Semantic image segmentation and cosegmentation

Zhu, Hongyuan

2015

Zhu, H. (2015). Semantic image segmentation and cosegmentation. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/62920>

<https://doi.org/10.32657/10356/62920>

SEMANTIC IMAGE SEGMENTATION AND COSEGMENTATION

ZHU HONGYUAN

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Doctor of Philosophy

2015

Acknowledgments

Pursuing a doctor degree has never been an easy story. When I took the first step into NTU, I reminded me of the challenges ahead, though the road to explore the truth of the vision world is far more longer and bumpier than I once imagined. My exploration just takes a few steps.

I would like to express my sincere thanks and appreciation to my supervisor, Prof. Cai Jianfei (SCE), and co-supervisor, Prof. Nadia Thalmann (IMI) and Prof. Zheng Jianmin (SCE) for their patient guidance, constant support and invaluable suggestions. Their knowledge, suggestions, and discussions help me to become a capable researcher. Their encouragement spangled the road of my research, which helps me to overcome the difficulties encountered in my career and daily life. Their enthusiasms toward research and life build up good examples for me to follow.

I would also like to thank Dr. Lu Jiangbo (ADSC, Singapore) to offer me the opportunity to work with him on several challenging topics in computer vision. During my internship at ADSC, I broadened my horizon and developed further understanding of my thesis topic. It is my honor to get to know a lot of excellent researchers at ADSC: Hongsheng, Nianjuan, Weiyong, Dongbo, Shenghua, Bingbing, Chris, Prof.Mihn... Having discussions with them always give me sudden inspiration for the challenge problem.

I also want to thank my colleagues in NTU, for their generous help. I am also indebted to Dr.Zhang Juyong (USTC, China), for his technical and personal suggestions for my first work. My gratitude also goes to Yuzhe, Yuanquan, Wenjing, Fanghui, Jingshang, Yingnan, Yangqin, Guibing for their friendship and support.

My thanks also goes to my family, for their constant love and encouragement. Thanks my parents for always believe in myself and gave me support during every important turning point of my life.

Last but not least, I would thank myself for my endurance and constant efforts. Actually, every tough moment in my life I strived through. It's hard to imagine what if I gave in before any of them (e.g. the first work stillborn, the car accident half year ago), my life will be another story... Thanks Hongyuan, please keep working hard!

Contents

Acknowledgments	i
List of Figures	v
List of Tables	xiii
1 Introduction	3
1.1 Background	3
1.2 Objectives and Scope	4
1.3 Contributions and Thesis Outline	6
2 Literature Review	9
2.1 Bottom-Up Image Segmentation	9
2.2 Conditional Random Field	13
2.2.1 Continuous Covex Relaxed Potts Model	15
2.2.2 Discrete Potts Model	17
2.3 Interactive Segmentation	20
2.4 Top-Down Object Segmentation	21
2.5 Multiple Image Segmentation	22
3 Object Segmentation using Low Level Cues	26
3.1 Related Work	27

3.2	Proposed Method	27
3.2.1	Global Structure and Boundary Feature Extraction	28
3.2.2	EM Initialization	30
3.2.3	Variational Segmentation	30
3.2.4	Selecting The Number Of Regions	34
3.3	Experiments	34
3.3.1	Test on Berkeley Benchmark	34
3.3.2	Visual Results	36
3.3.3	Limitations	36
4	Multiple Foreground Recognition and Cosegmentation	38
4.1	Relations and Comparison with Previous Work	40
4.2	Problem Formulation and Our CRF Model	42
4.2.1	Proposed CRFs Framework for MFRC	43
4.3	Incorporating Object Cues	44
4.3.1	Fast Object Detectors with Boosted Color Bins	44
4.3.2	Detector-Based Robust Consistency Potentials	45
4.3.3	Contour for Object Boundary Reasoning	47
4.4	Segment-Based Label Consistency Potential	50
4.5	Experimental Results and Discussions	51
5	Multiple Human Identification and Cosegmentation	54
5.1	Related Works	55
5.2	Problem Formulation and Proposed CRF Framework	58
5.2.1	Proposed CRF Framework for MHIC	59

5.3	Cross Region Based Pixelwise Unary Potential	60
5.4	Contour Based Pairwise Smoothness Potential	61
5.5	Incorporating Human Detector Based Cues	62
5.5.1	Efficient Color Line and Texton Histogram Based Poselet Human Detector	63
5.5.2	Edge-Aware Soft Shape Map with GGDT and JBF	64
5.5.3	Detector-Based Robust Consistency Potential	67
5.6	The NTU-MHIC Dataset	68
5.7	Experimental Results and Discussions	76
6	Conclusions and Future Work	79
6.1	Conclusions	79
6.2	Future Directions	80
6.2.1	Fine Grained Object Segmentation and Cosegmentation	80
6.2.2	Automatic Multiple Foreground Recognition and Cosegmentation in Video	81
	Appendix - List of Publications	82
	References	83

List of Figures

1.1	An image from Berkeley segmentation dataset [MFM04] with hand labels from different human subjects.	5
1.2	1.2.b and 1.2.c: Existing low-level cues based segmentation methods such as Ncut [SM00] and Felz-Hutt [FH04] often over-segment the image or cannot align well with true object boundary. 1.2.d: The original Potts model relying on local color feature mistakenly treats part of the mountain and chair, the ground and the sky as the same segments.	6
2.1	Examples of interactive segmentation by using bounding box (a $\sim$ b) [RKB04] or scribbles (c $\sim$ d) [YCZL10]	21
2.2	Example of the object class agnostic segmentation [CS12].	22
2.3	Examples to illustrate the ambiguity of the “object” notion, which can vary according to scale, context and repetitiveness.	23
2.4	Examples of semantic segmentation [KK11] which is trained by using pictures with human labeled ground truth similar to (b) to segment the test image in (c) and produce the final segmentation in (d).	23
2.5	Simultaneously segmenting one common foreground object from a set of image (a) is a well studied problem [CLL11]. Recently, Kim et al. [KXLK11] considers multi-class cosegmentation problem (b).	24

2.6	Given a user’s photo stream about certain event, which consists of finite objects e.g. reg-girl, blue-girl, blue-baby and apple basket. Each image contain an unknown subset of them, which we called “Multiple Foreground Cosegmentation” problem. Kim [KX12] proposed the first method to extract these irregularly occurred objects from the photo stream (b) by using few bounding box provided by user in (a).	25
3.1	(a), (c): An image and its top four non-zero eigenvalue eigenvectors. (b): The gPb contour map generated from eigenvector captures clean and salient image boundaries	29
3.2	An example to compare the segmentation results by using different features. 3.2.e:Combining global spectral, color and spatial features achieves the best result	32
3.3	Comparison of the results of our method using the two different $C_i(x)$ definitions in (2.24) and (3.6), respectively. Some boundary problems due to using (2.24) are circled in (c).	33
3.4	Examples of some images randomly selected from the BSDS500 test dataset and their corresponding segmentation results of different methods with optimal parameters tuned over the training set. Note that red and yellow contours depict the region boundaries. For Blobworld, it uses white contours for boundaries and the gray regions indicate unlabelled pixels. . . .	37

4.1	Visual comparison between the state-of-the-art MFC method [KX12] and our method on two images from the FlickrMFC dataset. The MFC method wrongly annotates the <i>apple+bucket</i> foreground as the <i>girl+red</i> , and the hair of the <i>girl+red</i> foreground as the <i>girl+blue</i> foreground. It also misclassifies the background as the <i>dog+white</i> in the second test image. Our method does a much better job in resolving the ambiguity in the MFRC task with a hierarchical CRF model using higher-level object cues.	39
4.2	Overview of our MFRC system. It integrates various object-level cues with low-level cues in a probabilistic CRF model.	41
4.3	Effects of the object detector-based robust consistency potential. (a) A user-annotated image with two foreground instances labeled with bounding boxes. (b) Applying a learned object detector to a novel image. Shown are top-scored bounding boxes. Segmentation result (c) without and (d) with using the proposed object detector-based label consistency potential.	46
4.4	Behavior of the object detector potential as a soft constraint. (a) Input image. (b,c) Two object detection results. Our results (d) without and (e) with using the object detector potentials.	46
4.5	Proposed contour-layout filter based on the <i>gPb</i> contour map. (a) Input image. The blue bounding boxes indicate example detection results for the <i>apple bucket</i> image. (b) The intensity-inverted gPb map. (c) Proposed contour-layout filter (see the text for the details). (d) Close-up views of false/true positive object detections.	48

4.6	SLIC segmentation and segment-level consistency potentials. (a) SLIC segments [ASS ⁺ 12]. (b) Segment-level and (c) pixel-level classifier response map for the <i>baby</i> foreground. (d) Robust region consistency constraint. The big orange dot denotes a segment assigned a label l_s . The blue dot denotes an outlier pixel.	50
4.7	Segmentation accuracy comparison between our method (MFRC) and other baselines (MFC [KX12], COS [KXLK11], DC [JBP10], LDA [RFE ⁺ 06]) for the FlickrMFC dataset. The S and U denote whether the method is supervised or unsupervised.	51
4.8	Some randomly drawn examples from seven groups of the FlickrMFC dataset. From top to bottom, each set presents its input images, color-labeled segmentation results. The colored tag below each set indicates which category each region is assigned to.	53
5.1	Overview of our MHIC algorithm. Given a set of event photos, a user annotates only a few of them to indicate human instances of interest with bounding-boxes in different colors. Our algorithm then jointly localizes, identifies and segments out all the human instances for the given image set. It tackles the MHIC task by integrating object-level cues with mid-level and low-level cues in a probabilistic CRF framework. The proposed algorithm yields the segmentation of human instances of interest in each image, and identifies them by taking on the same color annotated earlier for the corresponding human class.	56

5.2	The proposed cross region based pixelwise unary potential outperforms the conventional single pixel based counterpart [ZLC ⁺ 14] for the MHIC task. (a) Two example adaptive cross regions defined at pixel p and p' . (b) A close-up view of the adaptive cross region defined at pixel p . (c) The color histograms H_c and texon histograms H_t extracted from the respective cross support regions at p and p' . (d) The conventional pixel-level classifier response map for the human class “boy-in-blue”. (e) The proposed adaptive region-based classifier response map for the same “boy-in-blue” class. (f) The final multi-class labeling result based on (d). (g) The final multi-class labeling result based on (e). (h) The ground truth label map.	58
5.3	Comparison between the two different contrast-sensitive smoothness measures. Left column: (a) Input color image. (b) Ground-truth label map. Middle column: (c) Canny edge detection map (intensity inverted) for the input image. (d) The final labeling result based on (c). Right column: (e) gPb contour detection map (intensity inverted) for the input image. (f) The final labeling result based on (e). The white dashed circles highlight three places where using a gPb based pairwise potential yields much better labeling results than the Canny edge based potential.	62

5.4	The proposed human instance detector allows to utilize the soft segmentation mask associated with each poselet template to define a human-sensitive shape prior. (a) Input image. (b) The final detected human bounding boxes (solid boxes). They are computed by running all the poselet detectors (three example dashed boxes) at various positions and scales as in [BMBM10]. (c) The detection scores for the three example poselet detectors. (d) The human class-agnostic response map computed by fusing the soft masks transferred from all poselet templates, with the fusion weights defined by the respective poselet detection scores. (e) The human class-specific response map (four classes in this example) by further considering the detection score evaluated by our proposed multi-class color line texon classifier.	63
5.5	Effects of the soft shape cue and the higher order object detector potential. (a) Input images. (b) Detected bounding boxes for corresponding human classes. (c) The initial transferred poselet masks. (d) The enhanced shape masks using generalized geodesic distance transform [CSRP10]. (e) The final soft shape maps generated by using an edge-aware joint bilateral filter [ABD10] to filter (d). (f) The labeling results without using soft shape cues. (g) The labeling results by combining shape cues and the cross region based unary term. (h) The results using higher order detector potentials. (i) The ground-truth label maps. By incrementally including the shape potential and the higher order detector potential, one can observe the improvements in labeling accuracy.	65

5.6	Ground-truth annotation for the NTU-MHIC dataset. (a) Input image. (b) A user segments out the first human (the green contour) and indicates him as the human of interest by placing a red stroke. (c) The generated ground-truth label for the first human class. (d) The second human in the image is annotated following the same procedure. The whole process to annotate an image takes 3 ~ 4 minutes.	69
5.7	The percentage of the images containing 0 ~ 5 human instances in each subset of the NTU-MHIC dataset.	70
5.8	Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each row presents input images, their corresponding ground-truth, color-labeled segmentation results for supervised MFC [KX12], MFRC [ZLC ⁺ 14] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.	70
5.9	Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each row presents input images, their corresponding ground-truth, color-labeled segmentation results for supervised MFC [KX12], MFRC [ZLC ⁺ 14] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.	71
5.10	Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each row presents input images, their corresponding ground-truth, color-labeled segmentation results for supervised MFC [KX12], MFRC [ZLC ⁺ 14] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.	72

5.11	Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each row presents input images, their corresponding ground-truth, color-labeled segmentation results for supervised MFC [KX12], MFRC [ZLC ⁺ 14] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.	73
5.12	Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each row presents input images, their corresponding ground-truth, color-labeled segmentation results for supervised MFC [KX12], MFRC [ZLC ⁺ 14] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.	74
5.13	Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each row presents input images, their corresponding ground-truth, color-labeled segmentation results for supervised MFC [KX12], MFRC [ZLC ⁺ 14] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.	75
5.14	Segmentation accuracy comparison between our method (MHIC) and other baselines (MFRC [ZLC ⁺ 14], MFC [KX12], CoS [KXLK11], and DC [JBP10]) for the MHIC dataset. The S and U denote whether the method is super- vised or unsupervised.	76

List of Tables

3.1	Boundary benchmarks on BSDS500.	35
5.1	Component evaluation of the proposed MHIC algorithm (column 3–5) in comparison with MFRC [ZLC ⁺ 14].	78

Summary

This thesis considers the challenging problem of automatically segmenting an image or a photo stream into a few number of regions that correspond to semantic objects or high-level structures, which can be highly beneficial for various computer vision tasks given existing huge amount of images over the internet. We tackle the challenge by dividing it into three subtasks and propose efficient solutions to each of them.

For the first task, we aim at segmenting an image into a small number of regions that correspond to higher level objects or parts in an unsupervised manner. While such object segmentation usually requires additional high-level knowledge or learning process, we explore what low level cues can produce for this purpose. Our idea is to construct a feature vector for each pixel, which elaborately integrates spectral attributes, color Gaussian Mixture Models and geodesic distance, such that it encodes global color and spatial cues as well as global structure information. Then we formulate the Potts variational model in terms of the feature vectors to provide a variational image segmentation algorithm that is performed in the feature space. We also propose a heuristic approach to automatically select the number of segments. The use of feature attributes enables the Potts model to produce regions which are coherent in color and position, comply with global structures corresponding to objects or parts of objects and meanwhile maintain a smooth and accurate boundary.

For the second task, we move forward to a higher level photo stream parsing task. This task is very challenging as it aims to parse a set of realistic event photos which contain irregularly occurring multiple foregrounds with high appearance and scene configuration variations. We cast the multiple foreground recognition and cosegmentation (MFRC) problem within a conditional random fields (CRFs) framework in a principled manner. We capitalize centrally on the key objective that MFRC is to segment out and annotate the foreground objects or the “things” rather than the “stuff”. To this end, we exploit a few complementary objectness cues (e.g. contours, object detectors and layout) and propose novel and efficient methods to capture object-level information. Integrating object potentials as soft constraints (e.g. robust higher-order potentials defined over detected object regions) with the low-level unary and pairwise terms holistically, we solve the MFRC task with a probabilistic CRF model by using minimal amount of user annotations on just a few example photos.

For the last task, we consider the problem to label and segment human objects instead of other non-human objects or the “stuff” from a photo stream. We refer to such a problem as Multiple Human Identification and Cosegmentation (MHIC). This problem is difficult as human exhibits much more deformation than other classes, and typical MFRC method failed on this case. To identify specific human subjects, we propose an efficient human instance detector by combining an extended color line model with a poselet-based human detector. Moreover, a novel soft human shape cue is proposed, which is initialized by the human detector, then further processed through a generalized geodesic distance transform, and refined finally with a joint bilateral filter. We also propose to capture the rich feature context around each pixel by using an adaptive cross region data structure, which gives a higher discriminative power than a single pixel-based estimation. Then all the cues are fused into a principled conditional random field (CRF) framework. We evaluate our method over a newly created NTU-MHIC human dataset, which contains close to 190 images with manually annotated ground-truth segmentation. Both visual and quantitative results demonstrate that our method achieves state-of-the-art performance for the challenging MHIC task.

Chapter 1

Introduction

1.1 Background

With the ubiquitousness of the internet, digital cameras and smart phones, seas of images and videos are uploaded and shared across different websites and platforms. To automatically discover valuable information from such big volume of visual data, people resort to develop computer vision algorithms for tasks such as object recognition, tracking, 3D reconstruction et al. to enhance computers' ability to understand the visual world. These tasks are so complex by themselves that direct feeding raw pixels or video frames to the existing machine learning algorithms are not computationally feasible and can lead to severe overfitting. On the other hand, there exist structures in nature images or photos which could be exploited with the the learning techniques. Segmentation, as one of the tools which can extract these structures and transform an image into more meaningful intermediate representations, can greatly benefit higher level vision tasks.

Image segmentation, as a basic operation in computer vision, refers to the process to divide a natural image into K non-overlapping meaningful entities (e.g., objects or parts). It has been proved quite useful in many image processing and computer vision tasks. For example, image segmentation has been applied in image annotation [CBGM02] by decomposing an image into several blobs correspond to objects. Superpixel segmentation, which transforms millions of pixels into hundreds or thousands of homogeneous regions [ASS⁺12, CM02], has been applied to reduce the model complexity, and increase speed and accuracy of some complex vision tasks, such as estimating dense correspondence field [LYMD13], scene parsing [KT07a] and body model estimation [Mor05]. [GLAM09, CS12, EH14] have used segmented regions to facilitate object recognition, which provide better localization than sliding windows. The techniques developed in image segmentation, such as Mean Shift [CM02] and Normalized Cut [SM00] also have been widely used and adapted in other areas, such as data clustering and density estimation.

Much research has been conducted on image segmentation in literature. Unsupervised segmentation, as one classic topic in the computer vision, has been studied since 70's. Early techniques focus on local region merging and splitting [OPR78,BF70,OTS79], which borrow ideas from the clustering community. Recent techniques, on the other hand, seek to optimize some global criteria [CM02,FH04,SM00,VS91,Bea06,CV01,OS88]. Interactive segmentation methods [RKB04,YCZL10,ACZZ12] which depend on user input have been applied in some industrial applications, such as Microsoft Office and Adobe Photoshop. The substantial development of the image classification [SWRC09], object detection [FGM10], superpixel segmentation [ASS⁺12] and 3D scene recovery [HEH08] in the past few years have boosted the research in the supervised scene parsing [KLT09,GFK09,LYT11,TL13,KK11,LRKT10]. With the emerging large scale image databases e.g. the ImageNet [DDS⁺09] and personal photo streams on Flickr, the cosegmentation methods [KX12,KXLK11,JBP12,RSLP12,JBP10,CLL11,MSP11,MSD09] which can extract recurring objects from the image sets has attracted increasing attentions these years.

1.2 Objectives and Scope

One would expect an segmentation algorithm to decompose an image into the “objects” or meaningful parts which are consistent with the user intention, image attributes and human perception. However, what makes an “object” or a part “meaningful” can be ambiguous. An “object” can be referred to as a “thing” (a cup, a cow, etc.), a kind of texture (wood, rocks) or even a “stuff” (a building or a forest). Sometimes, an “object” can also be part of the other “objects”. Lacking a general definition of the “object” makes purely bottom-up segmentation a challenging and ill-posed problem. Fig. 1.1 illustrates one such example, different human subjects have various ways to interpret the image. In this sense, what makes a ‘good’ segmentation result needs to be properly defined.

Research in human perception has provided some useful guidelines to develop segmentation algorithms which can match human’s intention. For example, cognition study [HS97] shows that human vision views part boundaries at those with negative minima of curvature and the part salience depends on three factors: relative size, boundary strength and degree of protrusion. Gestalt theory and other psychological studies have also developed various principles reflecting human perception, which include: (1) human tends to group elements which have similarities in color, shape or other properties; (2) human favors linking contours whenever elements of the pattern establish an implied direction. Many unsupervised methods have been developed with some of the guidelines in mind [CM02,FH04,SM00,VS91,Bea06,CV01,OS88], although there is a significant performance gap between the unsupervised bottom-up methods and the human perception to extract object like regions or parts. To reduce this gap, different levels of supervision

Figure 1.1: An image from Berkeley segmentation dataset [MFM04] with hand labels from different human subjects.

have been involved. For example, the interactive methods [RKB04, YCZL10, ACZZ12] and fully supervised methods [KLT09, GFK09, LYT11, TL13, KK11, LRKT10] which involve careful human labeling have achieved the state-of-the-art performance by designating the objects-of-interest, though the labeling cost is relatively high. Cosegmentation [KX12, KXLK11, JBP12, RSLP12, JBP10, CLL11, MSP11, MSD09] defines the object-of-interest as the one which is repeatedly appear in the images. Thus, it can be treated as a kind of weakly supervised methods. Its labeling cost and accuracy lies between the unsupervised method and the fully supervised method.

Another challenge which makes “object” segmentation difficult is how to effectively represent the “object”. When human perceives an image, elements in the brain will be perceived as a whole, but most images in computers are currently represented based on low-level features such as color, textures, curvatures, convexity et al. Such low-level features reflect local object information, which are difficult to capture global object shapes. These features are also sensitive to lighting and perspective variations, which can further cause current algorithm to over-segment the image into trivial regions. An example of the limitation of such low level cues is illustrated in Fig. 1.2. Fully supervised methods aforementioned can learn higher level and global cues, but they can only handle limited class of objects and require per-class labeling.

Therefore, the ultimate goal of this thesis is to design and develop more robust, efficient and accurate algorithms satisfying the perceptual criteria mentioned above. We start with the classic unsupervised bottom-up image segmentation problem. Unlike the methods which aim at producing superpixel over-segmentation, we aim at producing regions which correspond to objects or high level structures. The core challenge of this problem is how to develop low level cues which can capture global object information and how to elaborately incorporate such cues into the model which can deliver desirable segmentation; we develop unique solution to handle this challenge. Then we consider the

Figure 1.2: 1.2.b and 1.2.c: Existing low-level cues based segmentation methods such as Ncut [SM00] and Felz-Hutt [FH04] often over-segment the image or cannot align well with true object boundary. 1.2.d: The original Potts model relying on local color feature mistakenly treats part of the mountain and chair, the ground and the sky as the same segments.

challenging photo stream parsing problem—Multiple Foreground Recognition and Cosegmentation (MFRC), we propose to solve the problem within a principled and efficient conditional random field framework by explicitly fusing the high level “object” notion with the low-level and mid-level cues, which has not been considered by the previous works. Finally, we extend the MFRC framework to the problem of parsing multiple foreground human instances from the photo stream, by incorporating the unique ‘human’ constraint, which greatly advances the performance for the human labeling tasks.

1.3 Contributions and Thesis Outline

The thesis has made several contributions to the research of semantic image segmentation and cosegmentation:

- We propose to solve the unsupervised object level segmentation problem by elaborately constructing feature vectors for an image from low level cues resulted from the state-of-the-art techniques which can reflect high level structures to some extent. The feature vectors consist of spectral attributes, global color and spatial information. Then, the Potts model is formulated in terms of the feature vectors for segmentation. A heuristic approach is proposed based on the stability of Ncut to select the number of segments which roughly correspond to the object/object parts

with considerable sizes. As a result, a new algorithm is developed, which can automatically segment a natural image into a small number of regions that are locally coherent, respect global structures, have smooth contours snapping to salient object boundaries, and correspond to meaningful objects. Experiments demonstrate that the proposed algorithm can achieve object segmentation to some extent.

- We propose to jointly localize, recognize, segment the objects from a user photo stream by formulating the multiple foreground recognition and cosegmentation (MFRC) problem within a conditional random fields (CRFs) framework in a principled manner for the first time. We capitalize centrally on the key objective that MFRC is to segment out and annotate the foreground objects or the “things” rather than the “stuff”. To this end, we exploit a few complementary objectness cues (e.g. contours, object detectors and layout) and propose novel and efficient methods to capture object-level information. Integrating object potentials as soft constraints (e.g. robust higher-order potentials defined over detected object regions) with low-level unary and pairwise terms holistically, we solve the MFRC task with a probabilistic CRF model. The inference for such a CRF model is performed efficiently with the graph cut based move making algorithms. With a minimal amount of user annotations on just a few example photos, the proposed approach produces spatially coherent, boundary-aligned segmentation results with correct and consistent object labeling. Experiments on the FlickrMFC dataset justify that our method achieves state-of-the-art performance.
- We extend the MFRC framework to the more challenging problem of multiple human identification and cosegmentation (MHIC), we propose to solve the problem with a principled CRF framework using a few weakly labeled images. A novel human instance detector is proposed by combining an extended multiple color line model [WSTS07] with the poselet-based human part detector [BMBM10]. We also propose a more robust unary potential based on an adaptive cross map structure. Moreover, an effective high-level human shape cue is proposed by applying in sequence the geodesic distance transform [CSRP10] and the joint bilateral filtering [ABD10] to the initial human instance detection response map and can be generalized to the other objects. The experiments show state-of-the-art performance on our proposed challenging NTU-MHIC dataset.

The rest of the thesis is organized as follows. Chapter 2 provides a brief literature review of the-state-of-art image segmentation methods and related solvers. Their advantages and disadvantages will be discussed. Chapter 3 gives a detailed introduction to the work “Semantic Segmentation Using Low Level Cues” which can segment an image into regions correspond to object or parts using global information accumulated from low-level cues without any learning process. Chapter 4 introduces the work “Multiple

Foreground Recognition and Cosegmentation” which can extract irregularly occurred objects from the user photo stream by using a principled random field which jointly reason with the novel proposed multi-class colorline object detectors, mid-level superpixel and contour cues and low level pixel cues. Chapter 5 extends the work in Chapter 4 to the more challenging multiple foreground human group identification and cosegmentation tasks. Finally, we conclude in Chapter 6, where the directions of future research will be discussed.

Chapter 2

Literature Review

In this chapter, we give a review for different kinds of image segmentation methods, which are classified according to the level of supervision involved. In the first place, we introduce the popular unsupervised bottom-up image segmentation methods. Then we discuss the conditional random field framework, which has been widely used in the following sections for interactive methods, top-down methods and multi-image based methods.

2.1 Bottom-Up Image Segmentation

The goal of bottom up image segmentation is usually defined to group nearby pixels which are homogeneous in the feature space e.g. color, texture or curvature by clustering those features based on fitting mixture models, mode shifting [CM02] or graph partitioning [FH04] [SM00] [VS91] [Bea06]. Moreover, the variational [CV01] and level set [OS88] techniques have also been used in segmenting images into sub-regions. The grouping criterion does not take into account the explicit notion of the object. This type of segmentation is often referred to as over-segmentation, and has proven to be very useful to either simplify image representation or to help discover the “objects” [CS07, LG10]. Below gives a brief summary of some recent popular methods due to their combinations of the reasonable performance and the public implementation:

- **Mixture of Gaussians:** This method and K-means are similar to each other. Given k initial centers which can be randomly selected, K-means first assigns each sample to one of the centers based on their feature space distance, and then the centers are updated. These two steps iterate until the termination condition is met. Within the mixture of gaussians, each cluster center is modeled by a covariance matrix. Assume a set of d -dimensional feature vector $x_1, x_2, \dots, x_n$ which are drawn from a Gaussian mixtures:

$$p(x|\{\pi_k, \mu_k, \sigma_k\}) = \sum_k \pi_k N(x|\mu_k, \Sigma_k) \quad (2.1)$$

where π_k is the mixing weight, μ_k, Σ_k are the means and covariance matrix, and

$$N(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right\} \quad (2.2)$$

is the normal equation [Sze11]. The parameters π_k, μ_k, Σ_k can be estimated by using the expectation maximization (EM) algorithm. The algorithm will iterate between two steps [Sze11]:

- (i) The *E step* estimates how likely a sample x_i was generated from the k th Gaussian clusters with current parameters

$$z_{ik} = \frac{1}{Z} N(x_i|\mu_k, \Sigma_k) \quad (2.3)$$

with $\sum_k z_{ik} = 1$ and $Z = \sum_k N(x_i|\mu_k, \sigma_k)$

- (ii) The *M step* will update the parameters:

$$\mu_k = \frac{1}{N_k} \sum_i z_{ik} x_i, \Sigma_k = \frac{1}{N_k} \sum_i z_{ik} (x_i - \mu_k)(x_i - \mu_k)^T, \pi_k = \frac{N_k}{N} \quad (2.4)$$

After the parameters are estimated, the segmentation can be formed by assigning the pixels to the most probable cluster.

- **Mean Shift:** Unlike the parametric methods such as K-means and Mixture of Gaussian which have assumptions over the cluster number and feature distributions, Mean-shift [CM02] is a non-parametric method which can automatically decide the cluster number and modes in the feature space.

Assume a data point x are drawn from some probability function, whose density can be estimated by convolving the data with a fixed kernel of width h :

$$f(x) = \sum_i K(x - x_i) = \sum_i k\left(\frac{\|x - x_i\|^2}{h^2}\right) \quad (2.5)$$

where x_i is a near-by sample and $k(\cdot)$ is the kernel function [Sze11]. After the density function is estimated, mean shift uses a multiple restart gradient descent method which starts at some initial guess y_k , then the gradient direction of $f(x)$ is estimated at y_k and uphill step is taken in the direction [Sze11]. Particularly, the gradient of $f(x)$ is given by

$$\begin{aligned} \nabla f(x) &= \sum_i (x_i - x) G(x - x_i) \\ &= \sum_i (x_i - x) g\left(\frac{\|x - x_i\|^2}{h^2}\right) \end{aligned} \quad (2.6)$$

where

$$g(r) = -k'(r) \quad (2.7)$$

and $k'(\cdot)$ is the first-order derivative of $k(\cdot)$. $\nabla f(x)$ can be re-written as

$$\nabla f(x) = \left[\sum_i G(x - x_i) \right] m(x) \quad (2.8)$$

where the vector

$$m(x) = \frac{\sum_i x_i G(x - x_i)}{\sum_i G(x - x_i)} - x \quad (2.9)$$

is called the mean shift vector

value of x . During the mean-shift procedure, the current mode y_k is replaced by its locally weighted mean,

$$y_{k+1} = y_k + m(y_k) \quad (2.10)$$

Final segmentation is formed by grouping pixels whose converge points are closer than h_s in the spatial domain and h_r in the range domain, and these two parameters are tuned according to the requirements of different applications.

- **Normalized Cut:** Many aforementioned methods form the segmentation with local image statistics, and thus such methods can produce trivial regions because the low-level features are sensitive to lighting and perspective changes. Normalized Cut [SM00] finds a segmentation to split the affinity graph which encodes global image information by minimizing the *Ncut* value between different clusters:

$$Ncut(S_1, S_2, \dots, S_k) := \frac{1}{2} \sum_{i=1}^k \frac{W(S_i, \bar{S}_i)}{vol(S_i)} \quad (2.11)$$

where $S_1, S_2, \dots, S_k$ form a k -partition of a graph, $\bar{S}_i$ is the complement of S_i , $W(S_i, \bar{S}_i)$ is the sum of boundary edge weights of S_i , and $vol(S_i)$ is the sum of the weights of all edges attached to vertices in S_i . The basic idea here is that big clusters have large $vol(S_i)$ and minimizing *Ncut* encourages all $vol(S_i)$ to be about the same, thus achieving a “balanced” clustering.

Finding the normalized cut is an NP-hard problem. Usually, an approximate solution is sought by finding eigenvectors of the generalized eigenvalue system $(D - W)v = \lambda Dv$, where $W = [w_{ij}]$ is the affinity matrix of an image graph with w_{ij} describing the pairwise affinity of two pixels and $D = [d_{ij}]$ be a diagonal matrix with $d_{ii} = \sum_j w_{ij}$. In the seminal work of Shi and Malik [SM97, SM00],

the pairwise affinity is chosen as the Gaussian kernel of the spatial and feature difference for pixels within a radius $\|x_i - x_j\| < r$:

$$W_{ij} = \exp\left(-\frac{\|F_i - F_j\|^2}{\sigma_F^2} - \frac{\|x_i - x_j\|^2}{\sigma_s^2}\right) \quad (2.12)$$

where F is a feature vector which consists of intensity, color and gabor features, σ_F , x_i is the spatial coordinate of pixel i , and σ_s are the variance of the feature and spatial position, respectively. In the later work of Malik et al. [MBSL99], they define a new affinity matrix using an intervening contour method. They measure the difference between two pixel i and j by inspecting the probability of an obvious edges along the line connecting these two pixels.

$$W_{ij} = \exp\left(-\max_{p \in \bar{ij}} mPb(p)/\sigma\right) \quad (2.13)$$

where $\bar{ij}$ is the line segment connecting i and j and σ is a constant, the $mPb(p)$ is the boundary strength defined at pixel p by maximizing the oriented contour signal $mPb(p, \theta)$ at different orientation θ :

$$mPb(p) = \max_{\theta} mPb(p, \theta) \quad (2.14)$$

The oriented contour signal $mPb(p, \theta)$ is defined as a linear combination of multiple local cues at orientation θ :

$$mPb(p, \theta) = \sum_s \sum_i \alpha_{i,s} G_{i,\sigma(i,s)}(p, \theta) \quad (2.15)$$

where $G_{i,\sigma(i,s)}(p, \theta)$ measures the χ^2 distance at feature channel i (brightness, color a, color b, texture) between the histograms of the two halves of a disc of radius $\sigma(i, s)$ divided at angle θ , and $\alpha_{i,s}$ is the combination weight by gradient ascent on the F-measure using the training images and corresponding ground truth.

The segmentation is achieved by recursively bi-partitioning the graph using the first nonzero eigenvalue's eigenvector [SM00] or spectral clustering of a set of eigenvectors [NJW01]. For the computational efficiency purpose, spectral clustering requires the affinity matrix to be sparse which limits its applications. Recent work of Cour et al. [CBS05] solves this limitation by defining the affinity matrix at multiple scales and then setting up cross-scale constraints which achieved better results. In addition, Arbelaez et al. [AMFM09] convolve the eigenvectors with Gaussian directional derivatives at multiple orientations θ to obtain oriented spectral contours responses at each pixel p :

$$sPb(p, \theta) = \sum_{k=1}^n \frac{1}{\sqrt{\lambda_k}} \cdot \nabla_{\theta} \mathbf{v}_k(p) \quad (2.16)$$

Since the signal mPb and sPb carries different contour information, Arbelaez et al. [AMFM09] proposed to combine them to globalize the contour information:

$$gPb(p, \theta) = \beta \cdot mPb(p, \theta) + \gamma \cdot sPb(p, \theta) \quad (2.17)$$

where the combination weights β and γ are also learned by gradient ascent on the F-measure using ground truth, later the gPb are used to generate a nested segmentation.

- **Contour Based Region Merging:** Such methods [BF70] [JTLB04] start from pixels or super-pixels, then two adjacent regions are merged based on the metrics which can reflect their similarities/dissimilarities such as edge strength or the color difference. Recently, Arbelaez et al. [AMFM11] proposed the gPb-OWT-UCM method to transform a set of contours into a nested partition of the image. The method first combines the local oriented edge signal with the spectral edge signal to form a probability edge map gPb which delineates the salient contour between objects, as formulated in Eq. 2.17. Then, it performs watershed over the topological space defined by the gPb to form the finest level segmentation. Finally, the edges between regions are sorted and merged in ascending order which forms the ultrametric contour map UCM . Thresholding UCM at a scale λ forms the final segmentation.
- **Graph Based Region Merging:** Unlike aforementioned methods which use *fixed* merging rules, Felzenszwalb and Huttenlocher [FH04] advocates a method which uses a *relative* dissimilar measure to produce segmentation. The method maps an image to a graph which is equivalent to a 8-neighbor Markov Random Field. The pixels denote the nodes, while the edge weights reflect the color dissimilarity between the nodes. Initially each node forms its own component. The internal difference $Int(C)$ is defined as the largest weight in the minimum spanning tree of a component C . Then the weight is sorted in ascending order. Two regions C_1 and C_2 are merged if the between-edge weight is less than $\min(Int(C_1) + \tau(C_1), Int(C_2) + \tau(C_2))$ where $\tau(C) = k/|C|$, and k is a coefficient that is used to control component size. Merging stops when the difference between components exceeds the internal difference.

2.2 Conditional Random Field

Before we go to the following sections, we first introduce the conditional random field framework, as many problems in image segmentation/cosegmentation can be formulated as finding the most probable labels problem in Markov or Conditional Random Field.

Random field provides a principled framework to model the complex interactions between hidden variables in an elegant and precise way. The power of such representation lies in the fact that the probability distribution over different labeling of the random variables can be factorized, which thus allows for efficient inference. The conditional random field has been applied in many methods to be introduced in Sec. 2.3, 2.4 and 2.5 and in Sec. 3, 4 and 5.

Consider a discrete random field $\mathcal{X}$ defined over a 1-D lattice $\mathcal{V} = \{1, 2, \dots, n\}$, where each lattice point $i \in \mathcal{V}$ will be associated with a random variable $X_i \in \mathcal{X}$ and takes a discrete value from the label set $\mathcal{L} = \{l_0, l_1, \dots, l_k\}$ [KLT09]. Each random variable X_i is also associated with its own neighborhood system $N_i, \forall i \in \mathcal{V}$ where N_i denotes the set of all adjacent variable of X_i according to certain criterion, e.g. Euclidean distance. A clique $\mathbf{X}_c$ is a set of random variables which are conditionally dependent on each other [KLT09]. Any possible labels assigned to the random variables is called a labeling, which is denoted by $\mathbf{x}$ and takes values from the set $\mathbf{L} = \mathcal{L}^n$

Two properties are needed to be satisfied for a MRF with respect to a neighborhood system $\mathcal{N} = \{N_v | v \in \mathcal{V}\}$: 1) the positivity property: $Pr(x) > 0 \forall x \in \mathcal{X}^n$ and 2) the Markovian property, i.e. [KT07a]:

$$Pr(x_v | x_u : u \in \mathcal{V} - \{v\}) = Pr(x_v | x_u : u \in N_v) \quad \forall v \in \mathcal{V} \quad (2.18)$$

A conditional random field (CRF) can be treated as an MRF globally conditioned on the observed data $\mathbf{D}$. The conditional distribution $Pr(\mathbf{x} | \mathbf{D})$ over the labelings of the CRF is a Gibbs distribution [KLT09], which can be written as:

$$Pr(\mathbf{x} | \mathbf{D}) = \frac{1}{Z} \exp\left(-\sum_{c \in C} \psi_c(\mathbf{x}_c)\right) \quad (2.19)$$

where Z is a normalization constant known as the partition function, which ensures the distribution falls within the range of $[0 \sim 1]$, and C is the set of all cliques. $\psi_c(\mathbf{x}_c)$ is known as the potential function where $\mathbf{x}_c = x_i, i \in c$ is the vector encoding the labeling of the variables constituting the clique. The corresponding Gibbs energy $E(x)$ is the negative log likelihood of the Gibbs distribution, where we can drop $-\log Z$ as it is a constant:

$$E(x) = -\log Pr(\mathbf{x} | \mathbf{D}) - \log Z \simeq \sum_{c \in C} \psi_c(\mathbf{x}_c) \quad (2.20)$$

The most probable or maximum a posteriori (MAP) labeling $\mathbf{x}^*$ of the random field is defined as [KLT09]:

$$\mathbf{x}^* = \operatorname{argmax}_{\mathbf{x} \in \mathbf{L}} Pr(\mathbf{x} | \mathbf{D}) \quad (2.21)$$

and can be found by minimizing the energy function E .

Potts model corresponds to a special case of such Markov Random Field over the image graph, which computes the graph partition with minimal total perimeters of the one-label region and doesn't favor any particular label orders. Potts model consists of a data term and a regularization term. The data term enforces the data to be similar to some physical observation. The regularization term requires the solution to be regular. One reason for such an interpretation is that the solution which lives up to the observation can be abundant and close to random, and thus the algorithm needs some criteria to justify which one is the best. Many data in real world exhibit some regular properties. In terms of image, it often means the solution should be spatially smooth and the boundary between two segments should be compact instead of jaggy.

To minimize the function/functional, either discrete combinatorial or variational optimization methods can be used. Discrete models treat the image as a discrete graph and define a variable for each node on the graph which takes a value from the finite set. On the other hand, continuous counterparts define the function over continuous domain which takes a continuous value. The methods to optimize the function/functional are also varied. Discrete methods often use combinatorial optimization methods to minimize the function. Among them, the fast min-cut and max-flow method proposed by Boykov and Kolmogorov [BK04] has been widely applied. On the other hand, continuous methods often use convex relaxation methods [PSG⁺08, YBT10] to solve the Euler-Lagrange equation which have demonstrated reasonable performance. There are both pros and cons for using discrete and continuous methods. Discrete methods are more numerical stable, however they might produce visual artifact as the grid might bias towards certain direction. Continuous methods have attracted more attention in these years, as they can better describe image geometry information (boundary length, curvature....), produce smooth and accurate boundaries and can be parallelized by GPU. On the other hand, incorporating higher order potentials in continuous models can lead to difficult to optimize functional, which is not as easy to solve as in discrete methods.

In the following parts, we will introduce both the continuous and discrete version of the Potts model correspondingly and the state-of-the-art solvers for the problem.

2.2.1 Continuous Convex Relaxed Potts Model

Given an image $I : \Omega \rightarrow R$, the continuous Potts model attempts to partition the image into n disjoint sub-regions $\{\Omega_i\}_{i=1}^n$ s.t. $\bigcup_{i=1}^n \Omega_i = \Omega, \Omega_k \cap \Omega_l = \emptyset, \forall k \neq l$ by minimizing the functional:

$$\min_{\{\Omega_i\}_{i=1}^n} \sum_{i=1}^n \int_{\Omega_i} p(l_i, x) dx + \alpha \sum_{i=1}^n C_i(x) |\partial\Omega_i|, \quad (2.22)$$

where the first term is the region term which measures the cost to assign label l_i to the data. A simple region term is given by

$$p(l_i, x) = |I(x) - c_i|, i = 1, \dots, n \quad (2.23)$$

where c_i corresponds to the mean intensity of l_i . The second term is the boundary term where $|\partial\Omega_i|$ correspond to the perimeter of each disjoint region Ω_i , and $C_i(x)$ is an edge detector function which is usually defined by

$$C_i(x) = \frac{1}{1 + |\nabla I(x)|^2}. \quad (2.24)$$

The region term and the boundary term are balanced by a tradeoff factor α . Minimizing the region term ensures the segmentation complying with some region coherence and minimizing the boundary term favors the segmentation with tight and smooth boundaries along the salient edges in the image.

By defining an indicator function $u_i(x)$ for each region $\Omega_i, i = 1 \dots n$, i.e.

$$u_i(x) = \begin{cases} 1, & x \in \Omega_i \\ 0, & x \notin \Omega_i \end{cases} \quad (2.25)$$

The Potts model (2.22) can be rewritten as

$$\min_{u_i(x) \in \{0,1\}} \sum_{i=1}^n \int_{\Omega} \{u_i(x)p(l_i, x) + \alpha C_i(x)|\nabla u_i|\} dx, \quad (2.26)$$

$$s.t \sum_{i=1}^n u_i(x) = 1, \forall x \in \Omega \quad (2.27)$$

whereas the second term $\int_{\Omega} |\nabla u_i| dx$ is to approximate the perimeter of each disjoint subdomin Ω_i . This model is nonconvex due to the binary configuration of each $u_i(x) \in \{0, 1\}$ which often causes computation to get stuck in local minimums. Thus, current methods [PCCB09] [BYT11] [LKY⁺09] propose to relax the binary constraint to the interval $[0, 1]$ and approximate (2.26) with the convex relaxed model:

$$\min_{u_i(x) \in [0,1]} \sum_{i=1}^n \int_{\Omega} \{u_i(x)p(l_i, x) + \alpha C_i(x)|\nabla u_i|\} dx, \quad (2.28)$$

$$s.t \sum_{i=1}^n u_i(x) = 1, \forall x \in \Omega \quad (2.29)$$

2.2.1.1 Multi-Label Continuous Max-Flow Method

To solve the convex relaxed Potts model (2.28), Yuan et al. [YBTB10] present a new multi-label continuous max-flow formulation, which is equivalent and dual to the original one.

The method maps the functional to n parallel copies $\Omega_i, i = 1 \dots n$ of the image domain Ω , which are linked to a common source node s and sink node t , where n is the number of labels. Then the minimization problem is transformed to the problem of finding a set $p_s, p = [p_1, p_2, \dots, p_n], q = [q_1, q_2, \dots, q_n], u = [u_1, u_2, \dots, u_n]$ for $\max_{p_s, p, q} \min_u L_c(p_s, p, q, u)$:

$$\begin{aligned} L_c(p_s, p, q, u) = & \int_{\Omega} [p_s + \sum_{i=1}^n \langle u_i, \text{div} q_i - p_s + p_i \rangle] dx \\ & - \frac{c}{2} \sum_{i=1}^n \|\text{div} q_i - p_s + p_i\|^2 \\ \text{s.t. } & |q_i(x)| \leq \alpha C_i(x), i = 1, \dots, n, \forall x \in \Omega \\ & p_i(x) \leq p(l_i, x), i = 1, \dots, n, x \in \Omega \end{aligned}$$

where $c > 0$ is a constant. In this new formulation, for each position $x \in \Omega$, $p_s(x)$ is the flow streams from the source s to x at each copy Ω_i , $p(l_i, x)$ serves as the capacity of the sink flow $p_i(x)$ directed from x at Ω_i to the sink t , whereas $\alpha C_i(x)$ is the capacity of spatial flows $q_i(x)$ defined within each Ω_i and $u_i(x)$ is the indicator function for each label i and works as the Lagrangian multiplier.

An algorithm based on the augmented Lagrangian method is introduced in [YBTB10] to find the solution. Final segmentation is formed by assigning each pixel x to the label i whose corresponding indicator function $u_i(x)$ has the largest magnitude. It has been shown that compared with the previous methods [ZGFN08] [LKY⁺09], the continuous max-flow method has a faster convergence rate, and it can be highly parallelized to achieve even faster processing speed. This algorithm is adopted in our approach in Chapter. 3 to solve the Potts model and detail solver can be found in Pseudocode. 1.

2.2.2 Discrete Potts Model

Similar to the continuous version, the discrete Potts model also have two terms defined on a 4-connected or 8-connected 2D grid. The objective function to solve is:

$$E(X) = \lambda \sum_{\{i,j\} \in N} \delta(x_i \neq x_j) E_1(x_i, x_j) + \sum_{i \in P} E_2(x_i) \quad (2.34)$$

where P denotes the pixels in an image, N denotes all the unordered pairs of $\{p, q\}$ of 4-connected neighboring pixels of P . $X = \{x_1, x_2, \dots, x_{|P|}\}$ is the indicator vector

Protocol 1 Multiplier-Based Maximal Flow Algorithm [BYTB14]

Choose some starting values for p^1 , q^1 and λ^1 , let $k = 1$ and start k -th iteration, which includes the following steps, until convergence:

- Optimize $q_i, i = 1, \dots, n - 1$, by fixing other variables

$$q_i^{k+1} := \operatorname{argmax}_{\|q\|_\infty \leq \alpha C(x)} - \frac{c}{2} \| \operatorname{div} q_i(x) + p_{i+1}^k - p_i^k - u_i^k/c \|^2 \quad (2.30)$$

This equation can be solved by Chambolle's projection algorithm [Cha04]

- Optimize $p_i, i = 1 \dots n$, by fixing other variables

$$p_i^{k+1} := \operatorname{argmax}_{p_i(x) \leq p(l_i, x)} - \frac{c}{2} \| p_i + \operatorname{div} q_i^{k+1} - p_s^k - u_i^k/c \|^2 \quad (2.31)$$

which can be computed at each $x \in \Omega$ in a closed form;

- Optimize p_s by fixing other variables:

$$p_s^{k+1} := \operatorname{argmax}_{p_s} \int_{\Omega} p_s dx - \frac{c}{2} \sum_{i=1}^n \| p_s - (p_i^{k+1} + \operatorname{div} q_i^{k+1}) + u_i^k/c \|^2 \quad (2.32)$$

which can be done by computing the maximum of the quadratic equation at $x \in \Omega$ pointwise

- Update multiplier $u_i, i = 1, \dots, n - 1$ by

$$u_i^{k+1} = u_i^k - c(\operatorname{div} q_i^{k+1} - p_s^{k+1} + p_i^{k+1}); \quad (2.33)$$

- Let $k = k + 1$ go to the $k + 1$ the iteration until convergence
-

whose components indicate pixels' assignment to the observed finite set of label $\mathcal{L} = \{l_1, l_2, \dots, l_n\}$

The $E_1(x_i, x_j) = V^{pq}(x_i, x_j)$ is the smoothness term which encourages neighboring pixels to share similar labels and is defined as follows:

$$E_1(x_i, x_j) = \exp \left(- \frac{\|v_i - v_j\|^2}{\sigma^2} \right) \quad (2.35)$$

where σ is the variance of all the pair wise differences. v_i and v_j represent the RGB values of pixels i and j .

The second term $E_2(x_i)$ is the data/region term which encodes the likelihood that each pixel x_i belongs to a certain label $l_k, k = 1, 2, \dots, n$ similar to the one in Eq. 2.23. The most commonly used data term is generated from the color Gaussian mixture model which can be learned from the user interaction [BJ01] or labeled databases [SJC08].

2.2.2.1 Approximate Optimization for Discrete Potts Model

Different function V^{pq} can lead to different optimization algorithms. When the set of feasible labels are binary (such that $x_i \in \{0, 1\}, \forall i \in V$), if V^{pq} satisfies the submodular condition i.e.:

$$V^{pq}(1, 1) + V^{pq}(0, 0) \leq V^{pq}(1, 0) + V^{pq}(0, 1), \forall (p, q) \in \mathcal{N} \quad (2.36)$$

which is similar to the convex condition in continuous model, then (2.34) can be optimized exactly by min-cut/max-flow methods. However, if V^{pq} is non-submodular, (2.34) is NP-Hard. Specifically, this is the case for the Pott's interaction potential. For non-submodular case, only approximate solution exists, which can be produced by several methods, e.g. iterated conditional modes [Bes86], simulated annealing [GG84], message passing [Kol06] and linear programming [KT07b]. Most popular are the graph cut based α expansion and $\alpha - \beta$ swap [BVZ01], which are widely considered state-of-the-art for minimizing such energy function that can cause large energy change.

The α expansion and $\alpha - \beta$ swap can only be applied under some assumptions on V^{pq} . The interaction potential $V^{pq}(\cdot, \cdot)$ is said to be semi-metric if for any $\alpha, \beta \in \mathcal{L}$ and all $(p, q) \in \mathcal{N}$, $V^{pq}(\alpha, \beta) = V^{pq}(\beta, \alpha) \geq 0$ and $V^{pq}(\alpha, \beta) = 0 \Leftrightarrow \alpha = \beta$. If $V^{pq}(\cdot, \cdot)$ further satisfies the triangle inequality, i.e. for any $\alpha, \beta, \gamma \in \mathcal{L}$, $V^{pq}(\alpha, \beta) + V^{pq}(\beta, \gamma) \leq V^{pq}(\alpha, \gamma)$, then $V^{pq}(\cdot, \cdot)$ is said to be a metric [BVZ01].

Protocol 2 α - β swap algorithm

- 1: Select initial labeling function u
 - 2: Repeat until $w_p = v_p, \forall p \in P$
 - 3: **for** each pair $\alpha, \beta \in \mathcal{L}$ **do**
 - 4: $u \leftarrow \arg_{\min} E(u)$, s.t. u is one $\alpha - \beta$ swap move from v
 - 5: set $v \leftarrow u$
 - 6: **end for**
 - 7: set $w \leftarrow v$
 - 8: Output labeling function w .
-

For any labeling function $u : P \rightarrow \mathcal{L}$, defines the set $P_i(u) = \{p \in P, u_p = l_i\}, i = 1, \dots, n$. Given a pair of labels $\alpha, \beta \in \mathcal{L}$, a move from a labeling function u to a labeling function u' is called an $\alpha - \beta$ swap if $P_i(u) = P_i(u')$ for all $l_i \in \mathcal{L} \setminus (\alpha \cup \beta)$, which indicates the only difference between u and u' are some pixels which are labeled as α (or β) change

Protocol 3 α expansion algorithm

- 1: Select initial labeling function u
 - 2: Repeat until $w_p = v_p, \forall p \in P$
 - 3: **for** each $\alpha \in \mathcal{L}$ **do**
 - 4: $u \leftarrow \arg_{\min} E(u)$, s.t. u is one α expansion move from v
 - 5: set $v \leftarrow u$
 - 6: **end for**
 - 7: set $w \leftarrow v$
 - 8: Output labeling function w .
-

to β (or α), respectively. For a label value $\alpha \in \mathcal{L}$ a move from u to u' is called an α expansion if $P_\alpha(u) \in P_\alpha(u')$ and $P_{l_i}(u') \subseteq P_{l_i}(u)$ for all $l_i \in \mathcal{L} \setminus \alpha$, which means some pixels which are not labeled as α in u is allowed to change to α in u' [KZ04].

The essential difference between α expansion and $\alpha - \beta$ swap lies in their optimization structures when different metrics are applied to the energy functions of the form (2.34). For any $\alpha, \beta \in \mathcal{L}$ and any u , the optimal labeling function v which is within one $\alpha - \beta$ swap from u and minimizes the energy (2.34) can be computed by binary graph cut, provided $V^{pq}(\cdot, \cdot)$ is a semi-metric. Similarly, for any $\alpha \in \mathcal{P}$ and any labeling function u , the optimal labeling function v which is within one α expansion from u and minimizes the energy (2.34) can be computed by binary graph cuts, provided $V^{pq}(\cdot, \cdot)$ is a metric. The $\alpha - \beta$ swap and α expansion algorithms start with an initial labeling function u and iteratively compute the $\alpha - \beta$ moves and the α expansion moves respectively, until the energy does not change. The algorithms are illustrated in Pseudocode 2 and 3.

2.3 Interactive Segmentation

After we introduced the conditional random field framework, then we introduce the interactive methods. Given absence of an absolute criterion for segmentation, user input can be highly desirable to resolve the ambiguities by specifying what he wants to segment and provide an answer based on some pre-defined rules. Some implicit assumptions are made about the object-of-interest and the user input: most methods assume the user input is totally correct and the object is often assumed to be “thing” which is distinct from the rest of the image. Many frameworks have been proposed to allow the user to give some partial information about the regions he want to segment either by drawing bounding boxes [RKB04] or scribbles [YCZL10, ACZZ12] on the major parts of the objects, then graph cut (Sec. 2.2.2.1), random walk [Gra06] or variational model (Sec. 2.2.1) are used to solve the corresponding optimization problem. Recently, Subr et al. [SPSK13] breaks the assumption that user input is totally correct and allow their method to handle inaccurate input by using fully connected CRF.

Fig 2.1 shows examples of the interactive segmentation. In practise, interactive segmentation works very well and has been implemented in commercial products, such as Microsoft Office and Adobe Photoshop.

Figure 2.1: Examples of interactive segmentation by using bounding box (a ~ b) [RKB04] or scribbles (c ~ d) [YCZL10]

2.4 Top-Down Object Segmentation

Top-down image segmentation has been proposed to solve the limitations of the bottom-up approach by leveraging the higher level concept of the “object” or the “thing” to separate the object regions from the “stuff”. One underlying reason for this definition is that the “things” tend to have clear size and shape (e.g. pedestrian, cars), as opposed to the “stuffs” which tend to be homogeneous or recurring pattern of fine scale structures. In this case, segmentation is, instead to separate the whole image into some meaningful regions, but to produce a pool of regions that correspond to the “objects” and leave away the “stuff” region. Therefore, how to make a distinction between the “stuffs” and the “things” is vital for the top-down segmentation.

One approach to incorporate the object notion is through the use of object detectors for certain classes. Such object detectors can be any bounding box detectors, e.g. famous Deformable Part Model (DPM) [FGM10] or Poselets [BMBM10]. A few works have been proposed to combine object detection and segmentation. Larlus and Jurie [LJ08] obtain the object segmentation by refining the bounding box using CRF. Gu et al. [GLAM09] has proposed to use hierarchical region trees for object detection, instead of the bounding boxes. However, these approaches can only be applied to a limited number of object classes, which is restrictive when applied to a large number of object classes, e.g. ImageNet with thousands of classes.

The other approaches consider the problem in an object class agnostic way to produce segmentation [CS12] [EH14] [ADF10]. Carreira et al. [CS12] and Endres et al. [EH14] proposed to generate region proposals by using uniform spatial grid or contour guided grid to act as foreground seeds and image border as the background seed followed by

Figure 2.2: Example of the object class agnostic segmentation [CS12].

graph cut (Sec. 2.2.2.1) to extract the region, and then using a learned model with different region and boundary features to give a score to each region according to its likelihood to be an object. In certain sense, this can be treated as a kind of salient object discovery problem, but with learned features. Such object agnostic segmentation has been applied in some recent video analysis [LKG11] or cosegmentation tasks [MLLN12]. On the other hand, the distinction between stuff and things can still be ambiguous which makes any “objectness” disputable. Fig 2.3 illustrates some interesting ambiguities: In Fig 2.3.(a), the trees in the middle can be considered as the objects given they are more distinctive. However, when considering with the surrounding forest, they can be treated as the “stuff”. In the Fig 2.3.(b), the cars may be considered as the “stuff” despite having “specific size and shape” as the pattern is quite repetitive.

2.5 Multiple Image Segmentation

Another way for semantic image segmentation is to learn a segmentation model based on the information associated with multiple images. Although also depending on the given information, the higher level representations learned based on the given set of images makes the essential difference compared with interactive segmentation. The learned models can then be used to predict the similar regions in new images. This type of approach is also different from the top-down approach in the sense that it aims to parse the image as a whole into the “thing” and “stuff” classes, instead of just producing possible “thing” candidates. According to the granularity of the information provided, it can be roughly divided into three subcategories: supervised methods, weakly supervised methods and cosegmentation methods.

Figure 2.3: Examples to illustrate the ambiguity of the “object” notion, which can vary according to scale, context and repetitiveness.

Figure 2.4: Examples of semantic segmentation [KK11] which is trained by using pictures with human labeled ground truth similar to (b) to segment the test image in (c) and produce the final segmentation in (d).

In supervised segmentation [KLT09] [GFK09] [LYT11] [TL13] [KK11] [LRKT10], a segmentation model is learned by using images with human labeled pixelwise ground truth. These regions are often associated with a particular visual category. An example is given in Fig 2.4. This type of approach gives good results in practice but it requires ground truth for each category to learn, which is quite time-consuming and expensive. It also makes the assumption that any given region must belong to the predefined category in the training set, and can not handle the categories those do not belong to this set.

On the other hand, weakly-supervised segmentation aims to achieve the same goal as the supervised segmentation but with a weaker form of information [VFB11]. In this context, training images are associated with some tags, but hand-labelled data is typically not available in the image. The goal is thus to simultaneously segment the images into regions representing the tags and learn a model.

Figure 2.5: Simultaneously segmenting one common foreground object from a set of image (a) is a well studied problem [CLL11]. Recently, Kim et al. [KXLK11] considers multi-class cosegmentation problem (b).

Finally, cosegmentation [KX12,KXLK11,JBP12,RSLP12,JBP10,CLL11,MSP11,MSD09] aims to simultaneously divide a set of images into $K \geq 2$ different object classes. It can be treated as a special case of weakly supervised segmentation. On the other hand, unlike weakly supervised segmentation, there is no explicit labeled images or pre-trained model in cosegmentation. Therefore, cosegmentation aims to learn a new model from a set of given images implicitly, which makes the method applicable to more flexible settings. Indeed, one would expect that cosegmentation can play a key role in effective automated object discovery techniques and part based approaches to object detection [LG10]. For this reason, this field has been quite active recently: early works focus on a pair of images with the same object instance but with different background [RMBK06] and optimize by graph cut (Sec. 2.2.2.1). Several works have extended the original model to multiple

Figure 2.6: Given a user’s photo stream about certain event, which consists of finite objects e.g. reg-girl, blue-girl, blue-baby and apple basket. Each image contain an unknown subset of them, which we called “Multiple Foreground Cosegmentation” problem. Kim [KX12] proposed the first method to extract these irregularly occurred objects from the photo stream (b) by using few bounding box provided by user in (a).

image setting [CLL11] [BKP⁺10] [MSP11]. More recent works explicitly aim at handling multiple object classes and images [KX12]. However, these methods make strong assumption that all images contain the objects of similar appearance, which is quite restrictive. A more common scenario in user photo stream is that each image contains a subset of the objects, which may not necessarily appear in all images. An example is illustrated in Fig 2.6. The album contains four common objects: girl in blue, girl in red, baby in blue and an apple basket. However, each image typically contains only a subset of them. To apply classic cosegmentation, users need to manually sort out the images into subsets so that each image contain the common object of interest, which is cumbersome. Recently, some works [KX12, ML13, ZLC⁺14] begin direct modeling this multiple foreground scenario explicitly and have achieved the state-of-the-art performance.

Chapter 3

Object Segmentation using Low Level Cues

Fast growing of web images require them to be transformed into more structured representations before they can be further exploited. This chapter investigates the problem of automatically segmenting an image into a small number of regions. Different from the conventional unsupervised segmentation relies on low-level representations which often leads to severe over-segmentation, our proposed method targets to produce a small set of regions that have a relatively large size and correspond to objects or parts conveying some semantics or high-level structure, in addition to certain homogeneity. The motivation for this target is that most images can be viewed as combinations of a few salient objects or their parts. Such segmentation can be beneficial for many computer vision tasks such as object detection and scene understanding, as demonstrated in [CBGM02] where images are transformed into the Blobworld representation composed of a small set of image regions and these regions are then used for image retrieval and querying. We hence call our process object-level segmentation, though this chapter does not discuss category-specific labeling that is often involved in the conventional semantic segmentation with learned representation and structure prediction models.

Developing an automatic algorithm that well mimics human brain which is good at abstracting semantically meaningful regions from visual cues is still very challenging. To perceive an image, elements must be perceived as a whole, but most images are currently represented based on local low-level features, e.g color, texture. In addition, the concepts of “objects” and human perception are quite subjective and content dependent. In general, automatically generating semantic segmentation is an ill-posed problem. It often requires global image information or high level knowledge, which may come from user input in interactive methods [YCZL10] or labeled database in learning/training based methods [SJC08]. However, we observe that high level knowledge and low level cues are not totally independent and actually some semantics are conveyed in various low level cues. For example, cognition study [HS97] shows that human vision views part boundaries

at those with negative minima of curvature and the part salience depends on three factors: the relative size, the boundary strength and the degree of protrusion. Gestalt theory and other psychological studies have also developed various principles reflecting human perception, which include: (1) human tends to group elements which have similarities in color, shape or other properties; (2) human favors linking contours whenever the elements of the pattern establish an implied direction. Therefore it is interesting to explore what low level cues can produce for high level semantic segmentation.

3.1 Related Work

This section gives a brief review of extensive research has been conducted for unsupervised image segmentation. One broad family of the methods makes use of local features such as color and texture for clustering. Examples are MeanShift [CM02] and Graph-Based Region Merging [FH04]. While these methods are usually very fast, they tend to produce over-segmentation. Alternatively, methods like spectral clustering [NJW01] and Normalized Cut [SM00] use eigenvectors for the clustering process. It has been shown that eigenvectors resulting from spectral clustering carry global contour information and thus these methods are able to capture semantic regions with considerable sizes. The gPb-owt-ucm method [AMFM11] combines multiple local cues into a globalization process using spectral clustering and then constructs a hierarchical region tree from a set of contours to achieve hierarchical image segmentation, which has demonstrated the state-of-the-art segmentation performance. The segmentation results with these methods are generally good. However, some visual artifacts can still be observed. For example, the region contours do not follow object boundaries very well and the large uniform or smooth regions may be split. On the other hand, the multi-label segmentation can also be formulated as a variational problem [BVZ01, MS89, PSG⁺08, BYTB14, YBTB10]. Variational methods have become popular since they can produce smooth and accurate region boundaries and many fast numerical solvers have been developed. In addition, many of these solvers can be parallelized, which is very suitable for GPU implementation. However, variational methods are in general sensitive to the initializations and bad initialization can result in local minimum [BYTB14].

3.2 Proposed Method

Given an input image, our goal is to automatically partition the image into a small number of regions that are coherent in color and structure. We divide this problem into two subproblems. The first one is how to segment the image into k regions for a given number k . The second one is how to automatically choose k , the number of regions,

which will be described in Section 3.2.4. Combining the solutions to both subproblems leads to an automatical variational image segmentation algorithm.

For the first subproblem, our basic idea is to construct some feature vectors from various low level cues for images and then formulate the Potts variational model in terms of the feature vectors for segmentation. Basically, this involves the following processes:

- **Global Structure/Boundary Feature Extraction:** Construct eigenvectors and globalized probability of boundary for each pixel which is later used to construct the Potts functional.
- **EM Initialization:** Model the distribution of color and eigenvectors with a k mixture of Gaussians and use the EM method to generate initial means for the Potts model.
- **Functional Formulation and Segmentation:** The global features that consist of global structure, color and spatial information are used to formulate the region and boundary terms of the continuous Potts functional in Sec. 2.2.1 to make the segmentation produce homogeneous regions and snap to accurate boundaries. The final segmentation is completed by solving the variational model using the continuous max-flow method [YBTB10], a detailed introduction can be found in Sec. 2.2.1.1.

These processes are elaborated in the next three sub-sections.

3.2.1 Global Structure and Boundary Feature Extraction

The Potts model works well under the assumption that the image is roughly piecewise constant. However low-level local features such as intensity, color, texture and curvature may not necessarily possess such characteristics, which makes the model sometimes produce trivial regions and false boundaries. Thus it is necessary to consider features that can better describe the underlying data. The eigenvectors resulting from an affinity matrix are such features because they carry global contour information, reflect significant structures and tend to be roughly piecewise constant, which can be observed in Figure 3.1.

To construct the eigenvectors, we basically adapt the work of [AMFM11] that describes a very nice globalization method for contour detection and spectral clustering. The main steps are as follows:

First, a multiscale extension mPb of the posterior probability of boundary at each image pixel x is computed, which considers gradients at different scales for image brightness, color, and texture channels in order to detect both fine and coarse structures. Second, a sparse symmetric affinity matrix $W = [w_{ij}]$ is constructed using the intervening contour

3.1.a: Original image

3.1.b: gPb Map

3.1.c: Eigenvectors

Figure 3.1: (a), (c): An image and its top four non-zero eigenvalue eigenvectors. (b): The gPb contour map generated from eigenvector captures clean and salient image boundaries

cue [MFM04, AMFM11] and the maximal value of mPb along a line connecting two pixels x_i and x_j with $w_{ij} = \exp(-\max_{p \in x_i x_j} \{mPb(p)\} / \rho)$, where $x_i x_j$ is the line segment connecting x_i and x_j and ρ is a constant (which is set to $\rho = 0.1$ in literature). Third, let $D = [d_{ij}]$ be a diagonal matrix with $d_{ii} = \sum_j w_{ij}$ and solve for the generalized eigenvectors of $(D - W)v = \lambda Dv$. We choose l eigenvectors corresponding to the l smallest nonzero eigenvalues. l is determined via an eigen-gap heuristic [vL07]. Finally, a spectral boundary detector sPb is defined at each pixel by the convolutions of each eigenvector with Gaussian directional derivative filters. As pointed out in [AMFM11], signal mPb fires at all the edges and spectral signal sPb extracts only the most salient edges in the image. Thus a linear combination of mPb and sPb is suggested to yield a globalized probability of boundary, gPb . A more detailed formulation of gPb can be found in Sec. 2.1. In our application, from this process we extract l eigenvectors and the gPb map, which will be used in the subsequent steps.

3.2.2 EM Initialization

One drawback of the Potts model is that it is sensitive to the choice of mean $c_i, i = 1, \dots, n$, in (2.23). Inappropriate initialization of c_i can make the model get stuck in local minimum [BYTB14]. Therefore here we propose a simple and efficient method to generate reasonable initial means.

As eigenvectors are nearly piecewise constant, thus for k -partition, we can assume they are drawn from k Gaussians in the mixture model and use the EM algorithm to determine the maximum likelihood parameters of the mixture of k Gaussians in the feature space. For each pixel, we construct a feature vector of length $(l + 3)$ that consists of RGB colors and eigenvectors $v(x) = (R, G, B, e_1, \dots, e_l)$. Then we perform the EM algorithm to estimate the parameters.

The EM algorithm proceeds by iteratively repeating two steps—the “E Step” and the “M step”—until a stopping criterion is reached. In the “E Step”, we estimate values to fill in for the Gaussian clusters to which the points in the feature space belong; in the “M Step”, we compute the maximum-likelihood parameter estimated based on the current clusters. To apply the EM algorithm, we need to initialize parameters. The initial mixing weights π_i are simply set to $\frac{1}{k}$. The covariance matrices Σ_i are set to the identity matrix. For the means $\mu_1, \mu_2, \dots, \mu_k$, we run K-means to generate k clusters and then use the means of these clusters as the initial means. After the EM iteration stops, each pixel is assigned to the label corresponding to the largest probability, thus delivering k initial regions.

Note that Carson et al. [CBGM02] use a similar method to generate blobworld for image retrieval. However the low-level feature (color, Gabor filter and pixel location) can prevent EM from approaching semantic objects. In addition, parameter initialization in their work is done in an ad-hoc manner which can lead to failure detection of the object of interest.

3.2.3 Variational Segmentation

The Potts model contains two terms: a region term (2.23) and a boundary term (2.24), as in Section 2. Below we show how to formulate these two terms to incorporate global information such as the spectral attributes:

3.2.3.1 Region Term Formulation

In [BYTB14], the variational model is defined in the RGB space. When natural images contain much variation of colors, local features such as color and texture are often insufficient to reflect the structure of the images. To generate a good partition, it is desirable to include some global information in the region term.

In our case, the EM clustering has provided a good approximation of k regions, which could be utilized to extract some useful region information for later variational segmentation. We first introduce the GMM to describe the color distribution of each label, which has demonstrated its success in interactive segmentations [YCZL10,RKB04]. Specifically, for each pixel x in the image, we can obtain a set of probabilities $[g_1(x), \dots, g_k(x)]$, where $g_i(x)$ denotes the probability that the pixel belongs to label l_i and is computed by

$$g_i(x) = \frac{-\log Pr(x|\bar{l}_i)}{-\log Pr(x|l_i) - \log Pr(x|\bar{l}_i)} \quad (3.1)$$

where $Pr(x|l_i)$ indicates the probability that pixel x fits the GMM of label l_i and $Pr(x|\bar{l}_i)$ is the probability that x fits the GMM of any label other than l_i .

To strengthen the region information when foreground and background colors are not well separable, we further introduce the geodesic probability to describe the spatial information of the seed regions. The geodesic probability indicates the likelihood of pixel x belonging to label l_i and is defined by [BS07]

$$\epsilon_i(x) = \frac{D(x, \bar{l}_i)}{D(x, l_i) + D(x, \bar{l}_i)} \quad (3.2)$$

where $D(x, l_i)$ is the geodesic distance from pixel x to the seed region of label l_i and $D(x, \bar{l}_i)$ is the geodesic distance from x to other seed regions.

Now we are ready to construct a feature vector for each pixel x . The feature vector is a $(2k + l)$ vector:

$$[g_1(x), \dots, g_k(x), \epsilon_1(x), \dots, \epsilon_k(x), e_1(x), \dots, e_l(x)] \quad (3.3)$$

where k is the number of labels, l is the number of the selected eigenvectors generated in Section 3.2.1, $g_i(x)$ is the GMM probability of label i defined by (3.1), $\epsilon_i(x)$ is the geodesic probability of label i defined by (3.2), and $e_i(x)$ is the element corresponding to x in the i -th eigenvector. The feature vector can be viewed as a point in a $(2k + l)$ -dimensional space called the feature space. The Potts model will be applied to this space. Thus we define the region function for each label l_i to be

$$p(l_i, x) = \alpha_1^i \cdot [\alpha_2^i \cdot (1 - g_i(x)) + (1 - \alpha_2^i) \cdot (1 - \epsilon_i(x))] + (1 - \alpha_1^i) \cdot d_{eigen}(l_i, x) \quad (3.4)$$

where α_1^i and α_2^i are the tradeoff factors for label l_i ,

$$d_{eigen}(l_i, x) = \sqrt{(e_1(x) - e_1^{l_i})^2 + \dots + (e_l(x) - e_l^{l_i})^2}$$

Figure 3.2: An example to compare the segmentation results by using different features. 3.2.e: Combining global spectral, color and spatial features achieves the best result

and $e_j^{l_i}$ is the mean of spectral attribute e_j for label l_i . The region function (3.4) has three terms: the first two describe the color and spatial information of regions and the third one describes the global structure information. The combination of them enhances the capability of differentiating regions.

It is important to properly set the tradeoff factors α_1^i and α_2^i . When GMMs provide enough information to distinguish one label from the others, the first term should dominate; Otherwise, eigenvectors and the geodesic probability should play a major role. Thus, as suggested in [YCZL10], we set α_1^i and α_2^i to be the Kullback-Leibler divergence between the current label's GMM and the rest labels' GMMs:

$$\alpha_1^i = \alpha_2^i = \frac{1}{n} \sum_{j=1}^n \left| \frac{\log Pr(x_j|l_i) - \log Pr(x_j|\bar{l}_i)}{\log Pr(x_j|l_i) + \log Pr(x_j|\bar{l}_i)} \right| \quad (3.5)$$

where $\bar{l}_i$ indicates the rest labels and n is the number of pixels.

An example which compare different combination of global features is shown in Figure 3.2. We can see that by using GMM, geodesic probability and eigenvectors, the result is more meaningful and accurate than using the local RGB information.

3.2.3.2 Boundary Term Formulation

The boundary term in (2.22) is a weighted total variation of function u . The weight $C_i(x)$ plays an important role. The definition of $C_i(x)$ in (2.24) favors the segmentation along the curves where the edge detection function takes small values. In our algorithm, we use the gPb proposed in [MBSL99, AMFM11] as the base map where the edge detector

Figure 3.3: Comparison of the results of our method using the two different $C_i(x)$ definitions in (2.24) and (3.6), respectively. Some boundary problems due to using (2.24) are circled in (c).

of (2.24) is applied. gPb is computed in the process of generating spectral attributes and it has proved to be powerful signal for edge information. Unlike other classical detectors, gPb makes use of the global information encoded in eigenvectors and thus it can capture the salient edges. However, gPb has limitations in that some weak edges may be missed due to the fact that eigenvectors may not capture small structures. Thus we propose to further incorporate the GMM probability map to enhance the edge detection:

$$C_i(x) = \beta^i \cdot g_c + (1 - \beta^i) \cdot g_e \quad (3.6)$$

where g_c and g_e are the results of applying the edge detector of (2.24) to the GMM probability map and the gPb map respectively, and β^i is a tradeoff factor which is defined in a similar way as α_1^i or α_2^i given in (3.5).

Figure 3.3 compares the results of our method with and without g_c . We can find that by incorporating the GMM probability map, the weak edges are enhanced and the segmentation is better snapped to the salient image boundaries. In addition, the result with the canny edge detector is the worst, as it captures too much trivial edges which make the algorithm snap to unsalient ones.

3.2.4 Selecting The Number Of Regions

We now discuss how to choose k , the number of regions. An ideal value of k should best fit the number of groups present in the image. However, the notion of best fitting is quite subjective. Here we presented a heuristic approach to compute k based on the stability of the $Ncut$ values.

Our observation is that if a good k -partition has been formed, increasing the number of segments to $k + 1$ will cause the existing segments to be split and merged to form a new segmentation, which usually results in a big change of the $Ncut$ values. This suggests a brute-force approach: perform clustering and compare the $Ncut$ values to select among different values of k . Considering that our goal here is to find the number of regions, we just use the EM clustering, based on which we perform $Ncut$ value stability analysis. Particularly, we choose the best k to be

$$\arg \max_{i \in [2, 15]} \{|Ncut(i + 1) - 2Ncut(i) + Ncut(i - 1)|\} \quad (3.7)$$

which maximizes the second order difference.

Experimental results shown in Figure 3.4 demonstrate that the number of regions determined by this heuristic approach leads to meaningful segmentations.

3.3 Experiments

3.3.1 Test on Berkeley Benchmark

We have also conducted experiment on the Berkeley segmentation dataset (BSDS) [AMFM11]. Particularly, we use the BSDS500 dataset that contains 500 images (300 images for training and 200 images for testing) and their corresponding human segmentations. The only parameter we calibrated over the training data is the tradeoff factor α in (2.22). In other words, we did not perform any training and α is empirically adjusted to a fix value ($\alpha = 0.1$) that achieves the best performance over the training images only. We evaluate the results using the precision and recall framework of [MFM04], where the *Precision* measures the proportion of how many machine generated boundaries can be found in human labelled boundaries and is sensitive to over-segmentation, while the *Recall* measures the proportion of how many human labelled boundaries can be found in machine generated boundaries and is sensitive to under-segmentation, the commonly used *F-measure* = $\frac{2 \times Precision \times Recall}{Precision + Recall}$ are also reported.

Table 3.1 lists the the scores of our method and some state-of-art algorithms including the gPb-owt-ucm method [AMFM11], Normalized Cut [SM00], Mean-Shift [CM02] and Graph Based Region Merging [FH04]. We also report the scores of the the EM method and the original Potts model which uses only local color feature for comparison. The score

Table 3.1: Boundary benchmarks on BSDS500.

Method	Precision	Recall	F-Measure
EM method	0.46	0.49	0.48
Potts with RGB	0.53	0.43	0.48
Potts with GMM	0.59	0.84	0.69
Potts with GMM+Geodesic	0.72	0.54	0.62
Our Method	0.86	0.58	0.693
gPb-owt-ucm	0.72	0.73	0.725
Mean Shift [CM02]	0.59	0.71	0.645
NCuts [SM00]	0.56	0.74	0.64
Felz-Hutt [FH04]	0.5	0.77	0.61

of the Potts model using different combination of global features are also demonstrated. The performance gain is obvious with effective integration of the global information with the Potts model and the EM method in our algorithm. It can be seen that our algorithm obtains the highest precision with an average precision value of 0.86 across the test dataset, which means most of the boundaries generated by our method match human segmentation. On the other hand, as our algorithms aims at producing sizable segments, it exhibits a certain degree of “under-segmentation” compared to other methods, and thus the recall value of our method is relatively lower, meanwhile our F-measure ranks the second place in the existing method.

We would like to point out that the human segmentations offered in BSDS500 are of fine granularity, which goes against the goal of our algorithm, and thus the recall and F-measure values do not fully reflect the performance of our method. We did experiment of removing some human segmentations with much finer granularity and recalculating the statistics of our results and obtained an increase in the recall and F-measure score. A similar observation has been made by Hou et al [HYK13] in contour detection, where a method can achieve a high recall by detecting less frequently human labeled boundaries. Thus we believe a new benchmark which targets object agnostic segmentation should be built with an unified ground-truth like PASCAL VOC or a new evaluation metric which can consider the human labelers’ consistency should be adopted (e.g. the contours can be weighted according to the labeling consistency), which are still open problems worthwhile for exploration.

Our algorithm is implemented in C++ and runs on a Laptop with an Intel Core i7 1.73GHz Quad Core mobile processor, Nvidia Geforce GTX460M mobile graphics card and 8GB RAM. The average time to handle an image in the BSDS is about 1~3s with GPU acceleration.

3.3.2 Visual Results

Figure 3.4 shows some randomly selected images from the BSDS500 test dataset and their corresponding segmentation results by using different algorithms. It can be seen that NCut often breaks smooth image regions since it requires a large input label number in order to obtain the correct boundaries. Felz-Hutt method usually produces many super-pixels, which causes more severe visual artifacts. Blobworld and Mean-shift produce displeasing segmentation results for complex images as they rely on local image features. As for the gPb-owt-ucm method, it can still produce trivial regions since it is constructed from region contours which can be of fine granularity. Compared with these existing methods, our proposed algorithm generates pleasing segmentation results with boundaries snapped to the geometry features of objects and reasonable number of segments matching global human perception. Moreover, we also show the results of the original Potts model which only relies on local color feature. We can observe that our proposed method that incorporates global color, spatial and structure information into the Potts model achieves much better visual results.

3.3.3 Limitations

Although the proposed method achieves very good visual results for most of the tested images, it still has some limitations. One limitation is that our method may ignore some small distinct regions due to the assumption of the method that the size of each segment is considerable. Another limitation is that for cluttered or camouflaged images that do not exhibit much structure information in eigenvectors, our algorithm does not perform well. We believe for such cases high-level knowledge should be involved in order to successfully segment the images.

Figure 3.4: Examples of some images randomly selected from the BSDS500 test dataset and their corresponding segmentation results of different methods with optimal parameters tuned over the training set. Note that red and yellow contours depict the region boundaries. For Blobworld, it uses white contours for boundaries and the gray regions indicate unlabelled pixels.

Chapter 4

Multiple Foreground Recognition and Cosegmentation

The prevalence of digital cameras and mobile phones allows people to record their daily life in a visually rich way. This chapter begins to explore developing tools which can help effectively manage, understand and exploit a set of photos a user takes for a certain event, which will be very useful and leads to many exciting applications. Typically, such event photos contain multiple foreground objects of interest, but only an unknown number of these objects appear in each image, where the background may also vary. These content misaligned sets of images pose great challenges for most cosegmentation algorithms [KX12, KXLK11, JBP12, RSLP12, JBP10, CLL11, MSP11, MSD09] as their objective function incorporate the assumption that the object-of-interest should appear in each input image, without explicitly considering the cases as aforementioned. To apply traditional cosegmentation methods to such a photo set, a cumbersome manually process need to adopt to divide the photo stream in to groups so that each group contains the object-of-interest. Recently, Kim and Xing [KX12] proposed an approach specifically designed to address this multiple foreground scenario, and achieved superior results in comparison with other existing none multiple foreground methods. However, though making the solution tractable, their design counting on coarse segmentation and the restriction imposed on generating foreground candidates are often over-simplified treatments and give only a sub-optimal solution for complicated realistic scenes. This chapter concerns the problem of localizing, recognizing, and segmenting multiple foreground objects jointly from a general user’s photo stream. We refer to such a problem as *Multiple Foreground Object Recognition and Cosegmentation* (MFRC). This MFRC task is challenging due to strong intra- and inter-object variation, background clutter and sharing features among different classes of objects, to name a few, in addition to the irregular foreground occurrence patterns mentioned earlier. Our work is motivated by the recent study of Kim and Xing [KX12], but it significantly advances the MFRC performance with several novel techniques (see Fig. 4.1).

Figure 4.1: Visual comparison between the state-of-the-art MFC method [KX12] and our method on two images from the FlickrMFC dataset. The MFC method wrongly annotates the *apple+bucket* foreground as the *girl+red*, and the hair of the *girl+red* foreground as the *girl+blue* foreground. It also misclassifies the background as the *dog+white* in the second test image. Our method does a much better job in resolving the ambiguity in the MFRC task with a hierarchical CRF model using higher-level object cues.

Inspired by the impressive recent advance in scene understanding [SWRC09, LRKT10, KLT09], object recognition, detection and segmentation [FPZ07, FGM10, ADF10, CS12, KGF12, KF12], we cast the MFRC problem similarly within a conditional random fields (CRFs) framework in a principled manner. At the heart of our proposed approach is the integration of the objectness notion into a probabilistic CRF model. Our key observation is that in general the common goal of MFRC is to segment out and annotate foreground objects or “things” (e.g. girl dressed in red, apple bucket) rather than “stuff” (e.g. sky, grass). Similar ideas of incorporating object-like proposals [ADF10] or object detectors [FPZ07, FGM10] in a conventional CRF framework have been successfully applied before to other vision tasks such as large-scale image segmentation [KGF12] and scene understanding [LSA⁺10]. However, the MFRC task considered here is unique and very challenging – the user only gives a minimal amount of annotations on just a few example photos, while the possible geometric and photometric variations that irregularly occurring multiple foregrounds exhibit across the photo set can be quite large. This chapter is hence triggered to answer how far we can achieve for the challenging yet useful MFRC task, leveraging recent advances from object detection [WSTS07] to robust higher-order CRFs inference [KLT09].

To handle the MFRC problem, we propose a few robust and complementary objectness cues and object-based labeling consistency constraints (e.g. contours, multi-class object detectors, layout patterns), and combine them with the low-level unary and the pairwise terms holistically in a CRF model. We further augment the CRFs by including robust higher-order potentials defined over detected object regions, which is beneficial to inference results but also can be solved efficiently with graph cut based move making algorithms [BVZ01]. Experiments on the FlickrMFC dataset demonstrates state-of-the-art performance of the proposed algorithm, which generates spatially coherent, boundary-accurate segmentation results with correct and consistent multiple foreground recognition.

4.1 Relations and Comparison with Previous Work

Cosegmentation. There is a vast amount of prior work on cosegmentation [KX12, KXLK11, JBP12, RSLP12, JBP10, CLL11, MSP11, BKP⁺11, CXGS12]. Most of the existing works focus on handling the binary cases, separating foreground(s) from the background, but few of them are designed for joint multi-class object recognition and segmentation. The unsupervised methods such as DC [JBP10], Cosand [KXLK11] and MC [JBP12] used low-level bottom-up features, so they cannot distinguish “stuff” from “objects” in presence of background clutter and sharing features among classes. To overcome the ill-defined nature of unsupervised methods, some user inputs are hence desired and also often necessary. One notable work is iCoseg [BKP⁺11], which solves binary

Figure 4.2: Overview of our MFRC system. It integrates various object-level cues with low-level cues in a probabilistic CRF model.

foreground cosegmentation using graph cut. Our proposed algorithm involves a minimal amount of user annotations on a very small fraction of the image set in the form of bounding boxes (or polylines) and object labels. But unlike the aforementioned methods, we do not require the user to carefully sort out a given event photo set manually to group images containing the same objects together. The MFC method [KX12] is one of the existing works which attempt to solve the irregularly occurring multiple foregrounds problem. Similar with MFC [KX12], our method also deals with the multiple foreground cosegmentation problem. However, we perform joint detection and segmentation of multiple objects for a set of event photos, which often exhibit high variability of foreground objects in shape, color and their complicated interactions with other objects or varying backgrounds. Technically, our algorithm incorporates the higher level non-local object cues into a probabilistic inference and optimization framework, which has not been explored before in previous cosegmentation works. Such non-local cues, which help to differentiate “stuff” and “objects”, are expressed as soft constraints. Thanks to the soft constraint, multiple hypotheses can compete to make our method robust to false positive detection hypotheses, so they do not affect the final results when strongly defended by the hypotheses based on pixels and segments. Recently, Ma and Latecki [ML13] proposed a semi-supervised graph based method to perform the MFC task with a new connectivity constraint and achieved state-of-the-art on the subset of FlickrMFC dataset. In fact, our work is theoretically complementary to Ma and Latecki’s work, which proves that higher order constraints are beneficial for the MFRC task. In addition, our method is scalable to large image datasets, while the method in [ML13] does not scale well due to its reliance on dense pair-wise image analysis. In terms of the experimental results, [ML13] does not report on the challenging “thinker+Rodin” group existing in the full FlickrMFC dataset, which features challenges such as strong intra object variation, background clutter and lighting and scale changes. In contrast, we reported the results on the full

FlickrMFC dataset, and achieved much better accuracy than the MFC method [KX12] on the “thinker+Rodin” group even by 50%.

Object recognition, localization and segmentation. The last few years have seen impressive progress for several areas such as object recognition [FPZ07, FGM10], generic object localization [ADF10, CS12], and object segmentation [KGF12, KF12]. For instance, objectness window [ADF10] have been successfully applied to single foreground segmentation propagation in ImageNet [KGF12]. Multiple foreground proposals [CS12] have also been applied to Sarah *et al.*’s binary foreground cosegmentation work [VRK11] and achieved state-of-the-art results. At the same time, combined multi-class object segmentation and recognition techniques have also been proposed to address the grand challenge of complete scene understanding [SWRC09, LSA⁺10]. Ladicky *et al.* [LSA⁺10] proposed to incorporate object detector-induced potentials into a CRF energy optimization framework as a soft constraint, which clearly improved the standard object class segmentation models that tend to underperform on the “things” classes for complex scenes. Inspired by these nice existing techniques, our work, however, also differs from them in several aspects. First of all, as explained earlier, the MFRC task is very unique and challenging due to the high variability of foreground objects across the given set of photos and the minimal supervision that is available. Second, geared towards this MFRC task, our algorithm has integrated and extended some selected technical modules. For example, we used discriminative color features [WSTS07] to train multiple object detectors. In addition, contour as an important object-oriented property has been novelly exploited in this chapter, which proves its effectiveness in the MFRC task.

4.2 Problem Formulation and Our CRF Model

Given a set of N input images $\mathcal{I} = \{I_1, \dots, I_N\}$, $m(m \ll N)$ of them $\mathcal{I}_t = \{I_t^1, \dots, I_t^m\} \subset \mathcal{I}$ are first annotated to specify the objects of interest and also their rough spatial extent in the form of bounding boxes or polylines. More specifically, each image from this small training set $\mathcal{I}_t$ with user supervisions contains a subset of annotated objects belonging to K different foregrounds $\mathcal{F} = \{F^1, \dots, F^K\}$. Each foreground F^l is associated with a numeric label $l \in \mathcal{L} = \{0, 1, \dots, K\}$, where 0 is used to denote the background for notational simplicity. We formulate the MFRC problem in terms of a global energy function defined on a conditional random field (CRF), for which the goal is to assign a random variable x_i to each pixel i in each image a label from $\mathcal{L}$. Our framework integrates various complementary object cues computed from different classifiers learned with low-level features, mid-level edge detectors and an interactive offline bounding box object detector. In fact, the proposed framework also allows to choose any state-of-the-art multi-class object detectors and classifiers, though we will present concrete modules in this chapter.

Fig. 4.2 illustrates the proposed framework, which consists of a few stages and several modules. During the preprocessing stage, various foreground cues such as unary multi-class pixel and segment classifiers, object detectors and gPb contour [AMFM11] are modeled and generated. Pixel, segment and object detectors classifiers are trained with user-drawn bounding boxes. The gPb contour map is generated by combining the edge signal from eigenvectors with low-level cues, and it captures mid-level object contours. After the preprocessing stage, based on the gPb signal, we specially design a contour-layout filter to reject false positive detector responses which are very likely to be “stuff”. With all the cues computed, we integrate them into a global energy function which enforces the labeling consistency between various level cues and finally produce the solution with fast expansion/move solvers. Once the initial segmentation is generated, our framework supports iteratively updating the learned models and performing the recognition and segmentation tasks to further improve the results.

4.2.1 Proposed CRFs Framework for MFRC

To make our algorithm linearly scalable with the image set size N , the recognition and segmentation inference is performed individually for each image $I_n \in \mathcal{I}$, similar to MFC [KX12]. We formulate the MFRC task as a multi-labeling problem with a CRF framework on a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ is the set of all image pixels of image I_n , while $\mathcal{E}$ corresponds to the set of all edges defined by a four or eight neighbor system. The proposed probabilistic CRF model is given by a Gibbs energy function as follows:

$$\begin{aligned} \mathbf{E}(\mathbf{x}) = & \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) \\ & + \sum_{s \in \mathcal{S}} \psi_s(\mathbf{x}_s) + \sum_{d \in \mathcal{D}} \psi_d(\mathbf{x}_d) . \end{aligned} \quad (4.1)$$

In (4.1), $\mathbf{x}$ denotes any valid label assignment to the random variables $\{x_i\}$, which takes a value from the object label set $\mathcal{L}$. $\mathcal{S}$ denotes a superpixel decomposition of the image I_n into a set of disjoint segments $\{s\}$, and $\mathbf{x}_s$ is the clique of pixels covered by the segment s . We denote the set of object detections with $\mathcal{D}$, which are typically returned in the form of bounding boxes enclosing objects. The pixels covered within the d -th detection bounding box are represented as $\mathbf{x}_d$. Our energy function consists of four terms: **1)** the pixel-based unary potential $\psi_i(x_i)$, evaluating the likelihood of a certain label assignment to pixel i ; **2)** the pairwise smoothness potential $\psi_{ij}(x_i, x_j)$, penalizing differently labeled adjacent pixels of similar appearance; **3)** the segment-level robust label consistency potential $\psi_s(\mathbf{x}_s)$, charging the label inconsistency cost robustly with the number of variables in the segment s not taking the segment label; **4)** the object detector potential $\psi_d(\mathbf{x}_d)$, enforcing a robust region label consistency constraint that is defined in a

similar way to $\psi_s(\mathbf{x}_s)$. These terms collectively capture the information for image/object representation and understanding from different levels in a complementary way. We will elaborate the last two terms modeled as robust high-order potentials in Sect. 4.3 and 4.4. A contour-based pairwise smoothness potential $\psi_{ij}(x_i, x_j)$ that improves the standard contrast-sensitive implementation [RKB04] will be presented in Sect. 4.3.3.

Pixel-based unary potential. The first term $\psi_i(x_i)$ is a unary potential defined on each pixel which indicates its cost of being assigned to a label $l \in \mathcal{L}$:

$$\psi_i(x_i) = -\omega_{pix} \log P(x_i | \mathcal{C}_{pix}); \quad (4.2)$$

where ω_{pix} is the weighting factor. $P(x_i | \mathcal{C}_{pix})$ denotes a normalized distribution returned by a random forest classifier $\mathcal{C}_{pix}$, which is an ensemble of weak decision trees [AG97]. The classifier is trained with the pixel-level features whose corresponding labels provided by the user. The features defined on each pixel is a seven-dimensional vector, which consists of six color features (RGB and L^*ab) and one texton feature. We generate textons by convolving the image with 17-dimensional filter banks at different scales, as in [SWRC09]. Then the filter bank responses are clustered using K-means algorithm into T_c code words to generate a texton map which encodes the final pixel-wise texton feature.

4.3 Incorporating Object Cues

This section presents a few complimentary object cues extracted with different technology, which characterize different aspects of an object in the proposed CRF model for the MFRC task. We also discuss the methods to define the corresponding object-oriented potentials.

4.3.1 Fast Object Detectors with Boosted Color Bins

The appearance of an image patch/segment by itself is often ambiguous when different objects and background contain similar local features, as it is incapable to capture the global configuration information of object class instances. This motivates us to address the MFRC challenge from higher and longer range grouping levels which have been proved to be useful in some image summarization and scene understanding research [KLT09, LRKT10]. A popular approach is to reason about the objects of interest with the help from rectangular bounding boxes which are generated from some detection methods [FPZ07, VGVZ09, FGM10]. But, such detections usually require a large number of training examples and often pose strong structured spatial layout constraints. Though deformable part models [FGM10] can relax the rigid spatial configuration constraint, they

are typically slow and not suitable for the MFRC task which shall parse a comparatively small set of images but with strong object variations.

To obtain bounding box proposals more robustly with the invariance to e.g. scale, rotation and non-rigid motion, we train a multi-class interactive offline color based object detector. Given a user drawn bounding box, we adopt the method of Wei *et al.* [WSTS07] by projecting all pixel colors onto a set of one dimensional (1D) lines in the RGB color space. These lines have different directions and pass through the point (128, 128, 128). The directions in color space are evenly sampled by 13 lines and then a 1D (normalized) histogram of the projected values is calculated on each line. We also use eight bins for each histogram through an empirical comparison and treat all $13 \times 8 = 104$ color bins as our features. Such features can be extracted using integral histogram [Por05] very efficiently in a constant time. For the multiple foreground recognition problem considered here, we use the Joint Boosting algorithm rather than Adaboosting adopted in [WSTS07], and train a multi-class bounding box classifier. The details of our learning procedure resembles closely with those described in [SWRC09]. Similar to [WSTS07], our training examples are generated from the user annotated images, which however have multiple class labels. To generate more positive samples and also be robust to object variations across images, the same appearance perturbation scheme [WSTS07] is employed, which perturbs the position of the object rectangles randomly by a small amount. Our negative examples are randomly sampled around the non-selected foreground regions using the bounding boxes of the same size as the user-specified ones. The bounding box proposals are generated by sweeping the object windows for a set of predefined scale levels in a test image. They are evaluated by the trained multi-class classifiers, whereas only the top-scored bounding boxes are retained.

4.3.2 Detector-Based Robust Consistency Potentials

A big difference between our energy function and that of conventional binary foreground segmentation is the higher-order bounding-box level potential involved. With the higher-order object information from bounding boxes, we can revolve some ambiguity which would otherwise be too hard to solve at a local level. The bounding box proposals are used to define a kind of soft constraint which works jointly with other hypotheses. We incorporate the object potential $\psi_d(\mathbf{x}_d)$ into our CRF framework by enforcing it as a robust region label consistency constraint defined in [KLT09]. Given the d -th detection bounding box $\mathbf{x}_d$ with a score R_d and the detected object label l_d , $\psi_d(\mathbf{x}_d)$ is defined as:

$$\psi_d(\mathbf{x}_d) = \begin{cases} N_d \frac{1}{Q_d} \gamma_{max} & \text{if } N_d \leq Q_d \\ \gamma_{max} & \text{otherwise,} \end{cases} \quad (4.3)$$

where $N_d = \sum_{i \in \mathbf{x}_d} \delta(x_i \neq l_d)$ is the number of variables in $\mathbf{x}_d$ not taking the dominant label l_d . The truncation parameter Q_d controls the maximum number of inconsistent

Figure 4.3: Effects of the object detector-based robust consistency potential. (a) A user-annotated image with two foreground instances labeled with bounding boxes. (b) Applying a learned object detector to a novel image. Shown are top-scored bounding boxes. Segmentation result (c) without and (d) with using the proposed object detector-based label consistency potential.

Figure 4.4: Behavior of the object detector potential as a soft constraint. (a) Input image. (b,c) Two object detection results. Our results (d) without and (e) with using the object detector potentials.

pixels. The cost γ_{max} , in the MFRC context, is now defined by a linear truncated function $f(\cdot)$, and it monotonically increases with the object classifier response R_d as

$$f(\mathbf{x}_d, R_d) = \omega_d |\mathbf{x}_d| \max(0, R_d - R_t) , \quad (4.4)$$

where R_t is a threshold and ω_d defines the detector potential weight. Our region consistency constraint is similar to the object detector term used in [LSA⁺10] for scene understanding. If a detector response is strong, the higher-order potential will encourage the pixels belonging to the bounding box $\mathbf{x}_d$ to take the label l_d . As the penalty is increased with the number of inconsistent pixels incrementally until the truncation threshold Q_d , this soft higher-order constraint produces better labeling results than the standard P^n Potts model [KKT09], which forbids other differently labeled pixels within the clique $\mathbf{x}_d$. The proposed object potential $\psi_d(\mathbf{x}_d)$ can be transformed to take the Robust P^n form [KLT09, LSA⁺10]:

$$\psi_d(\mathbf{x}_d) = -f(\mathbf{x}_d, R_d) + \min(f(\mathbf{x}_d, R_d), k_d \sum_{i \in \mathbf{x}_d} \delta(x_i \neq l_d)) , \quad (4.5)$$

where k_d is a slope parameter defined in the same way as in [LSA⁺10]. Including this term to a CRF model is implemented by adding two auxiliary nodes into the graph, and the augmented energy function can be efficiently minimized with the graph cut algorithms. Interested readers are referred to [KLT09] for the graph optimization details. Fig. 4.3 demonstrates the strength of the object detector-based potential when integrated into our CRF framework. Without using the detector-based potential, the black dog can only be partly annotated and segmented due to the weak low-level hypotheses based on pixels and segments. The object detector potential provides complementary high-level evidence and integrating it into the CRF model results in a more accurate result of recognizing the missed dog parts.

4.3.3 Contour for Object Boundary Reasoning

According to the cognition study [HS97], human vision views object part transitions at those with negative minima of curvature and the part salience depends on three factors: the relative size, the boundary strength and the degree of protrusion, so part transitions convey some mid-level information to help differentiate “object” and “stuff”. Conventional contour detectors capture part transitions by finding local extrema, which usually result in a high recall but low precision contour detection result. Recently, Arbelaz et al. [AMFM11] proposed to combine the contour signal from eigenvectors with the low-level contour signal and achieved the state-of-the-art contour detection results. The eigenvectors v are generated by solving an eigen-system $(Z - W)v = \lambda Zv$, where $W = \{w_{ij}\}$ is a sparse symmetric affinity matrix encoding the pairwise similarity between

Figure 4.5: Proposed contour-layout filter based on the gPb contour map. (a) Input image. The blue bounding boxes indicate example detection results for the *apple bucket* image. (b) The intensity-inverted gPb map. (c) Proposed contour-layout filter (see the text for the details). (d) Close-up views of false/true positive object detections.

image elements i and j based on the intervening contour cues [MFM04, AMFM11]. The diagonal matrix $Z = [z_{ij}]$ is defined with $z_{ii} = \sum_j w_{ij}$. As the affinity matrix W captures the global image information and the eigenvectors of the eigen-system are the solution to minimize the Ncut criteria [SM00], the eigenvectors capture the contour belonging to the transition between large object parts. This nice property makes gPb valuable for higher level image analysis. A more detailed formulation of gPb can be found in Sec. 2.1. Fig. 4.5(b) shows an example gPb contour map C .

Contour-layout filters to reject false objects. Since our object detectors presented earlier use only color features for the robustness reason, the detected object proposals would unavoidably contain false positive detection results that belong to “stuff” such as sky. Inspired by the aforementioned part salience theory, we propose to exploit the gPb signal to define a novel objectness measure, which we call *contour-layout filters*. The basic idea is that if a detected bounding box falls on a non-object region, the contour distribution around the region tends to quite monotone, so we can reject this kind of detection results with high confidence. To extract such a distribution, we first enlarge a detected bounding box B_p centered at pixel p by a ratio γ , while preserving the original aspect ratio of B_p . The resulting relaxed rectangle B_p^γ defines the out-of-box bound. Next, we quantize the region around pixel p into H directions, and shoot H rays distributed evenly apart in angle (i.e. $2\pi/H$) from the center pixel p . If the ray for a quantized orientation bin $o \in \{0, 1, \dots, H-1\}$ hits a salient gPb signal (with a strength above a threshold τ_{gPb}) within B_p^γ , we assign a value of 1 to the corresponding o -th component of a vector $V_p = [v_0, v_1, \dots, v_{H-1}]^T$, otherwise 0 is assigned. In this chapter, we set $H = 8$, as shown in Fig. 4.5(c). To be robust to the spatial and orientation sampling discretization, we consider the contributions of the gPb responses of neighboring pixels in a small circular patch around the ray sampling location. Given this vector V_p , our contour-layout filters finally classify the detected object bounding box B_p as a false positive result, if the L1 norm of the vector V_p is less than an empirically predefined threshold $\tau_{cl} = 0.4$ for all classes. As the detector based potential is designed as a soft constraint in the inference, our method is not sensitive to the threshold and works well. We find this simple scheme is very effective in rejecting false object detections (see Fig. 4.5(d)) and preventing them from confusing the CRF inference, though more sophisticated methods to compute the objectness measure using the gPb contour can also be employed.

Contour-based pairwise potential. Observing that the gPb contour map provides more reliable and higher-level reasoning of salient contours, we propose to compute the pairwise potential $\psi_{ij}(x_i, x_j)$ as follows,

$$\psi_{ij}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ \omega_a(1 - \|\nabla C(i, j)\|^2) & \text{otherwise,} \end{cases} \quad (4.6)$$

where ω_a gives the weight of the pairwise potential. $\nabla C(i, j)$ measures the gPb signal contrast between two adjacent pixels i and j . We observe this new pairwise term reduces

Figure 4.6: SLIC segmentation and segment-level consistency potentials. (a) SLIC segments [ASS⁺12]. (b) Segment-level and (c) pixel-level classifier response map for the *baby* foreground. (d) Robust region consistency constraint. The big orange dot denotes a segment assigned a label l_s . The blue dot denotes an outlier pixel.

the possibility of incorrect boundary alignments compared with average color contrast based pairwise term [RKB04].

4.4 Segment-Based Label Consistency Potential

The pixel-level features are usually too local to capture the change of neighborhood patterns, so we include an additional level of variables which consist of super-nodes/segments. We choose the SLIC algorithm [ASS⁺12] to over-segment the image into homogeneous regions. SLIC segments have been showed to give superior performance in terms of boundary adherence and segmentation compactness. Based on the generated super-nodes, we train a segment-based random forest classifier $\mathcal{C}_{seg}$. The feature computed at the segment level is the histogram of the textons with a dimension of T_c , which is defined earlier when producing the pixel-level features.

Now we present the formulation of the super-node based higher-order terms. Basically, we follow Kohli *et al.*'s method [KLT09] to build a multi-layer hierarchical CRF model (two layers in our case), where the base layer consists of pixels and the second layer is made up of super-nodes which encode mid-level region cues. Such a construction enforces a soft constraint on the pixels belonging to a segment, encouraging them to be labeled as the same as their parent, but it also allows some outlier pixels (see Fig. 4.6(d)). Using a soft constraint makes our approach robust to the super-node quantization artifacts, while leveraging segments' grouping power and complementary cues extraction from a higher level for the given image. We have also tested hierarchical CRF models with more levels of super-nodes, and found that the results obtained are similar but at more computational costs. Fig. 4.6 shows a visual comparison between the segment-level classifier response map (color-coded as a heat map) and its pixel-level counterpart. One can notice that the

Figure 4.7: Segmentation accuracy comparison between our method (MFRC) and other baselines (MFC [KX12], COS [KXLK11], DC [JBP10], LDA [RFE⁺06]) for the FlickrMFC dataset. The S and U denote whether the method is supervised or unsupervised.

segment-level classifier response is often stronger and more reliable than the pixel-level response, which tends to be noisy though with more details.

Given a segment $s \in S$ and its associated clique of pixels $\mathbf{x}_s$, let $N_s = \sum_{i \in \mathbf{x}_s} \delta(x_i \neq l_s)$ denote the number of variables in $\mathbf{x}_s$ not taking the segment label l_s , the super-node potential is designed as a linear truncated function [KLT09]:

$$\psi_s(\mathbf{x}_s) = \omega_s \cdot \begin{cases} N_s \frac{1}{Q_s} (\rho_{max} - \rho_{l_s}) + \rho_{l_s} & \text{if } N_s \leq Q_s \\ \rho_{max} & \text{otherwise,} \end{cases} \quad (4.7)$$

where ω_s is the weighting factor for the super-node potential. $\rho_{l_s} = -\log P(\mathbf{x}_s | \mathcal{C}_{seg})$ indicates the cost charge for a super-node $\mathbf{x}_s$ to take the label l_s . $P(\mathbf{x}_s | \mathcal{C}_{seg})$ is given by a random forest super-node classifier $\mathcal{C}_{seg}$. ρ_{max} is the maximum cost charge when a number of Q_s pixels do not take the label l_s . This segment potential can also be finally transformed to the Robust P^n form:

$$\psi_s(\mathbf{x}_s) = \min \left\{ \min_{l_s} (\rho_{l_s} + k_{l_s} \sum_{i \in \mathbf{x}_s} \delta(x_i \neq l_s)), \rho_{max} \right\}, \quad (4.8)$$

where k_{l_s} is the slope parameter similarly defined as k_d in (4.5). Similar to the object detector potential, this term can be minimized by including two auxiliary node in the graph and solved efficiently with graph cut [KLT09], a brief introduction can be found in Sec. 2.2.2.1.

4.5 Experimental Results and Discussions

We evaluate our method using the FlickrMFC dataset [KX12]. This dataset is the only MFC dataset consists of 14 groups of images with manually labeled ground-truth. Each group includes 10~20 images which are sampled from a Flickr photostream. This dataset is challenging as it contains a finite number of repeating subjects that are not presented in every image and there are strong lighting variation, pose change and background clutters

in the images. The parameters are empirically fixed as: $\omega_{pix} = 1, \omega_a = 10, \omega_s = 0.2, \omega_d = 0.1, R_t = 0.5, \rho_{max} = -\log(0.1)$. The overall time (including preprocessing, detection and segmentation) to process each image is around 20~30 seconds on a desktop Intel Core i5 3.2GHZ and 8GB RAM.

Quantitative Results: We compare our method with some baselines: MFC [KX12], CoSand [KXLK11] (COS), Discriminative Clustering(DC) [JBP10], LDA [RFE⁺06]. We adopt the procedure introduced in MFC [KX12] for evaluation. For supervised methods such as our method and MFC’s supervised version, we randomly pick 20% of the input images to annotate. For the unsupervised methods, e.g. CoSand [KXLK11], DC [JBP10] and LDA [RFE⁺06], the dataset is divided into several subgroups such that the images in each subgroup contain the same objects of interest, the methods are applied to each subgroup individually. We evaluate the segmentation accuracy by the standard intersection-over-union metric $\frac{(GT_i \cap R_i)}{(GT_i \cup R_i)}$.

Fig. 4.7 summarizes the segmentation accuracy on the 14 groups of the FlickrMFC dataset. The left most bar set presents the average segmentation accuracy on 14 groups. Since COS, DC, LDA and MFC-U are unsupervised methods which count on low-level cues, they failed to capture the real objects of interest, so their performance is not competitive in most cases. We hence focus on the comparison with the state-of-the-art MFC method. As shown in the bar chart, our algorithm’s average accuracy is around 10% higher than the MFC method [KX12]. Some datasets like *cheetah*, *butterfly*, *liberty*, we achieve around 20% accuracy improvement. For the *thinker* dataset, the accuracy gap reaches even 50%!

We have also evaluated the average accuracy gain contributed by including higher-order segment and detector potentials into the CRF model, which is about 2.3%. This numerical small gain has also been observed in Shotton *et al.* [SWRC09] and Kohli *et al.* [KLT09] in scene understanding research. As also indicated in [SWRC09, KLT09], we observe that including these potentials often bring a pronounced increase in perceived accuracy, especially for the challenging cases such as Fig. 4.1, 4.3 and 4.4.

Visual Results: Fig. 4.8 shows some visual results from seven groups of FlickrMFC dataset. For each set, the input images and color-coded segmentation results are displayed in the first two rows from top to bottom. The regions which are labeled with the same color in each set indicate they belong to the same category. The tags below each set explain the meaning of each color. From the images, one can observe that our method can handle irregularly appearing objects and produce smooth and accurate segmentation results. The images with no foregrounds are also correctly identified, e.g the *liberty* dataset. On the other hand, our current model still cannot handle some camouflage cases very well, such as the *butterfly* dataset.

Figure 4.8: Some randomly drawn examples from seven groups of the FlickrMFC dataset. From top to bottom, each set presents its input images, color-labeled segmentation results. The colored tag below each set indicates which category each region is assigned to.

Chapter 5

Multiple Human Identification and Cosegmentation

Given a set of personal photo stream about certain event, this chapter furthers the study in previous chapter to human identity grouping and segmentation, which is a highly desirable tool to understand and manipulate these photos which often focus on persons. With this tool, users will be equipped with the potential to manage their photos based on consistent appearance patterns of the persons-of-interest or to propagate editing operations among certain human groups, and many other exciting vision and graphics applications can be further developed. Face detection and recognition based human grouping has become popular in recent commercial applications such as Google Picasa or Facebook. Such face-based grouping methods work well for frontal faces, but experience difficulties for profile faces and occlusion. Some recent methods incorporate cues from torsos based on a face detector to improve the human identification performance [AcLGS07]- [ZCLZ03], but they still inherit the limitation of face detectors. This chapter aims to develop a holistic approach to localize, identify and segment multiple humans jointly in a personal photo album, which we refer to as a *Multiple Human Identification and Cosegmentation* (MHIC) problem.

The MHIC problem is challenging due to its unique irregular object-occurring patterns, strong variations in foreground/background and feature sharing among different classes, which inherit from a similar *Multiple Foreground Recognition and Cosegmentation* (MFRC) problem in Chapter 4. It is actually very different from the classic cosegmentation task studied by many existing algorithms [KXLK11, JBP12, RSLP12, JBP10, CLL11, MSP11, MSD09], which assume object-of-interests to appear in all input images. Although some recent works [KX12, ML13, ZLC⁺14] achieve certain progress in the MFRC scenario, their performance for the “human” class is still far from satisfactory due to the large variation of human bodies, self-occlusion and unique structures of human bodies compared with other objects.

Inspired by the impressive recent advances in scene understanding [SWRC09, LRKT10, KLT09], human detection [BMBM10, FGM10], tracking [WSTS07] and segmentation [LTZ13, VWLT11], we solve the MHIC problem with a conditional random fields (CRFs) framework in a principled manner. At the heart of our approach is the integration of the human notion into a probabilistic CRF model, which is implemented with a few innovative human object cues proposed in this work. Our key observation is that the essential goal of MHIC is to segment out and annotate “humans” rather than other objects or “stuff” (e.g. sky, grass). Such a human-centric constraint has not been explored in the previous works [KX12, ML13, ZLC⁺14], and we propose an effective and efficient framework to address the MHIC task with significantly improved performance. Similar ideas of incorporating object-like proposals [ADF10] or object detectors [FPZ07, FGM10] in a conventional CRF framework have been successfully applied before to other vision tasks such as large-scale image segmentation [KGF12], scene understanding [LSA⁺10] and co-segmentation [ZLC⁺14]. However, the MHIC task considered here is unique and very challenging – the user only gives a minimal amount of annotations on just a few example photos, and possible geometric and photometric variations that irregularly occurring human instances exhibit across the photo set can be quite large. This chapter is hence triggered to answer how far we can achieve for the challenging MHIC task, leveraging recent advances from human detection and tracking [WSTS07, BMBM10] and robust higher-order CRFs inference [KLT09].

5.1 Related Works

Human Detection and Segmentation: Human detection and grouping is a fundamental task with many real applications. The tree-structured pictorial framework [FGM10, YR13] is well-known for human detection and achieves leading performance on several PASCAL VOC challenges. However, this framework is not good at handling foreshortening and partial occlusion. Bourdev et al’s work [BMBM10] eschewed the pictorial structure by learning poselets for human parts which are tightly clustered in the appearance and configuration space and achieved more accurate localization. It also provides richer information from training data, e.g. by transferring the binary segmentation mask from the training data to the test image, so it can generate a rough, aggregated belief mask which indicates the location of certain human parts. Human detection is widely applied in recent human grouping tasks [SLJ13, GSRS11, TBS12, SEZ09, SZS06] in image and video tracking and achieves better results due to using larger spatial support than previous works using features extracted from human face detection [AcLGS07, SL06, ZCLZ03], as the latter has difficulties with profile poses and positions. However, these recent works do not perform any pixelwise labeling and they do not exploit any background modeling to boost identification.

Figure 5.1: Overview of our MHIC algorithm. Given a set of event photos, a user annotates only a few of them to indicate human instances of interest with bounding-boxes in different colors. Our algorithm then jointly localizes, identifies and segments out all the human instances for the given image set. It tackles the MHIC task by integrating object-level cues with mid-level and low-level cues in a probabilistic CRF framework. The proposed algorithm yields the segmentation of human instances of interest in each image, and identifies them by taking on the same color annotated earlier for the corresponding human class.

Human segmentation, on the other hand, intends to assign each pixel a label to indicate whether the pixel belongs to a human or background. Traditional semantic segmentation [SWRC09] cannot be directly applied to the MHIC task as all the pixels will be assigned the same label “person”. Recently, Ladicky *et al.* [LTZ13] and Vineet *et al.* [VWLT11] proposed to segment human instances from a single image or video using conditional random field and human detection. However, these methods essentially made prior decisions on human hypotheses, because they are initialized by human detectors and are sensitive to false detections. Different from these methods, our method makes a joint decision on human instances by taking account of the cues from various levels, which is more robust to false and duplicate human detections. In addition, the existing method also do not perform any identification on human instances.

Cosegmentation: There is a vast amount of prior work on cosegmentation [KX12, KXLK11, JBP12, RSLP12, JBP10, CLL11, MSP11, BKP⁺11, CXGS12]. Most of the existing works focus on handling the binary cases, separating foreground(s) from the back-

ground, but few of them are designed for joint multi-class object recognition and segmentation. The unsupervised methods such as DC [JBP10] and Cosand [KXLK11] used low-level bottom-up features, so they cannot distinguish “stuff” from “objects” in presence of background clutter and sharing features among classes. To overcome the ill-defined nature of unsupervised methods, some user inputs are hence desired and also often necessary, and one notable work is iCoseg [BKP⁺11].

The aforementioned methods, however, require the user to carefully sort out a given event photo set manually to group images containing the same objects together. Recently, Kim and Xing [KX12] proposed the first method to handle irregularly occurred multiple objects cosegmentation problem – *Multiple Foreground Cosegmentation (MFC)*. They used a combinatorial auction approach with spanning tree-based pruning, which is an over-simplified model and produces sub-optimal results. Ma and Latecki [ML13] proposed to solve the MFC problem using a semi-supervised graph transduction framework which enforces connectivity in the labeling result, but this method is weak in scalability due to the reliance on dense pair-wise image analysis. Both of the aforementioned methods do not model the concept of “objects” explicitly, and they frequently label regions belonging to “stuff” such as “sky” and “grass” as foreground objects. To overcome this problem, our previous work [ZLC⁺14] in Chapter 4 includes higher level, non-local object cues into a probabilistic inference and optimization framework. The object cues derives from an object detector based on color-line modeling without any shape information. Although our previous work can handle objects exhibiting certain degrees of rotation, scale, and illumination changes and produce state-of-the-art performance for MFC tasks, it still generates unsatisfactory results for the “human” class due to strong geometric and photometric variations of human bodies as well as background clutters.

To detect specific human instances and better handle body variations, we propose to tackle the challenge by extending our previous work in Chapter 4, and combine a color line based object detector model [ZLC⁺14] and a poselet-based human detector [BMBM10] in a novel way. In addition, a soft shape mask map for humans is newly generated for each input image, which captures the spatial distribution of articulated human bodies probabilistically. Employing a shape prior has been proved to be very useful for scene understanding [TL13], but usually a rigid shape model is used [TL13,SWRC09]. Finally, we propose to compute pixelwise features using larger spatial supports provided by the cross region structure [ZLL09,LSM⁺12], which have been successfully applied to dense stereo [LSM⁺12] and saliency detection tasks [SWLL13], achieving robustness to noise and providing a better discriminative power. Beyond these novel algorithmic designs, this work not only has the scalability and long-range object modeling advantages of our previous work, but also greatly advances its performance in differentiating human classes from other non-human objects and “stuff”.

Scene Understanding: In the last few years we have also seen impressive progress in combining multi-class object segmentation and recognition techniques to address the

grand challenge of complete scene understanding [SWRC09,LSA⁺10]. Ladicky *et al.* [LSA⁺10] proposed to incorporate object detector-induced potentials into a CRF energy optimization framework as a soft constraint, which clearly improved the standard object class segmentation models that tend to underperform on the “things” classes for complex scenes. Recently, Tighe and Lazebnik [TL13] utilized the rigid shape information transferred by Exemplar SVM [MGE11] in scene understanding tasks, which achieved state-of-the-art performance. Inspired by these nice existing techniques, our work, however, also differs from them in several aspects. First of all, as argued in our previous work [ZLC⁺14], the MHIC task is very unique and challenging due to the high variability of foreground objects across the given set of photos and the minimal supervision that is available. Second, geared towards this MHIC task, our algorithm is designed with some novel and critical technical modules that explicitly model and handle “human” classes, whose non-rigid motions and complex interactions among them and also with the background create several new challenges.

Figure 5.2: The proposed cross region based pixelwise unary potential outperforms the conventional single pixel based counterpart [ZLC⁺14] for the MHIC task. (a) Two example adaptive cross regions defined at pixel p and p' . (b) A close-up view of the adaptive cross region defined at pixel p . (c) The color histograms H_c and texton histograms H_t extracted from the respective cross support regions at p and p' . (d) The conventional pixel-level classifier response map for the human class “boy-in-blue”. (e) The proposed adaptive region-based classifier response map for the same “boy-in-blue” class. (f) The final multi-class labeling result based on (d). (g) The final multi-class labeling result based on (e). (h) The ground truth label map.

5.2 Problem Formulation and Proposed CRF Framework

Given a set of N input images $\mathcal{I} = \{I_1, \dots, I_N\}$, assume m ($m \ll N$) of them $\mathcal{I}_t = \{I_t^1, \dots, I_t^m\} \subset \mathcal{I}$ are annotated with bounding boxes or polylines to delineate the spatial extents of certain humans of interest. Each image from this training set $\mathcal{I}_t$ contains a

subset of annotated humans belonging to K different humans $\mathcal{H} = \{H_1, \dots, H_K\}$. Each human H_i is associated with a label $l_i \in \mathcal{L} = \{0, 1, \dots, K\}$, where 0 is used to denote the background. The MHIC problem is formulated in terms of a global energy function defined on a conditional random field (CRF), for which the goal is to assign a random variable x_i for each pixel i in each image a label from $\mathcal{L}$. Our framework captures complementary cues of “human” classes from low level features on adaptive spatial support to mid-level contour detectors and high-level human detectors learned with different classifiers. In fact, the proposed framework is generic and flexible, and also allows to choose other multi-class object detectors and classifiers.

Fig. 5.1 illustrates the proposed framework, which consists of two main stages and several specific modules. During the preprocessing stage, various cues such as a cross region based pixel classifier, a color line based human instance detector and gPb contour are modeled and generated. Pixel and object detectors are trained with user-drawn bounding boxes. After all the cues are computed, we integrate them into a global energy function which enforces the labeling consistency between various level cues and finally produce the solution with fast expansion/move solvers. Once the initial segmentation is generated, our framework supports iteratively updating the learned models and performing the recognition and segmentation tasks to further improve the results.

5.2.1 Proposed CRF Framework for MHIC

To make the MHIC problem tractable, we make two practical assumptions: i) each image I_n contains a subset of human $\mathcal{H}$, and ii) persons with consistent appearance in terms of color and shape should be assigned the same human identity label. If one dresses differently in each image, it is nearly unlikely to label them consistently across input images.

The MHIC task is formulated as a multi-labeling problem within a CRF framework on a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ is the set of all image pixels of image I_n , while $\mathcal{E}$ corresponds to the set of all edges defined by an eight-connected neighborhood system. The notations are similar defined in Chapter. 4. The proposed energy function is defined as follows:

$$\begin{aligned} \mathbf{E}(\mathbf{x}) = & \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) \\ & \sum_{i \in \mathcal{V}} \psi_i^s(x_i) + \sum_{d \in D} \psi_d(\mathbf{x}_d) . \end{aligned} \quad (5.1)$$

In Eq. (5.1), $\mathbf{x}$ denotes the valid label map assigned to the random variables $\{x_i\}$, which takes a value from the label set $\mathcal{L}$. We denote the set of human detections with D , which are returned bounding boxes enclosing potential human instances. The pixels

covered within the d -th detection are represented as the clique $\mathbf{x}_d$. The energy function consists of four terms: **(1)** the pixel-based unary potential $\psi_i(x_i)$, which is trained by using the features extracted from a pixelwise adaptive cross support region [LSM⁺12]; **(2)** the pairwise smoothness potential $\psi_{ij}(x_i, x_j)$ based on a *gPb* contour detector [AMFM11]; **(3)** the soft shape potential $\psi_i^s(x_i)$ that evaluates the likelihood of each pixel to lie within each potential human shape, where the shape is adapted according to the internal image structure; **(4)** the object detector potential $\psi_d(\mathbf{x}_d)$, where $\mathbf{x}_d$ is the label of the bounding box, charging the label inconsistency cost robustly with the number of variables in the bounding box d not taking the detector label. These terms collectively capture the information for human instances in a complementary way. We will elaborate the four terms in following sections.

5.3 Cross Region Based Pixelwise Unary Potential

The first term $\psi_i(x_i)$ is a unary potential defined on each pixel which indicates its cost of being assigned a label $l \in \mathcal{L}$:

$$\psi_i(x_i) = -\omega_{pix} \log P(x_i | \mathcal{C}_{pix}) , \quad (5.2)$$

where ω_{pix} is the weighting factor. $P(x_i | \mathcal{C}_{pix})$ is the normalized probability evaluated by a Random Forest (RF) classifier $\mathcal{C}_{pix}$ [AG97]. Given the human class label provided in the form of user-drawn bounding boxes, a RF classifier is typically trained using color and texon features extracted from a single pixel, which is also the scheme used in Eq. 4.2. However, the pixel-level features are usually too local to capture the change of neighborhood patterns, often resulting in a noisy and weak classifier response (see Fig. 5.2(d)). As shown in Fig. 5.2(f), this weak response signal leads to an unsatisfactory segmentation result. This observation motivates us to make use of a larger spatial context for each pixel when training a RF classifier. Though superpixels appear to be an option here, partitioning an image into non-overlapping local regions suffers from the superpixel quantization artifacts. As a result, the segment-based classifier response is typically coarse and does not preserve spatial boundaries and details well.

Recently, pixelwise adaptive spatial supports—cross regions [ZLL09, LSM⁺12]—have been successfully used in stereo matching cost aggregation [ZLL09], image filtering [LSM⁺12] and saliency detection [SWLL13], which achieved more robust and accurate results than single pixel or superpixel based estimation. Another advantage of adaptive cross regions is that they can be very efficiently computed for each pixel densely. In this work, we investigate incorporating this flexible data structure to evaluate the pixelwise classification cost.

To make the manuscript self-contained, we give a brief introduction to the construction of cross regions, and more technical details can be found in [LSM⁺12]. An adaptive

cross support region is constructed with the following steps. First, for each pixel p , four varying support arm lengths $\{h_p^0, h_p^1, h_p^2, h_p^3\}$ are decided based on the guidance image I , which is called as a cross skeleton in [ZLL09]. An improved strategy for adaptive scale selection is proposed in [LSM⁺12]. Once such a pixelwise cross skeleton is decided, a shape-adaptive full support region Ω_p is readily available as an area integral of multiple horizontal segments $H(q)$ spanned by pixel q [LSM⁺12]. Specifically, $\Omega_p = \bigcup_{q \in V(p)} H(q)$, where q is a support pixel located on the vertical segment $V(p)$ defined for pixel p . Two example cross regions are shown in Fig. 5.2(a).

With the cross map constructed, we extract a pixelwise color histogram H_c and a texton histogram H_t from the adaptive support region around each pixel, as shown in Fig. 5.2(c). The color histogram H_c is generated by quantizing each channel of L^*a^b to 8 bins. The texton histogram H_t is generated by convolving the image with 17-dimensional filter banks at different scales, and then the responses are clustered using the Euclidean-distance K-means algorithm into $T_c = 92$ code words. Therefore, each pixel will be represented by a 116-dimensional feature vector. Based on these densely computed region-level features, we train a multi-class human classifier. An example response map (color-coded as a heat map) for the “boy-in-blue” class is shown in Fig. 5.2(e), which gives a much stronger and reliable response for those true pixels covered by the boy in blue. Such an improved pixelwise unary potential also leads to a much better identification and segmentation result as shown in Fig. 5.2(g).

5.4 Contour Based Pairwise Smoothness Potential

Conventional contour detectors typically capture part transitions by finding local extrema, which usually produce a high-recall but low-precision contour detection result. Recently, Arbelaz et al. [AMFM11] proposed a new method called *gPb*, which combines the local contour with the contour signal from eigen vectors that considers the region size and contour strength. The *gPb* method achieves the state-of-the-art contour detection results.

As the *gPb* contour map [AMFM11] provides more reliable and higher-level reasoning of salient contours, we replace the classical color contrast based pairwise potential with a *gPb* based potential $\psi_{ij}(x_i, x_j)$, which is similarly defined as in Eq. 4.6,

$$\psi_{ij}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ \omega_a(1 - \|\nabla C(i, j)\|^2) & \text{otherwise} . \end{cases} \quad (5.3)$$

where ω_a gives the weight of the pairwise potential. $\nabla C(i, j)$ measures the *gPb* signal contrast between two adjacent pixels i and j . A detail formulation of *gPb* can be found in Sec. 2.1. We observed using *gPb* as the base edge map makes the labeling snap to salient object boundaries, where an example is demonstrated in Fig. 5.3.

Figure 5.3: Comparison between the two different contrast-sensitive smoothness measures. Left column: (a) Input color image. (b) Ground-truth label map. Middle column: (c) Canny edge detection map (intensity inverted) for the input image. (d) The final labeling result based on (c). Right column: (e) gPb contour detection map (intensity inverted) for the input image. (f) The final labeling result based on (e). The white dashed circles highlight three places where using a gPb based pairwise potential yields much better labeling results than the Canny edge based potential.

5.5 Incorporating Human Detector Based Cues

The information carried by an image patch or segment by itself is often too local and hence ambiguous, which can be easily interfered when background contains similar local patterns, as it is incapable to capture the global configuration of object instances. This motivates us to address the MHIC challenge with higher and longer range grouping cues which have been proved to be useful in recent image summarization and scene understanding research [KLT09, LRKT10, TL13]. A popular approach is to reason about the objects of interest with the help from rectangular bounding boxes generated by some detection methods [FPZ07, VGVZ09, FGM10, BMBM10] or using rigid shape templates [TL13]. Both methods have limitations. Firstly, most object detectors have a little or no information of the shape that the detected objects cover, so existing methods often rely on a heuristic Grabcut approach [RKB04] to generate the shape mask. However, Grabcut itself is sensitive to the background/foreground color modeling when an image

Figure 5.4: The proposed human instance detector allows to utilize the soft segmentation mask associated with each poselet template to define a human-sensitive shape prior. (a) Input image. (b) The final detected human bounding boxes (solid boxes). They are computed by running all the poselet detectors (three example dashed boxes) at various positions and scales as in [BMBM10]. (c) The detection scores for the three example poselet detectors. (d) The human class-agnostic response map computed by fusing the soft masks transferred from all poselet templates, with the fusion weights defined by the respective poselet detection scores. (e) The human class-specific response map (four classes in this example) by further considering the detection score evaluated by our proposed multi-class color line texton classifier.

contains clutter. On the other hand, the shape template previously learned or human labeled [TL13] exhibits strong rigidity: though the high template response area can hit part of the real object, the response map produced by this approach still does not overlap with real object locations sufficiently and is not aligned to object boundaries due to the sliding window step size. This section will introduce a novel framework to generate an edge-aligned soft shape mask based on the poselet human detector [BMBM10] and a multiple color line model [WSTS07] for the MHIC task. In fact, our technique is general, and it can also be used with other state-of-the-art detectors e.g. the deformable part model [FGM10].

5.5.1 Efficient Color Line and Texton Histogram Based Poselet Human Detector

To detect humans of a certain identity with a desired degree of invariance to e.g. scale and rotation changes, we train two classifiers. One is the poselet human detector trained with the annotated H3D dataset [BMBM10], and the other is the multi-class interactive offline color and texton histogram based object detector.

The poselet detector is trained by finding patches with similar keypoint configurations in the training objects to guarantee the semantic consistency of detected parts. Each

poselet comes with a soft mask by averaging all aligned masks of training examples. In our work, a pre-trained human model and its associated soft masks from [BMBM10] is used in our implementation. Example poselet soft masks are shown in Fig. 5.4(c).

Given a user-drawn bounding box, to train the interactive offline color and texton histogram based object detector, we first adopt the method of Wei *et al.* [WSTS07] by projecting all pixel colors onto a set of one-dimensional (1D) lines in the RGB color space. These lines are evenly sampled in 13 directions which pass through (128, 128, 128) and then a 1D (normalized) histogram of the projected values is calculated on each line. Through an empirical comparison, we use eight bins for each line and treat all 13×8 color bins as the final color line feature, which can be efficiently extracted by using integral histogram [Por05]. To better handle background clutter, we additionally extract a texton histogram for each bounding box as the texture features, whose dimension is the same as the histogram used in Sec. 5.3, and the final feature dimension used for training is 196 dimensions. For the multiple foreground recognition problem we consider here, we use the JointBoosting algorithm [TMF04] rather than Adaboosting used in [WSTS07] to train a multi-class bounding box classifier. The details of our learning procedure resemble closely with those described in [WSTS07]. Similar to [WSTS07], the positive training samples are provided by the user annotated bounding boxes with multi-class labels. To generate more positive examples and also be robust to variations across images, the same appearance perturbation scheme [WSTS07] is employed, which perturbs the position and lighting scale of the object rectangles randomly by a small amount. The negative examples are randomly sampled around the non-selected foreground regions using the bounding boxes of the same size as the user-specified ones, also with simulated scale and lighting variations.

The detection proposals are generated by first sweeping different poselet templates in the test image and then the activations are merged into the final bounding boxes as in [BMBM10]. Then each bounding box is evaluated by both the trained multi-class color line texton classifier and poselet detector, the corresponding detection scores s_c^l and s_p are linearly combined using an empirically set trade-off factor $\varepsilon = 0.1$ to form the final score $R_d = (1 - \varepsilon) \times s_c^l + \varepsilon \times s_p$. Only the top 100 scored bounding boxes for each class l are retained, which form the final bounding boxes hypothesis set D in Eq. (5.1).

5.5.2 Edge-Aware Soft Shape Map with GGDT and JBF

As rigid shape templates have proved to greatly improve the performance of scene understanding tasks, it is interesting to explore whether a shape template can improve the more challenging MHIC task concerning non-rigid humans. Since the training dataset provided by [BMBM10] includes foreground/background annotation, each poselet template t comes with a soft mask $M_t \in [0, 1]$ by averaging the binary segmentation annotations among all

Figure 5.5: Effects of the soft shape cue and the higher order object detector potential. (a) Input images. (b) Detected bounding boxes for corresponding human classes. (c) The initial transferred poselet masks. (d) The enhanced shape masks using generalized geodesic distance transform [CSRP10]. (e) The final soft shape maps generated by using an edge-aware joint bilateral filter [ABD10] to filter (d). (f) The labeling results without using soft shape cues. (g) The labeling results by combining shape cues and the cross region based unary term. (h) The results using higher order detector potentials. (i) The ground-truth label maps. By incrementally including the shape potential and the higher order detector potential, one can observe the improvements in labeling accuracy.

example patches used for training the respective template. Therefore, for each bounding box from the top 100 scored ones for each class, we overlay the soft masks of the merged poselets on the test image, which are weighted by their corresponding poselet template detection score p_t and then further weighted by their corresponding bounding box score R_d . Iterating over all the detected top human bounding boxes, this process produces a pixelwise soft belief map $M^l : \Omega \rightarrow [0, 1]$ for each human class l , where Ω is the discrete image 2D domain. Fig. 5.4 illustrates the process how the soft masks are generated. Fig. 5.5(c) shows the rough hit maps of different human classes.

The belief map M^l (see Fig. 5.5(c)) for each class l is still blurry and contains some mistakes caused by sliding window offsets and false detection, despite that the proposed detector succeeds in eliminating most non-object regions. Based on this observation, we propose to use the color image's internal structure to refine and enhance the initial belief map. Another prior about humans and other objects is that they are often compactly clustered in space instead of being distributed around. In this work, we apply two efficient approaches to enhance the derived belief map M , and we explain them next.

Given two points sharing similar appearance, if the shortest path connecting them has a high geodesic length (e.g. it cuts through a high image gradient), this gives a strong hint that they are likely to belong to different regions or objects. However, exactly evaluating the geodesic distance between a pair of pixels is computationally expensive. Fortunately, this dense pairwise geodesic distance can be approximately estimated by the generalized geodesic distance transform method (GGDT) proposed by Criminisi *et al.* [CSRP10]. The GGDT method assigns each pixel a geodesic distance $\mathcal{Q}$ to the soft belief map M , given a gray-valued image J :

$$\mathcal{Q}(\mathbf{m}; M, \nabla J) = \min_{\mathbf{m}' \in \Omega} (\delta(\mathbf{m}, \mathbf{m}') + \nu M(\mathbf{m}')) , \quad (5.4)$$

where Ω is the image 2D domain, and ν is an amplification parameter. The geodesic distance between pixels $\mathbf{m}$ and $\mathbf{n}$ is given as:

$$\delta(\mathbf{m}, \mathbf{n}) = \inf_{\Gamma \in \mathcal{P}_{\mathbf{m}, \mathbf{n}}} \int_0^{l(\Gamma)} \sqrt{1 + \gamma^2 (\nabla J(s) \cdot \Gamma'(s))^2} ds , \quad (5.5)$$

where Γ is a sub-path of all the paths $\mathcal{P}_{\mathbf{m}, \mathbf{n}}$ connecting two points $\mathbf{m}$ and $\mathbf{n}$. The parameter γ controls the relative importance of the spatial distance to the image gradient.

By applying the GGDT to the initial belief map M , this weak signal is clearly enhanced, as shown in Fig. 5.5(d), because the detected human instance probabilities are propagated along pairwise shortest geodesic distances. Despite of this improvement, the obtained belief map is still spatially inaccurate with respect to the true object locations.

To produce an edge-aware soft belief map, we propose to use an efficient high dimensional joint bilateral filter [ABD10] to filter the result produced by GGDT to yield the final soft spatial mask, see Fig. 5.5(e). This result highlights the pixelwise extent of objects. In our current formulation in Eq. (5.1), we choose to integrate the soft shape mask as an additional unary potential $\psi_i^s(x_i)$ to compete with other hypothesis evaluations. As the belief map is produced from shapes, it is complementary to the color and texton based cues. Denoting by M_s the final soft shape maps for all the intended human classes, we introduce and compute the shape potential as follows,

$$\psi_i^s(x_i) = -\omega_{shape} \log P(x_i|M_s) , \quad (5.6)$$

where ω_{shape} is the weighting factor. With the help of such a soft shape term, the segmentation result is significantly improved. Fig. 5.5(f) and Fig. 5.5(g) provide some visual comparison between the results with and without the shape cue.

5.5.3 Detector-Based Robust Consistency Potential

Although the proposed shape cue can greatly improve the result, to make our system more robust and allow to detect instances of certain humans, we further incorporate higher order label consistency constraints from the detected bounding boxes. With the help of such a higher order constraint, we can revolve ambiguities which would otherwise be too hard to solve at a local level. The bounding box proposals are designed as a kind of soft constraint which works jointly with other hypotheses to overcome false positives and over-counting of object instances. Given the d -th detection bounding box $\mathbf{x}_d$ with a confidence score R_d belonging to a certain class as presented in Sec. 5.5.1, $\psi_{detector}(\mathbf{x}_d)$ is similarly defined as Eq. 4.3 of Chapter. 4:

$$\psi_{detector}(\mathbf{x}_d) = \begin{cases} N_d \frac{1}{Q_d} \gamma_{max} & \text{if } N_d \leq Q_d \\ \gamma_{max} & \text{otherwise ,} \end{cases} \quad (5.7)$$

where $N_d = \sum_{i \in \mathbf{x}_d} \delta(x_i \neq l_d)$ is the number of variables in $\mathbf{x}_d$ not taking the dominant label l_d . The truncation parameter Q_d controls the maximum number of inconsistent pixels. The cost γ_{max} is defined by a linear truncated function $f(\cdot)$, and it monotonically increases with the object classifier response R_d as

$$f(\mathbf{x}_d, R_d) = \omega_d \varepsilon_d |\mathbf{x}_d| \max(0, R_d - R_t) , \quad (5.8)$$

where R_t is a threshold. In Eq. (5.8), ω_d defines the detector potential weight, and ε_d is the aggregation of the weights w_p^i of the inconsistent pixels i in $\mathbf{x}_d$. This detector-based label consistency constraint is similar to the object detector term used in [LSA⁺10] for scene understanding. If a detector response is strong, the higher-order potential will encourage

the pixels belonging to the bounding box $\mathbf{x}_d$ to take the dominant bounding box label l_d . As the penalty is increased with the number of inconsistent pixels incrementally until the truncation threshold Q_d , this soft higher-order constraint produces better labeling results than the standard P^n Potts model [KKT09], which forbids other differently labeled pixels within the clique $\mathbf{x}_d$. The proposed object potential $\psi_d(\mathbf{x}_d)$ can be transformed to take the robust P^n form [KLT09, LSA⁺10]:

$$\psi_d(\mathbf{x}_d) = -f(\mathbf{x}_d, R_d) + \min(f(\mathbf{x}_d, R_d), k_d \cdot \sum_{i \in \mathbf{x}_d} w_p^i \delta(x_i \neq l_d)) , \quad (5.9)$$

where k_d is a slope parameter defined in the same way as in [LSA⁺10], and w_p^i is the same as in ε_d of Eq. (5.8).

In Eq. 4.3 of Chapter 4, the weight w_p^i is assumed uniform across all the pixels within the bounding box. This uniform setting may produce visual artifacts and increase the inference cost, because it does not model the spatial extent of the underlying object and cannot treat the foreground human and the background differently. Given the soft shape maps previously produced, we can reuse the maps to impose a spatial weighting to bias the graph cut to expand the label where it is more likely to be the potential human. This basic idea is to sum up the background belief maps of all the human instances $M_{bg} = \sum M_{bg}^l, l \in L$, and then use the pixelwise weighting map $M_w = 1 - M_{bg}/L$ to replace the constant w_p^i .

Including the detector term to a CRF model is implemented by adding two auxiliary nodes into the graph, and the augmented energy function can be efficiently minimized with the graph cut algorithms, a brief introduction can be found in Sec. 2.2.2.1. Interested readers are referred to [KLT09] for the graph optimization details. Fig. 5.5(h) demonstrates the strength of the object detector-based potential when integrated into our CRF framework. Without using the detector-based potential, the girl's face in Fig. 5.5(g) can only be partly segmented due to the competition between the hypotheses on pixels and shapes. The object detector potential provides a complementary high-level evidence, and integrating it into the CRF model results in a more accurate result of recognizing the missing face part.

5.6 The NTU-MHIC Dataset

To evaluate our proposed approach and to establish a new benchmark for future work, we introduce the first and the largest multiple human foreground cosegmentation and identification dataset – the NTU-MHIC dataset. In fact, till now there exists no benchmark dataset to evaluate a method's performance for the MHIC task. The FlickrMFC dataset [KX12] contains some subsets with human classes, but some of them are not very

Figure 5.6: Ground-truth annotation for the NTU-MHIC dataset. (a) Input image. (b) A user segments out the first human (the green contour) and indicates him as the human of interest by placing a red stroke. (c) The generated ground-truth label for the first human class. (d) The second human in the image is annotated following the same procedure. The whole process to annotate an image takes $3 \sim 4$ minutes.

challenging for the MHIC task. In addition, other object classes are also mixed with the human classes in this dataset. The CoDel dataset [SLJ13] is intended for a multiple foreground human detection task by sampling representative video frames in “Big Bang Theory”, but it only comes with bounding box labels.

As one of our main contributions, we create a new dataset by collecting image subsets which match the MHIC scenario from both the FlickrMFC and CoDel datasets. We manually annotate the images with pixelwise class labels, which serve as the ground-truth to evaluate various MHIC algorithms. The obtained NTU-MHIC dataset contains 13 subsets with a total of 192 images, and each subset includes around $10 \sim 20$ images for the same event. This dataset is challenging, because it contains both indoor and outdoor human activities (e.g. sports, child play, and group chat) with large viewpoint change and displacement, background/lighting variations and occlusion, and also only a finite number of repeating subjects are present in each image (see Fig. 5.8~ 5.10). Figure 5.7 illustrates the statistics of the percentage of the images that contain $0 \sim 5$ human instances in each subset of the NTU-MHIC dataset. For different image subsets, there is clearly a high diversity in terms of the distribution of the human instance number per image. For those

Figure 5.7: The percentage of the images containing 0 ~ 5 human instances in each subset of the NTU-MHIC dataset.

Figure 5.8: Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each row presents input images, their corresponding ground-truth, color-labeled segmentation results for supervised MFC [KX12], MFRC [ZLC⁺14] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.

subsets sampled from the CoDel dataset [SLJ13], more images containing at least 2 or more human instances are observed.

Dataset Annotation: The ground-truth pixelwise labels for each image in the dataset are generated by an annotator using a labeling tool. This human labeler first needs to delineate a closed contour around the object-of-interest, and then a foreground stroke is marked over the object to conduct a final object cutout. The labeling tool allows the user to add/delete a stroke to further refine the annotation. Fig. 5.6 shows the labeling process. The average time to annotate an image is around 3 ~ 4 minutes. We will make this dataset and pixelwise annotations publicly available to facilitate future work.

Figure 5.9: Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each row presents input images, their corresponding ground-truth, color-labeled segmentation results for supervised MFC [KX12], MFRC [ZLC⁺14] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.

Figure 5.10: Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each row presents input images, their corresponding ground-truth, color-labeled segmentation results for supervised MFC [KX12], MFRC [ZLC⁺14] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.

Figure 5.11: Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each row presents input images, their corresponding ground-truth, color-labeled segmentation results for supervised MFC [KX12], MFRC [ZLC⁺14] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.

Figure 5.12: Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each row presents input images, their corresponding ground-truth, color-labeled segmentation results for supervised MFC [KX12], MFRC [ZLC⁺14] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.

Figure 5.13: Some randomly drawn examples from seven groups of the MHIC dataset. From top to bottom, each row presents input images, their corresponding ground-truth, color-labeled segmentation results for supervised MFC [KX12], MFRC [ZLC⁺14] and our MHIC method. The colored tag below each set indicates which category each region is assigned to.

Figure 5.14: Segmentation accuracy comparison between our method (MHIC) and other baselines (MFRC [ZLC⁺14], MFC [KX12], CoS [KXLK11], and DC [JBP10]) for the MHIC dataset. The S and U denote whether the method is supervised or unsupervised.

5.7 Experimental Results and Discussions

We set the parameters in our algorithm empirically and fix them across all the tests as: $\omega_{pix} = 1, \omega_a = 10, \omega_s = 0.2, \omega_d = 0.3, R_t = 0.3, \rho_{max} = -\log(0.8)$. The overall time (including preprocessing, detection and segmentation) to process an image is around 10 ~ 20 seconds on a laptop with Intel Core i7 Q740 1.7GHZ and 22GB RAM.

Quantitative Results: We first compare our method with some baselines: MFRC [ZLC⁺14], MFC [KX12], CoSand (COS) [KXLK11], and Discriminative Clustering (DC) [JBP10]. We adopt the procedure introduced in MFC [KX12] for evaluation. For supervised methods such as our method, MFRC, and MFC’s supervised version (MFC-S), we randomly pick 20% of the input images to annotate. For unsupervised MFC (MFC-U), we run them by changing the foreground number K from two to eight, and report the best scores for each subset. For the unsupervised binary class methods COS [KXLK11] and DC [JBP10], the dataset is divided into several subgroups such that the images in each subgroup contain the same objects of interest, and the methods are applied to each subgroup individually.

Fig. 5.14 summarizes the segmentation accuracy on the 13 groups of the MHIC dataset. We evaluate the segmentation accuracy by the standard intersection-over-union metric $\frac{(GT_i \cap R_i)}{(GT_i \cup R_i)}$. The leftmost bar set presents the average segmentation accuracy on 13 groups. Some interesting results can be observed from the chart: 1) For unsupervised methods, COS and DC’s accuracies are better than the unsupervised version of MFC. The underlying reason is that DC and COS enforce a strong assumption that the set of images provided must have one common foreground in each image. 2) The methods with some supervision from a human, such as supervised MFC, MFRC and our proposed method are better than unsupervised methods in most of the cases, which on one hand proves that supervision can be beneficial for the task. 3) The higher accuracy of MFRC and our newly proposed method in comparison with supervised MFC demonstrates the clear benefits of explicit modeling of the objectness constraint. On average, MFRC and MHIC achieves 16% and 28% improvements over the supervised-MFC (MFC-S). In some

subset, such as *bb6* (where *bb* stands for “Big Bang Theory”), we achieve nearly 45% improvement!.

We have also evaluated the average accuracy gain contributed by including the proposed shape and higher-order detector potentials into the CRF model in Table 5.1, and also compared these models with our previous MFRC method [ZLC⁺14]. By using a cross region-based unary potential alone, we achieve comparable performance with our previously proposed MFRC, which includes more complex higher order superpixel potentials with higher inference costs. The incorporation of the shape cue contribute nearly 4% improvement in accuracy, while including the higher-order detector term brings an additional 3% improvement. The small numerical gain by the higher order detector term has also been observed in Shotton *et al.* [SWRC09] and Pushmeet *et al.* [KLT09] in scene understanding research. As also indicated in [SWRC09, KLT09], we observe that including this potential often brings a pronounced increase in perceived accuracy in boundary areas. In addition, we should also notice that introducing the shape cue occasionally reduce the accuracy, but this only happens for one subset *bb6*, and recovered by using higher order term. One potential reason for such a phenomenon is the shape potential is derived from a wrong human detection result, which is not consistent with other hypotheses, so our method chooses to label it as the background. Further human intervention should improve the result, which is left as one future direction.

Visual Results: Fig. 5.8~ 5.13 show some visual results from seven groups of the MHIC dataset. For each set, the input images and color-coded segmentation results are displayed in the first two rows. We also show visual results from supervised MFC (MFC-S) [KX12], MFRC [ZLC⁺14] and our proposed MHIC method. The regions which are labeled with the same color in each set indicate they belong to the same category. The tags below each set explain the meaning of each color. From these images, one can observe that MFC-S produced quite obvious segmentation and human identification errors. Its reliance on a coarse superpixel segmentation prevents it from correcting the errors made in the initial superpixel generation process. Without including human-specific modeling (or potentials), both MFC-S and MFRC frequently misclassified some non-human regions as human instances, though MFRC performs much better than MFC-S due to the inclusion of object notions/terms in the CRF model. Thanks to the novel human-centric cues as well as our region-based unary terms, MHIC achieves superior human segmentation and identification quality over other competing methods for this challenging MHIC dataset. Our method can handle irregularly appearing humans and produce smooth and more accurate segmentation results. The images with no foregrounds can also be identified, e.g., the *baseball* dataset. On the other hand, our current model still cannot very well handle the cases of humans in small size or from a highly profiled view, or having a great appearance overlap between foreground and background, such as the *bb3* and *bb5* image sets. This remains as a future research direction.

Table 5.1: Component evaluation of the proposed MHIC algorithm (column 3–5) in comparison with MFRC [ZLC⁺14].

Subset	MFRC ($\times 100\%$)	Unary + Pairwise ($\times 100\%$)	Unary + Pairwise + Shape ($\times 100\%$)	Full Model ($\times 100\%$)
apple	0.50	0.58	0.61	0.64
baseball	0.52	0.5	0.61	0.62
cow	0.66	0.65	0.66	0.67
bb1	0.51	0.59	0.67	0.73
bb2	0.45	0.37	0.4	0.41
bb3	0.48	0.56	0.58	0.64
bb4	0.55	0.6	0.62	0.64
bb5	0.38	0.42	0.53	0.57
bb6	0.53	0.63	0.59	0.64
bb7	0.43	0.59	0.62	0.62
bb8	0.37	0.45	0.48	0.55
bb9	0.41	0.42	0.44	0.51
bb10	0.5	0.53	0.57	0.62
average precision	0.48	0.53	0.57	0.60

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Image segmentation/cosegmentation will play more and more important roles in high level computer vision and machine learning tasks. In previous chapters, we conduct literature surveys on existing methods, and then propose to solve three open-ended challenges: one is to segment a single image into a small number of regions corresponding to object or parts; another is to jointly localize, recognize and segment multiple irregularly occurred object from the user photo stream; the third one considers a more challenging problem to jointly localize, identify and segment human objects from a photo album, which exhibit more deformations than other objects.

For the first challenge, we propose to construct spectral attributes, color Gaussian Mixture Models and geodesic distance from the state-of-the-art techniques as selected low level cues and elaborately integrated them to form feature vectors for image pixels. These feature vectors encode global color and spatial features as well as global structure information. Then, an automatic multi-label segmentation algorithm based on the Potts variational model and the stability of $Ncut$ value using the feature vectors is developed, which can produce a small number of regions reflecting local color and spatial coherence and global semantic structure in one framework.

For the second challenge, we propose to solve the multiple foreground recognition and cosegmentation (MFRC) problem within a conditional random fields (CRFs) framework in a principled manner. Our objective is to segment out and annotate the foreground objects or the “things” rather than the “stuff”. We exploit a few complementary objectness cues (e.g. contours, object detectors and layout) and propose novel and efficient methods to capture object-level information. Integrating object potentials as soft constraints (e.g. robust higher-order potentials defined over detected object regions) with the low-level unary and pairwise terms holistically, we solve the MFRC task with a probabilistic CRF model. With a minimal amount of user annotations on just a few example photos,

the proposed approach produces spatially coherent, boundary-aligned segmentation results with correct and consistent object labeling, which proves the advantage of explicit modeling of the “object” notion in MFRC problem.

For the third challenge, we propose to solve the challenging MHIC problem within a principled CRF framework for the first time, where only a small fraction of input images are weakly labeled with bounding boxes. We propose an efficient human instance detector by combining an extended color line model [WSTS07] with the poselet detector [BMBM10]. We propose to use adaptive cross regions [LSM⁺12] as basic spatial supports to evaluate features of each pixel, therefore providing more discriminative and robust pixelwise features. We propose an effective high-level human shape cue generated by applying in sequence the geodesic distance transform [CSRP10] and joint bilateral filtering [ABD10] to the initial human instance detection response map. The shape cue is further used to weight the detector’s higher-order potential, which produces better results in terms of quantitative measure and visual quality. We create a novel *NTU-MHIC* dataset to facilitate benchmarking the performance of various algorithms on the MHIC task. This dataset consists of nearly two hundred images with ground-truth human instance labeling, featuring multiple human instances in the MHIC scenario with various poses and scale change.

6.2 Future Directions

In this section, we list some potential directions for future research.

6.2.1 Fine Grained Object Segmentation and Cosegmentation

Existing methods label the whole object as one label, while ignoring the fine grained part information. For example, given test images in MHIC task, our method in Chapter 5 can localize, identify and segment the human instances from the photo stream. However, there is little way to retrieve the human parts after parsing. This applies to the other objects which contain distinctive parts. Retrieving part information can facilitate a lot of graphics and vision applications: people can edit part of the segmented instances and then propagate the editing to other images. Training part based detector can alleviate part of the problem, however it requires large number of training data and is not scalable when large number of object classes are available. Therefore, class agnostic part detector is a highly desirable tool for the tasks. The method in Chapter 3 provide one option worth for further exploration, however as indicated in the chapter, it can ignore small parts. Some of our recent research discovers that we can retrieve object/object parts with some simple signal processing techniques. It is interesting to explore further along this direction.

6.2.2 Automatic Multiple Foreground Recognition and Cosegmentation in Video

Many cameras now equip user with the capability of shooting short videos. One natural extension of Chapter 4 and Chapter 5 is to jointly localize, identify and segment the multiple foregrounds in video. However, directly extending the methods in previous chapters to video is not feasible, which is mainly due to the divergent purpose when people take photos and videos. When people take photos, they usually focus on the most informative content; when taking videos, there are more noisy content inside. On the other hand, a video has one natural advantage over an image is that it records more sequential information which is not available in the image. Thus developing holistic method which can automatically eliminate the noisy information, while simultaneously localize the object of interest without user interaction and make full use of the video's sequential characteristics in one principled framework can be highly desirable.

Publication List:

- **Hongyuan Zhu**, Fanman Meng, Jianfei Cai, Shijian Lu: Beyond Pixels: A Comprehensive Survey from Bottom-up to Semantic Image Segmentation and Cosegmentation, Submitted to JVCI and Under Review
- **Hongyuan Zhu**, Jiangbo Lu, Jianfei Cai, Jianmin Zheng, and Nadia Magnenat-Thalmann: Multiple Human Identification and Cosegmentation: A Human-Oriented CRF Approach with Poselets, Under Revision and To Re-Submit
- **Hongyuan Zhu**, Jiangbo Lu, Jianfei Cai, Jianmin Zheng, and Nadia Magnenat-Thalmann: Poselet-Based Multiple Human Identification and Cosegmentation, ICIP 2014.
- **Hongyuan Zhu**, Jiangbo Lu, Jianfei Cai, Jianmin Zheng, Nadia Magnenat-Thalmann: Multiple foreground recognition and cosegmentation: An object-oriented CRF model with robust higher-order potentials. WACV 2014
- **Hongyuan Zhu**, Jianmin Zheng, Jianfei Cai, Nadia Magnenat-Thalmann: Object-Level Image Segmentation Using Low Level Cues. IEEE Transactions on Image Processing 22(10): 4019-4027 (2013)
- **Hongyuan Zhu**, Jianfei Cai, Jianmin Zheng, Jianxin Wu, Nadia Magnenat-Thalmann: Salient object cutout using Google images. ISCAS 2013
- Anran Wang, **Hongyuan Zhu**, Jianfei Cai, Jianxin Wu: Multi-class Cosegmentation with Pairwise Active Learning. PCM 2013

References

- [ABD10] Andrew Adams, Jongmin Baek, and Myers Abraham Davis. Fast high-dimensional filtering using the permutohedral lattice. *Comput. Graph. Forum*, 29(2), 2010.
- [AcLGS07] Dragomir Anguelov, Kuang chih Lee, Salih Burak Göktürk, and Baris Sumengen. Contextual identity recognition in personal photo albums. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [ACZZ12] Nguyen Thi Nhat Anh, Jianfei Cai, Juyong Zhang, and Jianmin Zheng. Robust interactive image segmentation using convex active contours. *IEEE Transactions on Image Processing*, 21(8), 2012.
- [ADF10] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [AG97] Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7), 1997.
- [AMFM09] Pablo Arbelaez, Michael Maire, Charless C. Fowlkes, and Jitendra Malik. From contours to regions: An empirical evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2294–2301, 2009.

REFERENCES

- [AMFM11] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5), 2011.
- [ASS⁺12] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11), 2012.
- [Bea06] Richard Beare. A locally constrained watershed transform. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(7):1063–1074, 2006.
- [Bes86] J. Besag. On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society*, 48(3):259–302, 1986.
- [BF70] Claude R. Brice and Claude L. Fennema. Scene analysis using regions. *Artif. Intell.*, 1(3):205–226, 1970.
- [BJ01] Yuri Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *IEEE International Conference on Computer Vision*, pages 105–112, 2001.
- [BK04] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1124–1137, 2004.
- [BKP⁺10] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

REFERENCES

- [BKP⁺11] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. Interactively co-segmentating topically related images with intelligent scribble guidance. *International Journal of Computer Vision*, 93(3), 2011.
- [BMBM10] Lubomir D. Bourdev, Subhransu Maji, Thomas Brox, and Jitendra Malik. Detecting people using mutually consistent poselet activations. In *European Conference on Computer Vision*, 2010.
- [BS07] Xue Bai and Guillermo Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *IEEE International Conference on Computer Vision*, 2007.
- [BVZ01] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11), 2001.
- [BYT11] Egil Bae, Jing Yuan, and Xue-Cheng Tai. Global minimization for continuous multiphase partitioning problems using a dual approach. *International Journal of Computer Vision*, 92(1):112–129, 2011.
- [BYTB14] Egil Bae, Jing Yuan, Xue Cheng Tai, and Yuri Boykov. A fast continuous max-flow approach to non-convex multilabeling problems. *Efficient Algorithms for Global Optimization Methods in Computer Vision*, 2014.
- [CBGM02] Chad Carson, Serge Belongie, Hayit Greenspan, and Jitendra Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(8):1026–1038, 2002.

REFERENCES

- [CBS05] Timothée Cour, Florence Bénézit, and Jianbo Shi. Spectral segmentation with multiscale graph decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
- [Cha04] Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1-2):89–97, 2004.
- [CLL11] Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [CM02] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002.
- [CS07] Timothée Cour and Jianbo Shi. Recognizing objects by piecing together the segmentation puzzle. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [CS12] João Carreira and Cristian Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7), 2012.
- [CSRP10] Antonio Criminisi, Toby Sharp, Carsten Rother, and Patrick Pérez. Geodesic image and video editing. *ACM Trans. Graph.*, 29(5), 2010.
- [CV01] Tony F. Chan and Luminita A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001.

REFERENCES

- [CXGS12] Maxwell D. Collins, Jia Xu, Leo Grady, and Vikas Singh. Random walks based multi-image segmentation: Quasiconvexity results and gpu-based solutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [DDS⁺09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [EH14] Ian Endres and Derek Hoiem. Category-independent object proposals with diverse ranking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(2), 2014.
- [FGM10] Pedro F. Felzenszwalb, Ross B. Girshick, and David A. McAllester. Cascade object detection with deformable part models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [FH04] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.
- [FPZ07] Robert Fergus, Pietro Perona, and Andrew Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *International Journal of Computer Vision*, 71(3), 2007.
- [GFK09] Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *IEEE International Conference on Computer Vision*, 2009.

REFERENCES

- [GG84] Stuart Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, Nov 1984.
- [GLAM09] Chunhui Gu, Joseph J. Lim, Pablo Arbelaez, and Jitendra Malik. Recognition using regions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [Gra06] Leo Grady. Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(11):1768–1783, November 2006.
- [GSRS11] Rahul Garg, Steven M. Seitz, Deva Ramanan, and Noah Snavely. Where’s waldo: Matching people in images of crowds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [HEH08] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Closing the loop in scene interpretation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [HS97] Donald D Hoffman and Manish Singh. Saliency of visual parts. *Cognition*, 63(1):29 – 78, 1997.
- [HYK13] Xiaodi Hou, Alan L. Yuille, and Christof Koch. Boundary detection benchmarking: Beyond f-measures. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
- [JBP10] Armand Joulin, Francis R. Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

REFERENCES

- [JBP12] Armand Joulin, Francis Bach, and Jean Ponce. Multi-class cosegmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [JTLB04] Anil K. Jain, Alexander P. Topchy, Martin H. C. Law, and Joachim M. Buhmann. Landscape of clustering algorithms. In *International Conference on Pattern Recognition*, 2004.
- [KF12] Daniel Küttel and Vittorio Ferrari. Figure-ground segmentation by transferring window masks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [KGF12] Daniel Küttel, Matthieu Guillaumin, and Vittorio Ferrari. Segmentation propagation in imagenet. In *European Conference on Computer Vision*, 2012.
- [KK11] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Conference on Neural Information Processing Systems*, 2011.
- [KKT09] Pushmeet Kohli, M. Pawan Kumar, and Philip H. S. Torr. P & beyond: Move making algorithms for solving higher order functions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(9), 2009.
- [KLT09] Pushmeet Kohli, Lubor Ladicky, and Philip H. S. Torr. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3), 2009.
- [Kol06] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10), 2006.

REFERENCES

- [KT07a] Pushmeet Kohli and Philip H. S. Torr. Dynamic graph cuts for efficient inference in markov random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2079–2088, 2007.
- [KT07b] Nikos Komodakis and Georgios Tziritas. Approximate labeling via graph cuts based on linear programming. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(8), 2007.
- [KX12] Gunhee Kim and Eric P. Xing. On multiple foreground cosegmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [KXLK11] Gunhee Kim, Eric P. Xing, Fei-Fei Li, and Takeo Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *IEEE International Conference on Computer Vision*, 2011.
- [KZ04] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):147–159, 2004.
- [LG10] Yong Jae Lee and Kristen Grauman. Collect-cut: Segmentation with top-down cues discovered in multi-object images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [LJ08] Diane Larlus and Frédéric Jurie. Combining appearance models and markov random fields for category level object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [LKG11] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *IEEE International Conference on Computer Vision*, 2011.

- [LKY⁺09] Jan Lellmann, Jörg H. Kappes, Jing Yuan, Florian Becker, and Christoph Schnörr. Convex multi-class image labeling by simplex-constrained total variation. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 150–162, 2009.
- [LRKT10] Lubor Ladicky, Christopher Russell, Pushmeet Kohli, and Philip H. S. Torr. Graph cut based inference with co-occurrence statistics. In *European Conference on Computer Vision*, 2010.
- [LSA⁺10] Lubor Ladicky, Paul Sturges, Kartee Alahari, Christopher Russell, and Philip H. S. Torr. What, where and how many? combining object detectors and crfs. In *European Conference on Computer Vision*, 2010.
- [LSM⁺12] Jiangbo Lu, Keyang Shi, Dongbo Min, Liang Lin, and Minh N. Do. Cross-based local multipoint filtering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [LTZ13] Lubor Ladicky, Philip H. S. Torr, and Andrew Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
- [LYMD13] Jiangbo Lu, Hongsheng Yang, Dongbo Min, and Minh N. Do. Patch match filter: Efficient edge-aware filtering meets randomized search for fast correspondence field estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
- [LYT11] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12), 2011.

REFERENCES

- [MBSL99] Jitendra Malik, Serge Belongie, Jianbo Shi, and Thomas K. Leung. Textons, contours and regions: Cue integration in image segmentation. In *IEEE International Conference on Computer Vision*, 1999.
- [MFM04] David R. Martin, Charless Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(5):530–549, 2004.
- [MGE11] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *IEEE International Conference on Computer Vision*, 2011.
- [ML13] Tianyang Ma and Longin Jan Latecki. Graph transduction learning with connectivity constraints with application to multiple foreground cosegmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
- [MLLN12] Fanman Meng, Hongliang Li, Guanghui Liu, and King Ngi Ngan. Object cosegmentation based on shortest path algorithm and saliency model. *IEEE Transactions on Multimedia*, 14(5), 2012.
- [Mor05] Greg Mori. Guiding model search using segmentation. In *IEEE International Conference on Computer Vision*, 2005.
- [MS89] David Mumford and Jayant Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42(5):577–685, 1989.

REFERENCES

- [MSD09] Lopamudra Mukherjee, Vikas Singh, and Chuck R. Dyer. Half-integrality based algorithms for cosegmentation of images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2028–2035, 2009.
- [MSP11] Lopamudra Mukherjee, Vikas Singh, and Jiming Peng. Scale invariant cosegmentation for image groups. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [NJW01] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Conference on Neural Information Processing Systems*, pages 849–856, 2001.
- [OPR78] R. Ohlander, K. Price, and D. R. Reddy. Picture segmentation using a recursive region splitting method. *Computer Graphics and Image Processing*, 8:313–333, 1978.
- [OS88] Stanley Osher and James A. Sethian. Fronts propagating with curvature dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, 1988.
- [OTS79] A threshold selection method from gray-level histograms. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(1):62–66, Jan 1979.
- [PCCB09] Thomas Pock, Antonin Chambolle, Daniel Cremers, and Horst Bischof. A convex relaxation approach for computing minimal partitions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [Por05] Fatih Murat Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

REFERENCES

- [PSG⁺08] Thomas Pock, Thomas Schoenemann, Gottfried Graber, Horst Bischof, and Daniel Cremers. A convex formulation of continuous multi-label problems. In *European Conference on Computer Vision*, pages 792–805, 2008.
- [RFE⁺06] Bryan C. Russell, William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
- [RKB04] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. ”grabcut”: interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.
- [RMBK06] Carsten Rother, Thomas P. Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
- [RSLP12] José C. Rubio, Joan Serrat, Antonio M. López, and Nikos Paragios. Unsupervised co-segmentation through region matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [SEZ09] Josef Sivic, Mark Everingham, and Andrew Zisserman. ”who are you?” - learning person specific classifiers from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [SJC08] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

REFERENCES

- [SL06] Yang Song and Thomas Leung. Context-aided human recognition - clustering. In *European Conference on Computer Vision*, 2006.
- [SLJ13] Jianping Shi, Renjie Liao, and Jiaya Jia. Codel: An human co-detection and labeling framework. In *IEEE International Conference on Computer Vision*, 2013.
- [SM97] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1997.
- [SM00] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.
- [SPSK13] Kartic Subr, Sylvain Paris, Cyril Soler, and Jan Kautz. Accurate binary image selection from inaccurate user input. *Comput. Graph. Forum*, 32(2), 2013.
- [SWLL13] Keyang Shi, Keze Wang, Jiangbo Lu, and Liang Lin. Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
- [SWRC09] Jamie Shotton, John M. Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1), 2009.
- [Sze11] Richard Szeliski. *Computer Vision - Algorithms and Applications*. Texts in Computer Science. Springer, 2011.

REFERENCES

- [SZS06] Josef Sivic, C. Lawrence Zitnick, and Richard Szeliski. Finding people in repeated shots of the same scene. In *British Machine Vision Conference*, 2006.
- [TBS12] Makarand Tapaswi, Martin Bäumel, and Rainer Stiefelhagen. "knock! knock! who is it?" probabilistic person identification in tv-series. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [TL13] Joseph Tighe and Svetlana Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
- [TMF04] Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Sharing features: Efficient boosting procedures for multiclass object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.
- [VFB11] Alexander Vezhnevets, Vittorio Ferrari, and Joachim M. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *IEEE International Conference on Computer Vision*, 2011.
- [VGZ09] Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple kernels for object detection. In *IEEE International Conference on Computer Vision*, 2009.
- [vL07] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
- [VRK11] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. Object cosegmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

REFERENCES

- [VS91] Luc Vincent and Pierre Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(6):583–598, 1991.
- [VWLT11] Vibhav Vineet, Jonathan Warrell, Lubor Ladicky, and Philip H. S. Torr. Human instance segmentation from video using detector-based conditional random fields. In *British Machine Vision Conference*, 2011.
- [WSTS07] Yichen Wei, Jian Sun, Xiaoou Tang, and Heung-Yeung Shum. Interactive offline tracking for color objects. In *IEEE International Conference on Computer Vision*, 2007.
- [YBT10] Jing Yuan, Egil Bae, and Xue-Cheng Tai. A study on continuous max-flow and min-cut approaches. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [YBTB10] Jing Yuan, Egil Bae, Xue-Cheng Tai, and Yuri Boykov. A continuous max-flow approach to potts model. In *European Conference on Computer Vision*, pages 379–392, 2010.
- [YCZL10] Wenxian Yang, Jianfei Cai, Jianmin Zheng, and Jiebo Luo. User-friendly interactive image segmentation through unified combinatorial user inputs. *IEEE Transactions on Image Processing*, 19(9):2470–2479, 2010.
- [YR13] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12), 2013.
- [ZCLZ03] Lei Zhang, Longbin Chen, Mingjing Li, and Hongjiang Zhang. Automated annotation of human faces in family albums. In *ACM Multimedia*, 2003.

REFERENCES

- [ZGFN08] Christopher Zach, David Gallup, Jan-Michael Frahm, and Marc Niethammer. Fast global labeling for real-time stereo using multiple plane sweeps. In *International Workshop on Vision, Modeling and Visualization*, pages 243–252, 2008.
- [ZLC⁺14] Hongyuan Zhu, Jiangbo Lu, Jianfei Cai, Jianmin Zheng, and Nadia Thalmann. Multiple foreground recognition and cosegmentation: An object-oriented crf model with robust higher-order potentials. In *IEEE Winter Conference on Applications of Computer Vision*, 2014.
- [ZLL09] Ke Zhang, Jiangbo Lu, and Gauthier Lafruit. Cross-based local stereo matching using orthogonal integral images. *IEEE Trans. Circuits Syst. Video Techn.*, 19(7), 2009.