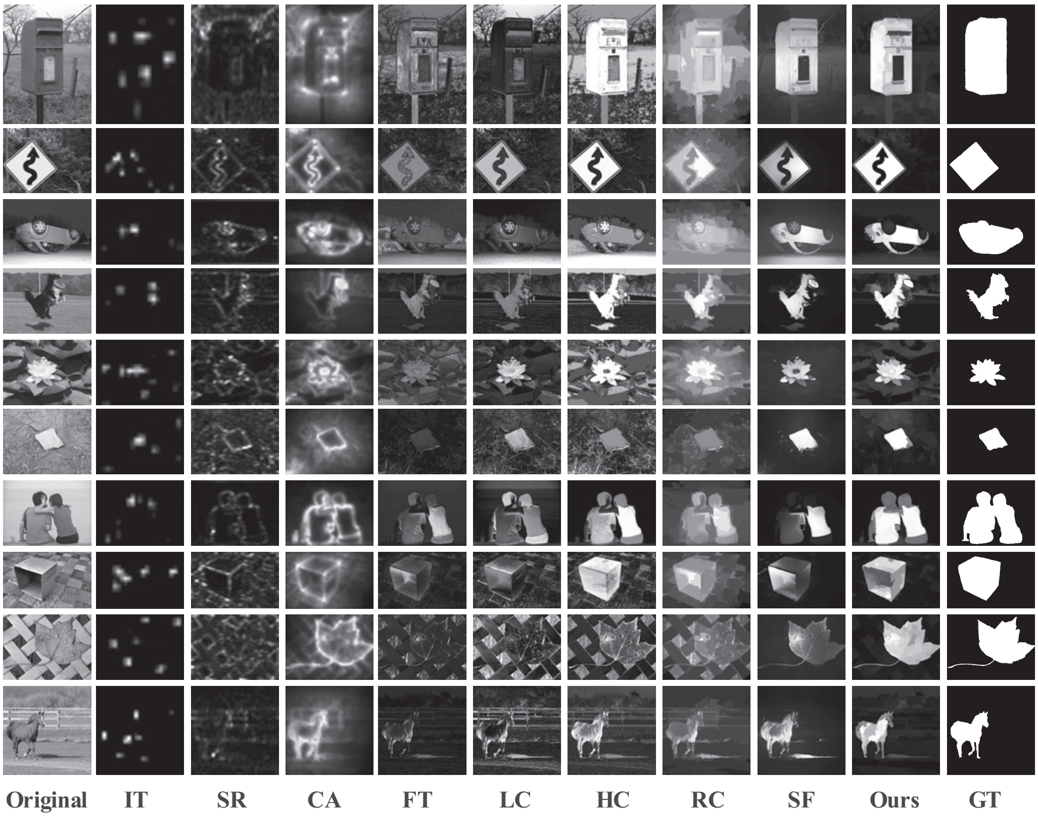

Figure 3.2: Visual comparisons of the proposed method with other existing methods (shown in columns). Images are from public dataset MSRA-1000 [15].

3.2 Paper 2: Salient Object Enhancement using Geodesic Propagation

Basic Ideas: Under the global color contrast hypothesis (e.g. [16] and paper 1), similar colors are supposed to obtain similar saliency after detection. However, this is not always true. Suppose an image (Figure 3.4) “a white sheep stands on green grass under white sky”, where the white sheep is the only salient object. The basic idea of this paper is to use geodesic distance-based propagation to maintain the local connectivity of objects. Since there is large geodesic distance between “sky” and “sheep”, it is possible to develop geodesic distance-based propagation to uniformly enhance the salient object (e.g. the “sheep” regardless of the “sky”). The aim of this paper is to develop a geodesic saliency propagation method to enhance salient objects from a set of coarse saliency maps.

Main contributions: We propose to first leverage an initial coarse saliency map that highlights potential salient regions, based on the observation that

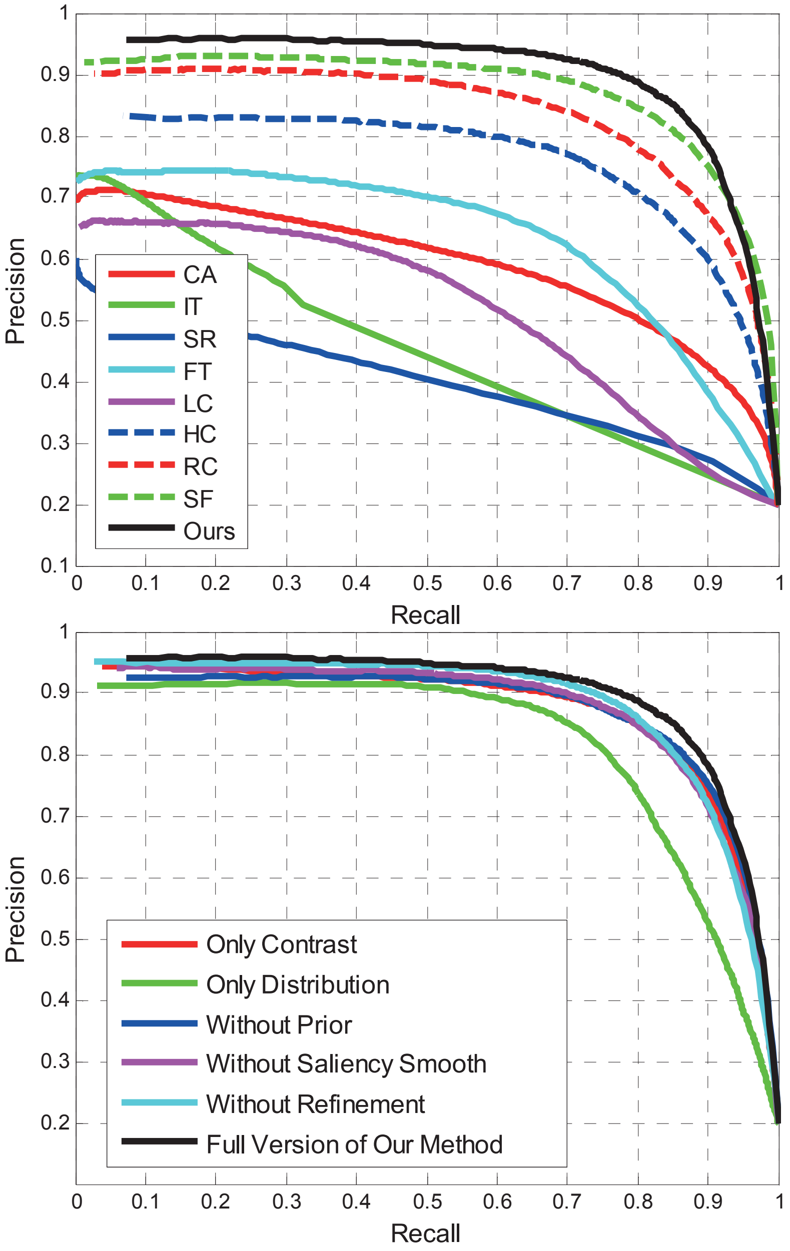

Figure 3.3: Performance evaluation and comparisons. Up: comparisons of precision-recall curves on dataset MSRA-1000. Down: The impact of individual phase on precision-recall curves on MSRA-1000.

salient objects are popped out from both background and clutter, and then propagate energy based on geodesic distance. The extent of propagation between two superpixels is controlled by using an exponential function that is monotonically decreasing in term of geodesic distance. The energy of a

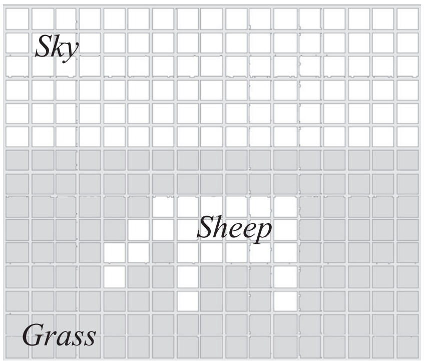

Figure 3.4: An illustrative example of object connectivity.

superpixel after propagation is the summation over all superpixels according to the geodesic distance. Superpixels lie in the same homogenous region would obtain the same saliency value after the propagation. The proposed method is hence capable of rendering a uniform object saliency map while suppressing the background.

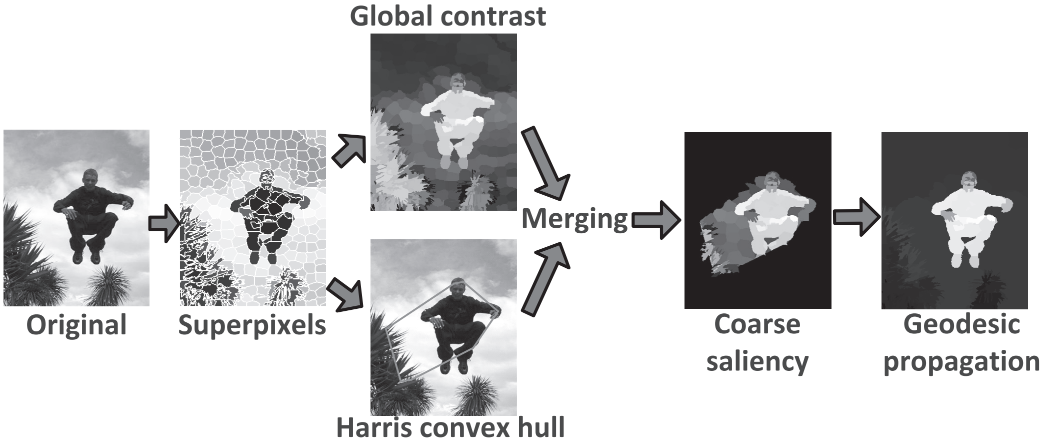

Figure 3.5: The block diagram for the proposed saliency propagation. Noting that background clutter in the color contrast map is suppressed and a missing object part outside the convex hull is recovered after the propagation.

Big picture: Figure 3.5 shows the diagram of the proposed method. An input image is pre-segmented into superpixels. Two cues, i.e. simple global contrast and Harris convex hull are computed. A coarse saliency map is obtained by merging these two cues by pruning saliency values of superpixels outside the convex hull to zero meanwhile maintaining saliency values of superpixels inside the hull. In the geodesic saliency propagation stage, energy from a certain image region is transmitted to its connected regions (e.g. from the body to the missing “feet” of the jumping boy in Figure 3.5 to guarantee them being highlighted). Background clutter inside the hull would be eliminated after propagation.

Results: Tests and comparisons are performed on a public dataset MSRA-1000 (1000 images) [15] and compared with 9 existing methods. Some results are shown in Figure 3.6. Our experiments show that the proposed method achieves comparable results with the 9 existing methods, where noticeable improvement is observed.

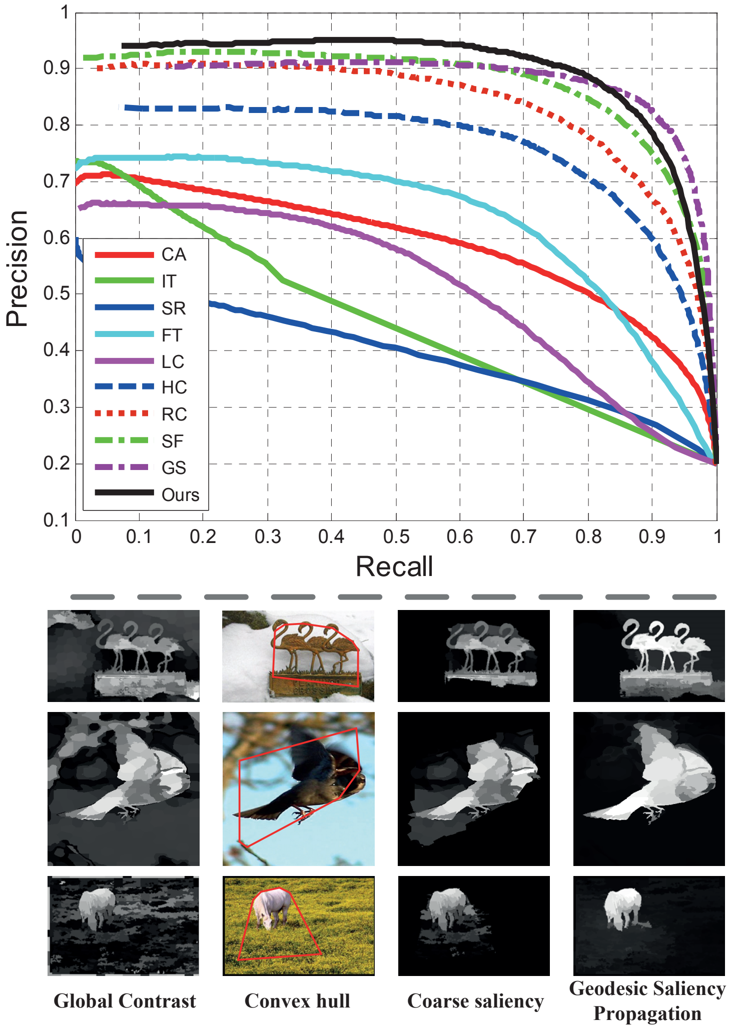

Figure 3.6: Performance evaluation and comparisons. Above the dash line: quantitative comparisons on precision-recall curves with 9 existing methods. Below the dash line: three more examples of propagation.

3.3 Paper 3: Coarse-to-fine Hierarchical Segmentation for Saliency Detection using 2-way Ncut

Basic Ideas: Over-segmenting images into non-overlapping regions to assist saliency computation is one of recent trends, e.g. as the region contrast proposed in [16]. Clustering-based or merging-based segmentation like Meanshift [38] and graph-based segmentation [39] are mostly employed. However, these segmentation methods may generate fragile regions, i.e. an object breaks up into small regions which ignore the object holism, making it difficult to enhance the entire object even though with the assistance of certain energy propagation to smooth the result. Besides, these segmentation methods involve tricky parameter tuning so as to achieve an ideal segmentation that trades off between over- and under-segmenting desired objects. To improve this situation, this paper suggests the use of Normalized cut (Ncut) [35], whose aim is to partition a graph into two discriminative parts, as a tool for segmentation. The basic idea is that the global energy minimization of Ncut guarantees strong discriminative ability to separate object-level contents, generating coarse segmentation that directly benefits saliency detection. Besides, less parameter tuning is required as it partitions a graph in a discriminative way rather than a clustering-based way. The aim in this paper is to generate coarse regions using Ncut and estimate regional saliency from them.
Main contributions: The proposed method uses Ncut to generate hierarchical segmentation, whose application is new to salient object detection. An undirected graph is first constructed from superpixels and is then iteratively partitioned using the 2-way Ncut by solving the second smallest eigenvector of system $(\mathbf{D}-\mathbf{W})\mathbf{u} = \lambda\mathbf{D}\mathbf{u}$. To separate different visual *concepts* (i.e. clusters) in an image, in each iteration the region that requires the lowest Ncut energy is bi-partitioned. The entire partition process leads to a binary tree structure. The effectiveness of the proposed method is experimentally validated. Combination of results from tree hierarchies yields more robust performance than using single hierarchy.
Big picture: An input image is pre-segmented into superpixels. An undirected graph $G = (V, E)$ is constructed by regarding superpixels as vertices and color similarity between them as graph edges. The graph is iteratively partitioned using the 2-way Ncut by solving the second smallest eigenvector of system $(\mathbf{D}-\mathbf{W})\mathbf{u} = \lambda\mathbf{D}\mathbf{u}$. To separate different visual *concepts* (i.e.) clusters in an image, in each iteration the region that requires the lowest Ncut energy is bi-partitioned. During each iteration, regions have the following properties are rendered high saliency:

- segments with closed boundaries (surroundedness).

- segments highly differentiating from the background, i.e. boundary superpixel set (figure-ground contrast).
- segments near the image center (center bias).

Finally, saliency maps from all iterations are combined and refined using CRF. The partitioning process using 2-way Ncut as well as intermediate saliency maps are shown in Figure 3.7.

Results: Tests and comparisons of the proposed method (named "SS" for short) are performed on three datasets including MSRA-1000 [15], SOD [40] and SED [41] with 13 existing methods. Some results are shown in Figure 3.8 and Figure 3.9. In Figure 3.8 visual comparisons, results generated by SS are close to the ground truth and also more consistent with human perceptions. With the assistance of N-cut, SS could handle images that make the state-of-the-art methods less satisfactory, such as the 4th, 7th and 8th row in Figure 3.8. Quantitative results on MSRA-1000 and SOD are shown in Figure 3.9. SS achieves the best precision and F-measure on MSRA-1000 under adaptive threshold, and competitive results on SOD. The mean absolute error (MAE) of SS is also comparable, i.e. second best on both MSRA-1000 and SOD.

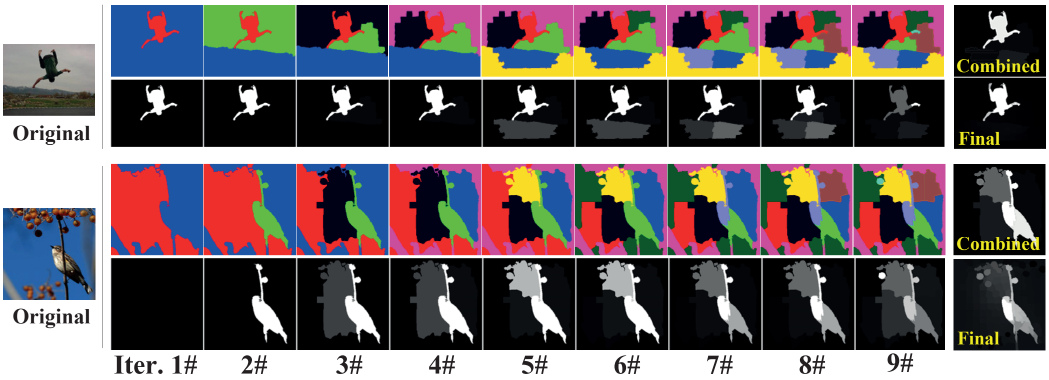

Figure 3.7: Partitioning process and intermediate saliency maps generated by our method. Different colors indicate different segments. The last column shows the integrated and refined final saliency maps.

3.4 Paper 4: Adaptive Region Merging Enhanced by Spectral Decomposition for Saliency Detection

Basic Ideas: To generate coarse segmentation for salient object detection, we consider from image edge viewpoint, since a salient object often presents

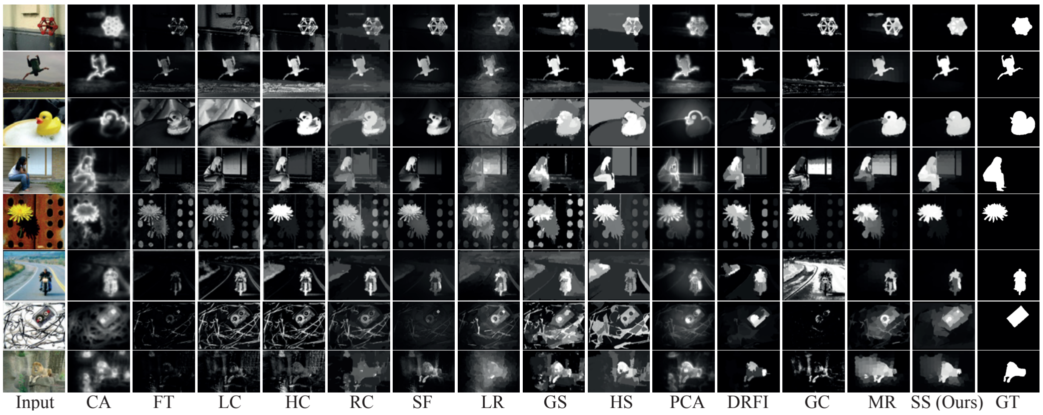

Figure 3.8: Visual comparisons of the proposed method with 13 existing methods (shown in columns). Images are from public dataset MSRA-1000 [15].

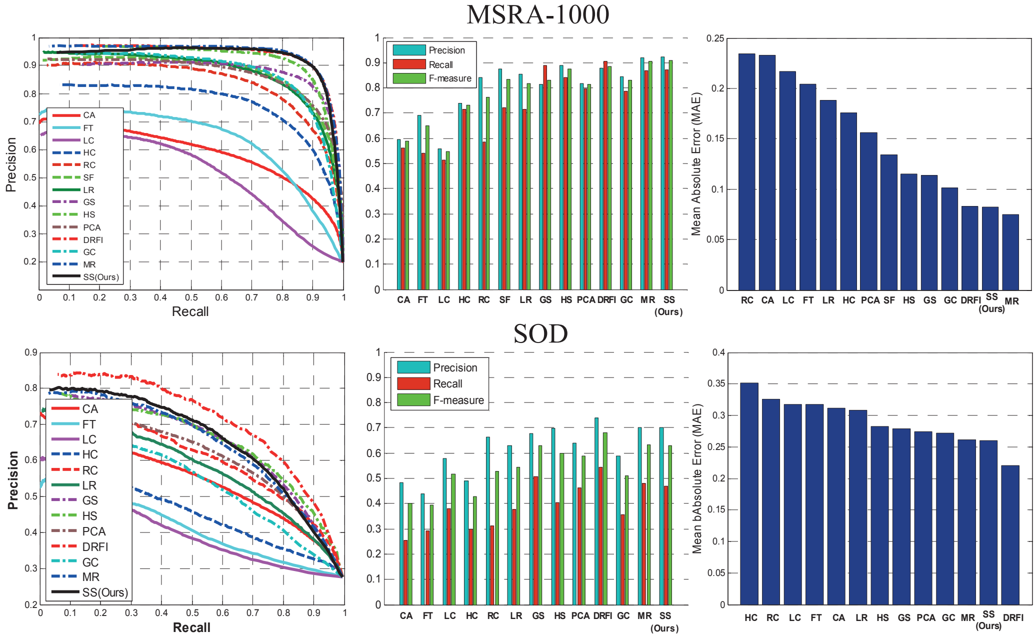

Figure 3.9: Quantitative comparisons of the proposed method with 13 existing methods on precision-recall curves, F-measure under adaptive threshold, and mean absolute error. From top to bottom are results on MSRA-1000 [15] and SOD dataset [40].

strong contours. Graph-based merging technique [39] could be used for this task but it might yield fragile segmentation (as discussed in Paper 3). Small fragile regions could bring noise and drastically affect the performance. To remedy this problem, this paper proposes to include global cues which provide concept-level contour information derived from Ncut. The basic idea is the continuous solutions (smallest eigenvectors) to the Ncut problem are soft indication vectors that distinguish between different clusters. Hence the derivation of eigenvectors results in edge information that may be used to enhance object-level contours. The aim of this paper is to develop a graph-based merging scheme further enhanced by graph-based spectral decomposition. It generates fine-to-coarse hierarchical segmentation in a merging-based manner.

Main contributions: An adaptive multi-level merging scheme for salient object detection from the "contour" point of view is proposed. An undirected graph is first constructed from superpixels, from which the merging starts. At each level of adaptive merging, two regions are fused if their shared contour strength, measured as the average graph edges connecting these two regions, is smaller than a threshold that is increasing w.r.t. the level. The graph edges are constructed by integrating the edge information in the first k smallest eigenvectors of the system $(\mathbf{D}-\mathbf{W})\mathbf{u} = \lambda\mathbf{D}\mathbf{u}$. This globalization procedure for enhancing salient contours is new to salient region detection.

Big picture: Figure 3.10 shows the block diagram of the proposed method. An input image is pre-segmented into superpixels. A preliminary graph affinity matrix $\mathbf{W}$ is first computed from color similarity of adjacent superpixels. An undirected graph $G = (V, E)$ is constructed by connecting adjacent superpixels and computing the graph edges by integrating the edge information from the first k smallest eigenvectors of the system $(\mathbf{D}-\mathbf{W})\mathbf{u} = \lambda\mathbf{D}\mathbf{u}$. Let $R^l = \{R^l_1, R^l_2, ...\}$ be a partition of V in the lth level and $R^l_k \in R^l$ corresponds to its kth part (namely region). At level l of adaptive merging, two components R^l_i, R^l_j are fused if the difference between them $D^l_{ij} \leq Th$, where threshold Th is to control the bandwidth of D^l_{ij} and is increased by a step in the next level. The criterion for measuring the pairwise difference, namely shared contour strength, of two regions R^l_i, R^l_j is defined as:

$$D^l_{ij} = D(R^l_i, R^l_j) = \text{mean}_{v_k \in R^l_i, v_m \in R^l_j, e_{km} \in E}\{e_{km}\} \tag{3.1}$$

where "mean" is averaging operation over graph edges connecting R^l_i and R^l_j. In each level, the merged regions are evaluated under simple saliency measures to generate an intermediate saliency map. Cross-scale integration is finally implemented to form the ultimate saliency map.

Results: Tests and comparisons of the proposed method are performed on three datasets including MSRA-1000 [15] (1000 images), SOD [40] (300 images) and SED [41] (SED1: one object set and SED2: two objects set, each containing 100 images) with 13 existing methods. Some results are shown in Figure 3.11 and Figure 3.12. In Figure 3.11 visual comparisons, our method effectively suppresses the background clutter and uniformly emphasizes the foreground objects, attributed mainly to the hierarchical

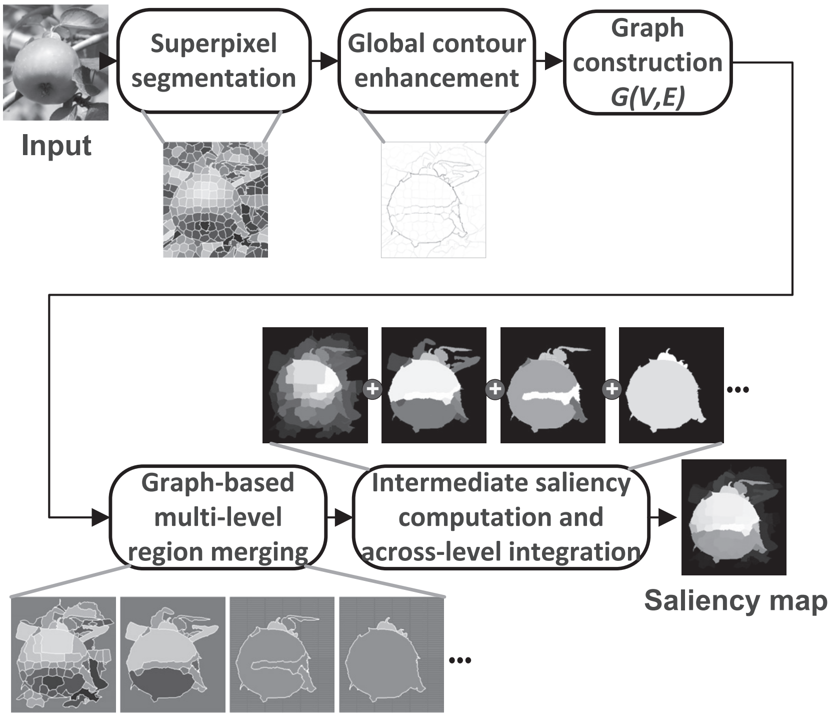
Input
Superpixel segmentation
Global contour enhancement
Graph construction G(V,E)
Graph-based multi-level region merging
Intermediate saliency computation and across-level integration
Saliency map
...

Figure 3.10: The block diagram of the proposed method.

region merging strategy. Further globalization helps pop out holistic salient contours. In Figure 3.12 quantitative comparisons, the performance of our method is comparable to the most recent state-of-the-art techniques, e.g. outperforms HS on MSRA-1000 and SED1 and achieves similar results on the rest. In the adaptive threshold experiment, our method achieves both the highest precision and F-measure on MSRA-1000, 3rd and 2nd F-measure on SOD and SED1. Besides, our method produces the lowest error on MSRA-1000, and consistently 2nd on the rest.

3.5 Paper 5: Graph-based Saliency Detection in Videos

Basic Ideas: While salient object detection in still images has gained a lot of attention, its video version still remains under-explored. This paper addresses salient object detection in videos. Since graph construction is a crucial issue for many salient region detection methods, the basic idea in this paper is to integrate both static appearance and motion cues for graph construction. Besides computing a saliency map intuitively for each video frame, object saliency in consecutive frames should be coherent. We propose to achieve this by using an energy propagation approach on a spatial-

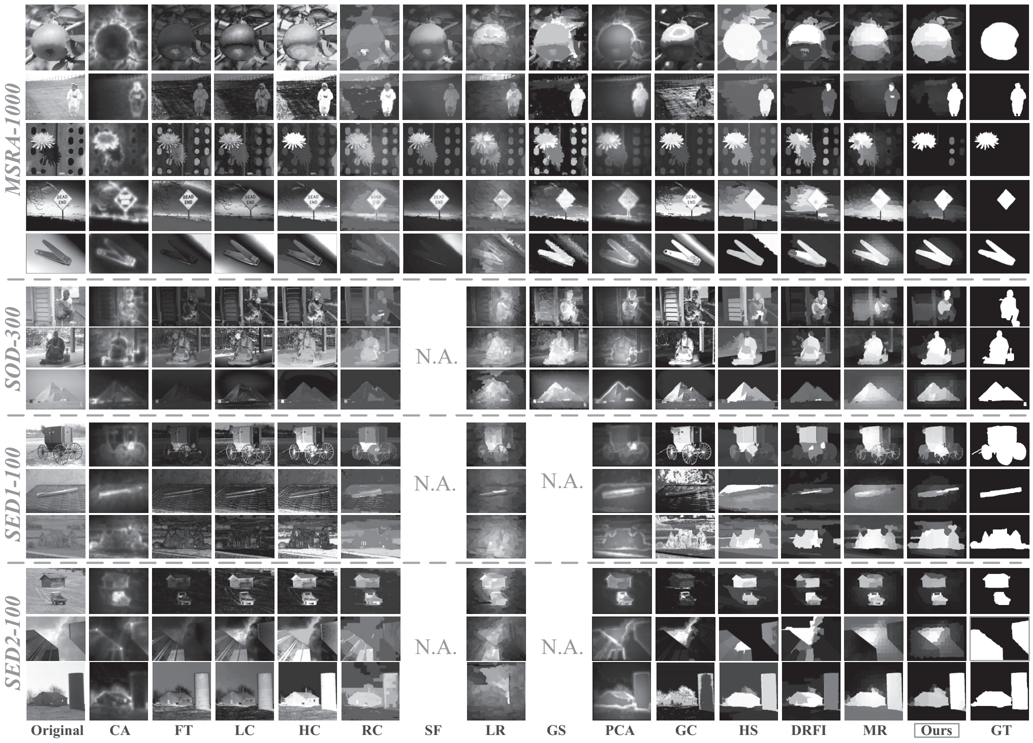

Figure 3.11: Visual comparisons on three benchmark datasets with 13 state-of-the-art methods. N.A. means neither results nor code are publicly available for a certain method.

temporal graph constructed from consecutive frames.

Main contributions: 1) A unified way for extending graph-based methods to video processing is proposed. Motion feature called *mean histogram of optical flows* (MHOF) is extracted from each superpixel. For situations where difference between superpixels is required as edge, the graph edge is constructed by weighted sum of the normalized color and motion difference. For cases where affinity between superpixels is required as edge, the graph edge could be computed from the aforementioned difference using an exponential function. The constructed graph could be leveraged by previous graph-based saliency detection methods to generate a saliency map in a single frame. 2) A method for spatial-temporal saliency smoothing is proposed. A two-frame graph is constructed by connecting superpixels in consecutive frames. The smoothing is conducted by employing Conditional Random Field (CRF) on the constructed two-frame graph, and the relaxed solution is achieved in a closed-form way.

Big picture: Figure 3.13 shows the block diagram of the proposed method. An input video frame is pre-segmented into superpixels. In each frame, an undirected graph $G = (V, E)$ is defined where vertices V are superpixels,

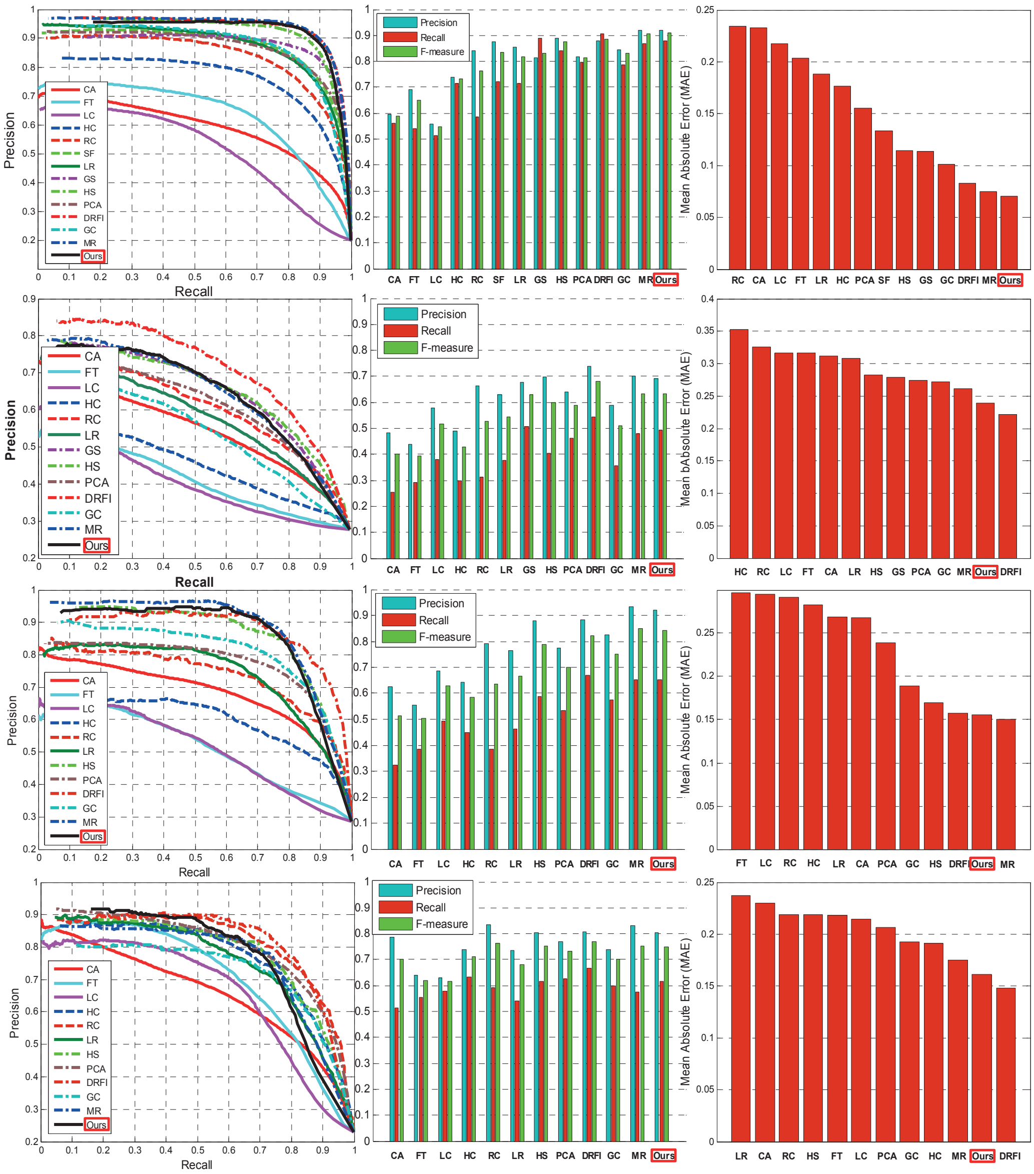

Figure 3.12: Quantitative evaluations on precision-recall curves, adaptive threshold and mean absolute error (MAE) on three benchmark datasets: from top to bottom are MSRA-1000, SOD, SED1, and SED2.

and E are graph edges. A superpixel only connects to its spatial neighbors in the graph. For situations where difference between superpixels is required as edge, E is computed as the weighted sum of the normalized color and motion difference:

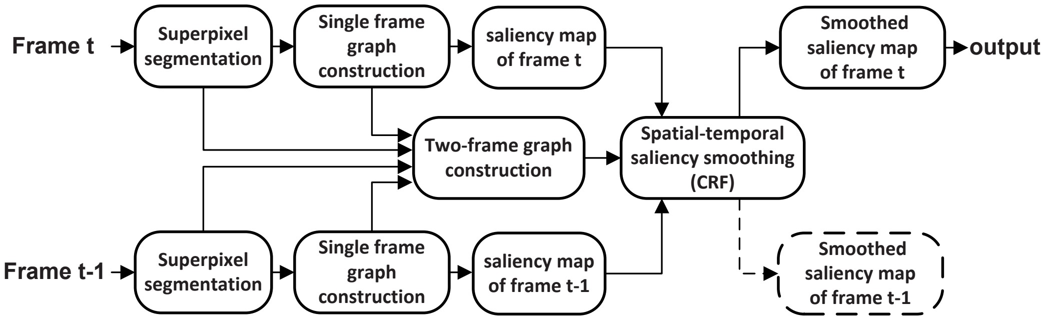
Frame t
Superpixel segmentation
Single frame graph construction
saliency map of frame t
Smoothed saliency map of frame t
output
Two-frame graph construction
Spatial-temporal saliency smoothing (CRF)
Frame t-1
Superpixel segmentation
Single frame graph construction
saliency map of frame t-1
Smoothed saliency map of frame t-1

Figure 3.13: The block diagram of the proposed method. Noting that after spatial-temporal saliency smoothing, we get both smoothed maps for t and $t-1$. But as t is the current frame, we only leverage the smoothed map of t.

$$\hat{d}_{ij} = (1-\alpha)\frac{||\mathbf{c}_i - \mathbf{c}_j||_2}{\max_{R_p \asymp R_q} ||\mathbf{c}_p - \mathbf{c}_q||_2} + \alpha\frac{||\mathbf{h}_i - \mathbf{h}_j||_2}{\max_{R_p \asymp R_q} ||\mathbf{h}_p - \mathbf{h}_q||_2} \tag{3.2}$$

where $\mathbf{c}_i$, $\mathbf{c}_j$ are mean colors of superpixels, while $\mathbf{h}_i$, $\mathbf{h}_j$ are *mean histogram of optical flows* (MHOF). "$R_i \asymp R_j$" means superpixel adjacency. For situations where affinity between superpixels is required as edge, E is defined as the kernel function with respect to $\hat{d}_{ij}$, e.g. an exponential function. After the graph is constructed, previous graph-based methods can be employed to generate a saliency map in a single frame. To achieve spatial-temporal saliency smoothing in Figure 3.13, a two-frame graph is constructed by connecting superpixels in consecutive frames t and $t-1$. Saliency maps generated from t and $t-1$ are processed by CRF on the two-frame graph. Finally the smoothed saliency map for frame t is obtained.

Results: Tests and comparisons of the proposed method are performed on 7 videos from two video datasets: SegTrack dataset and GaTech video segmentation dataset. The constructed graph in each single frame is processed by a manifold ranking-based salient object detection method [29]. Results are shown in Figure 3.14 and Figure 3.15. In Figure 3.14, one can observe that results of the proposed graph construction ("*appearance* + *motion*") are consistently better than the original method [29] ("*appearance*"). The proposed spatial-temporal saliency smoothing ("*appearance* + *motion* + *smoothing*") has been shown useful for improving performance in a noticeable margin, especially in "Birdfall,"Girl" and "Skater" videos. In Figure 3.15 visual comparisons, results are gradually improved from left to right.

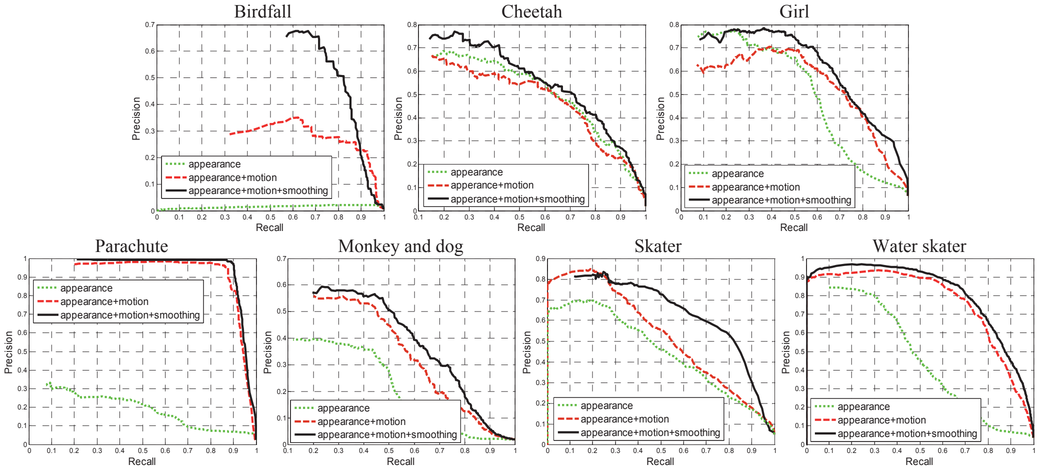

Figure 3.14: Quantitative evaluation on seven videos.

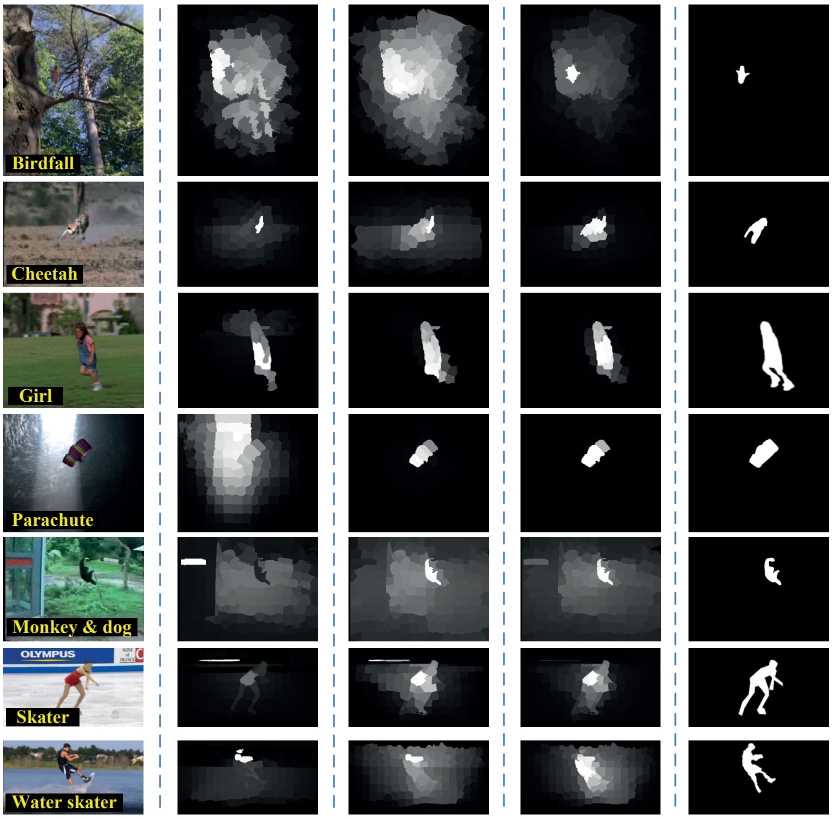

Figure 3.15: Visual comparisons on seven videos. Columns left to right: original frames, "*appearance*", "*appearance* + *motion*", "*appearance* + *motion* + *smoothing*", ground truth.

Chapter 4

Conclusion

The proposed saliency detection method by color contrast and color distribution in Paper 1 effectively combines contrast and distribution cues into a computational superpixel-based framework and renders high quality saliency maps. The exploited distribution prior and saliency smoothing procedure are both proved advantageous and achieve improvement. Complementary results of different cues are also validated. The limitation is that similar to previous global color contrast methods, the proposed method assumes the same color have the same saliency, which may not be true in complex scenarios.

The proposed geodesic saliency propagation in Paper 2 offers an effective way for enhancing object saliency. A coarse map is employed through combining global contrast and Harris convex hulls, followed by propagating the saliency energy to whole image areas through using the geodesic distance between superpixels. The coarse map is not restricted to certain cues and any other hypotheses can be employed. Observation is found that relaxing the assumptions on global contrast and Harris convex hulls to some extent would not significantly impact the propagated results.

The proposed hierarchical segmentation uses Ncut as a pre-segmentation technique for salient object detection, which helps effectively discover the object holism. Paper 3 introduces a coarse-to-fine partition framework accompanied by incorporating saliency measurement for segments based on Gestalt laws and statistical prior. The proposed method achieves better or comparable results on three widely used datasets with 13 state-of-the-art solutions. Since Ncut is a kind of balanced cut, assumption is made that desired salient contents can be split out after certain hierarchy. The limitation is that it is not true for very small objects.

Regarding to the proposed saliency detection scheme based on adaptive multi-level region merging in Paper 4, the core is adaptive region merging and globalization by graph-based spectral decomposition to enhance salient contours. The former combines potential foreground and background regions and the latter improves contour completions. When combined together, they greatly improve the accuracy on detecting holistic objects and effectively suppress the background. Experiments have shown the proposed method achieves state-of-the-art performance on three commonly used benchmark datasets. The limitation is that since in each level of merging an intermediate saliency map is computed, the computation cost scales with the level number. Even under a limited level number, the computation cost is still heavy compared to the methods in Papers 1-3.

Finally, graph-based methods are extended to video processing in Paper 5. The proposed graph construction has integrated both static and motion cues by using a novel feature: *mean histogram of optical flows* (MHOF) that effectively captures the statistical motion information in each superpixel. The advantage of the proposed method in video processing is shown by applying the manifold ranking-based method to constructed graphs on seven videos. The proposed spatial-temporal smoothing operation which incorporates the spirit of CRF is shown to make saliency output more coherent, and to enhance the final performance.

Comparing the methods proposed in Papers 1-4 as they are all designed for salient region detection in still images, in term of detection performance, Paper 3 and Paper 4 are better than Paper 1 and Paper 2, since the formers are the latest and employ Ncut to generate good segmentation. Meanwhile incorporating new measures such as “surroundness” for removing regions cropped by image boundaries also improves the performance. To categorize these methods, methods in Paper 1 and Paper 2 should be divided into heuristic color contrast-based type whereas the methods in Paper 3 and 4 should belong to multi-scale segmentation type for increasing the detection reliability. Compared to Paper 3 and Paper 4, Paper 1 and Paper 2 have advantages on detecting salient objects from relatively simple background in fast speed, but Paper 3 and Paper 4 show more robustness in complex scenarios.

Future work could include: applying methods in Paper 3 and Paper 4 to video processing by incorporating the graph construction from the method in Paper 5. Additionally, the graph affinity in this thesis is based on average colors from superpixels, which only give approximation of local color statistics without considering texture. This could degrade detection performance when colors of objects and the background are similar. This can be improved by constructing the graph affinity from edge detection that integrates local brightness, color, and texture cues, so that better delineation between objects and background can be obtained.

References

[1] A. Triesman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[2] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurbiology*, vol. 4, pp. 219–227, 1985.

[3] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 1, pp. 185–207, 2013.

[4] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[5] J. Han, K. Ngan, M. Li, and H. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, vol. 16, no. 1, pp. 141–145, 2006.

[6] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.

[7] Y. Ma, X. Hua, L. Lu, and H. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[8] F. Stentiford, "Attention based auto image cropping," in *Workshop on Computational Attention and Applications, ICVS*, 2007.

[9] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[10] Y. Ding, X. Jing, and J. Yu, "Importance filtering for image retargeting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[11] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[12] T. Chen, M. Cheng, P. Tan, A. Shamir, and S. Hu, "Sketch2photo: Internet image montage," *ACM Transactions Graph*, vol. 28, no. 5, pp. 1–10, 2006.

[13] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 20, no. 11, pp. 1254–1259, 1998.

[14] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," *ACM Multimedia*, pp. 815–824, 2006.

[15] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[16] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[17] F. Perazzi, P. Krahenbul, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[18] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[19] K. Shi, K. Wang, J. Lu, and L. Lin, "Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[20] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[21] T. Liu, Z. Yuan, J. Sun, J. Wang, and N. Zheng, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, no. 2, pp. 353–367, 2011.

[22] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[23] Y. Lu, W. Zhang, H. Lu, and X. Xue, "Salient object detection using concavity context," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[24] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[25] M. Cheng, J. Warrell, W. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[26] J. Lafferty, A. McCallum, and F. Pereira, "Submodular salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[27] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Random walks on graphs for salient object detection in images," *IEEE Transactions on Image Processing (IP)*, vol. 19, no. 12, pp. 3232–3242, 2010.

[28] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *European Conference on Computer Vision (ECCV)*, 2012.

[29] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[30] B. Jiang, L. Zhang, H. Lu, C. Yang, and M. Yang, "Saliency detection via absorbing markov chain," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[31] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[32] Y. Xie and H. Lu, "Visual saliency detection based on bayesian model," in *IEEE International Conference on Image Processing (ICIP)*, 2011.

[33] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[34] A. Criminisi, T. Sharp, C.Rother, and P. Pérez, "Geodesic image and video editing," *ACM Transactions on Graphics*, vol. 29, no. 5, p. 134, 2010.

[35] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 22, no. 8, pp. 888–905, 2000.

[36] U. von Luxburg, "A tutorial on spectral clustering. statistics and computing," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[37] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *International Conference on Machine Learning (ICML)*, 2001.

[38] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24, no. 5, pp. 603–619, 2002.

[39] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision (IJVC)*, vol. 59, no. 2, pp. 167–181, 2004.

[40] V. Movahedi and J. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *IEEE Computer Society Workshop on Perceptual Organization in Computer Vision*, 2010.

[41] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.