Visual saliency detection in image and video data

Luo, Ye

2013

Luo, Y. (2013). Visual saliency detection in image and video data. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/61800>

<https://doi.org/10.32657/10356/61800>

Visual Saliency Detection in Image and Video Data

Luo Ye

School of Electrical & Electronic Engineering

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

2013

Declaration of Authorship

I, Ye Luo, hereby certify that the thesis titled, *Visual Saliency Detection in Image and Video Data*, and the work embodied in it is the result of original research done by me and has not been submitted for a higher degree to any other University or Institute.

.....

Date

.....

Luo Ye

*To my parents
for their encouragements and love.*

Acknowledgments

I would like to express my deepest appreciation and sincere gratitude to my supervisors: Assistant Professor Yuan Junsong, Associate Professor Xue Ping and Dr. Tian Qi, for their invaluable inspiration, guidance and support. Prof. Yuan likes an order brother who teaches me a lot about doing the research and the career planning. His attitudes and strict requirements to both the research and the life set a good example to me, and will continue to influence me in the future. Prof. Xue Ping acts as a father who encouraged me a lot especially when I got depressed. Dr. Tian a senior researcher who makes me realize that the diligence and the persistence are very important to a researcher. This work would not have been possible without them.

My thanks also go to Dr. Zhao Gangqiang, Yu Gang, Wang Lu, Zhao Lifan, Wang Yangtao, Wang Junyan and Huang Likun, for their help and support. Together with them, the days to pursuing my Ph.D. degree in NTU become more colorful and unforgettable.

I would like to acknowledge School of Electrical & Electronic Engineering, Nanyang Technological University, for awarding me the research scholarship. Its excellent facilities and resources have also been essential to this work.

Finally, I would like to thank my parents and my husband for their love, understanding, and support through my years of study away from home.

Contents

Acknowledgments	i
Abstract	vi
List of Figures	ix
List of Tables	xv
1 Introduction	1
1.1 Motivations and Objectives	1
1.2 Major Contributions of the Thesis	4
1.3 Organization of the Thesis	5
2 Literature Review and Background	10
2.1 Human Visual Saliency	10
2.1.1 Visual Attention	11
2.1.2 Saliency Analysis by Experiment	12
2.1.3 Computational Visual Saliency Models	13
2.2 Salient Object Detection and Its Applications	16
2.2.1 Salient Object Detection Based on Saliency Map	16
2.2.2 Salient Object Detection Along with Saliency Map Estimation . .	18
2.2.3 Applications of Salient Object Detection	20
3 Spatio-Temporal Enhanced Sparse Feature Selection for Video Saliency Estimation	24

3.1	Introduction	24
3.2	Video Saliency by Sparse Features Selection	27
3.2.1	Dictionary Learning	29
3.2.2	Spatial Saliency Estimation	29
3.2.3	The Proposed Spatio-temporal Saliency	32
3.3	Experimental Results and Analysis	37
3.3.1	Experimental Setup	37
3.3.2	Result Comparisons	39
3.4	Discussion	43
3.4.1	The Clip-level Saliency vs. the Frame-level Saliency	43
3.4.2	Spatial Saliency vs. Temporal Saliency	44
3.4.3	Multi-scale Saliency vs. Single-scale Saliency	45
3.4.4	Parameters to Dictionary Learning	46
3.5	Conclusion	49
4	Saliency Density Maximization for Salient Objects Discovery	52
4.1	Introduction	52
4.2	Salient Object Detection in Images	55
4.2.1	Existing Schemes	55
4.2.2	Our New Approach	56
4.3	Salient Objects Detection in Videos	60
4.3.1	Salient Object Detected in Each Frame	61
4.3.2	Importance Map and Globally Optimal Salient Sub-image	61
4.4	Experimental Results	62
4.4.1	Salient Object Detection in Images	62
4.4.2	Salient Object Detection in Videos	71
4.5	Discussion	81
4.5.1	From Saliency Map to Salient Objects	82
4.5.2	Multiple Salient Objects Detection	82
4.5.3	Limitations of the Proposed Method	84
4.6	Conclusion	85

5	Salient Region Detection via Optimal Path Discovery	86
5.1	Introduction	86
5.2	Discriminative Map for Salient Video Object	88
5.2.1	Video Saliency Estimation	88
5.2.2	What's Salient Object	90
5.3	Salient Path Discovery	91
5.3.1	Temporal Coherence	92
5.3.2	Fusing Temporal Coherence to Saliency Density	94
5.3.3	The Proposed Method	96
5.4	Experiments	99
5.4.1	Experimental Setting	99
5.4.2	Experimental Results	100
5.5	Conclusion	111
6	Thematic Object Discovery via Subgraph Mining	112
6.1	Introduction	112
6.2	The Proposed Method	114
6.2.1	Salient Segment Selection	115
6.2.2	Segment Affinity Graph Model	117
6.3	Experiments	119
6.3.1	Datasets	119
6.3.2	Experimental Setting	121
6.3.3	Performance Evaluation	122
6.4	Discussion and Conclusion	126
6.4.1	Discussions	126
6.4.2	Conclusion	128
7	Conclusion and Future Work	130
7.1	Conclusion	130
7.2	Further Research	133

Bibliography**136**

Abstract

The study of psychology and cognitive science has shown that the human perception is selective. When seeing images or watching videos, we mainly focus on a salient sub-region of an image, e.g. the salient object, and follow this salient object in image sequences. Efficient saliency analysis and accurate salient object detection are fundamental problems to various computer vision applications, i.e. image/video retargeting, object recognition and etc.

In this thesis, a systematic study is performed for visual saliency analysis in images and videos.

First of all, in order to identify the important visual contents in videos, we propose a novel video saliency estimation method by fusing spatio-temporally selected sparse features. Since the proposed method can accurately represent each frame by the learned dictionary and consistently incorporate the temporal information, our method outperforms three state-of-the-art methods and achieves good performance on two public datasets.

Although different saliency models have been proposed for images/videos, due to the cluttered background, it is not easy to accurately locate the salient object and crop it out from a noisy saliency map. We further propose a novel saliency density maximization method to detect salient objects in saliency maps. Without a prior knowledge of the salient object, our method can adapt to different sizes and shapes of the objects, and is less sensitive to the cluttered background.

Moreover, considering that the temporal coherence of salient objects in consecutive frames is usually strong, we extend salient object detection from images to videos by formulating salient object detection as a salient path discovery problem. A global op-

timal solution can be obtained by the proposed dynamic programming algorithm. The comparisons with two state-of-the-art object detection methods and one tracking method further demonstrate the efficiency of the proposed method on salient object detection in videos.

At last, similar to salient object in an image, many videos contain the thematic objects (e.g. the bride and the groom in a wedding ceremony video), which appear frequently in a video scene and thus retain our impression after watching it. We propose a subgraph mining method which incorporates the spatio-temporal video context to find the thematic objects and label the salient positions that thematic objects appear. Experimental results on two public datasets and one self-collected eye-tracking dataset show the efficacy of the proposed method on thematic object detection.

List of Figures

1.1	Outline of his thesis: arrows represent the dependencies between the various concepts that are introduced in the chapters.	6
2.1	An example of fixations and saccades over text. This is the typical pattern of eye movement during reading. The eyes never move smoothly over still text.	13
3.1	Fixation points of five people by different colors are overlaid on one video clip from CRCNS-ORIG dataset (the first column). Our video saliency maps by sparse feature selection that comprise both of the spatial saliency map (the second column) and the temporal difference saliency map (the third column) are shown in the forth column. Results from Hou's method [1] are also provided for comparison in the last column. Better to see in color.	27
3.2	Illustration of the proposed video saliency estimation. Feature selections and the saliency maps built by the selected features at each step are also shown, e.g. spatial saliency map, temporal consistency saliency map and temporal difference saliency map. Here, SMap is for short of Saliency Map.	28
3.3	Sample results of temporal consistency saliency maps (second row) and temporal difference saliency maps (third row). Temporal consistency saliency maps highlight the shared saliency (e.g. the bodies of the tap and the fox) while temporal difference maps focus on the changes (e.g. the dropping water and the moving legs).	35
3.4	Sample results of our video saliency detection in employed video datasets. The last video clip shows psychological pattern displayed in a video. . . .	36
3.5	Video saliency detection comparisons between models [2, 3, 1] and ours on "beverly03" from CRCNS-ORIG dataset via KL-divergence.	40

3.6	Video saliency detection results of two video clips on DB2. The first row shows the original key-frames and the second row shows our results. Video saliency estimated from Hou's method [1] and Li's method [4] are shown in the third and the fourth rows, respectively.	42
3.7	The ROC comparisons of our method with Hou's method [1] (ICL for short) and Li's method [4] (CRC for short) on B2.	43
3.8	Sample results of spatial saliency maps at clip level (the second row) and frame level (the third row). As can be seen, for a video with complex content (first video clip), features selected at clip level can characterize the saliency but they are not so discriminative and only the salient global structures are highlighted. For a video clip with simple background (second video clip), both methods perform similarly well.	44
3.9	The ROC curves of our frame-level spatial saliency map, temporal correspondence saliency map and temporal difference based saliency map on DB2.	45
3.10	Sample results of single-scale based saliency maps (second row), multi-scale based saliency maps (third row) and original key frames in the first row.	46
3.11	The ROC performance comparisons for our multi-scale and single-scale saliency map estimation on DB1.	47
3.12	The ROC performance comparisons of our method with different number of randomly sampled patches to learn a dictionary of size 1200 for each video clip on DB1.	48
3.13	The ROC performance comparisons of our method with different dictionary sizes on DB1. For each video, the dictionary is learned from 10000 randomly sampled patches.	49
3.14	The ROC performance comparisons of our method with $L^*a^*b^*$ color feature and intensity feature on DB1.	50
4.1	Our salient object detection result using the saliency map proposed in [1]. The first image is the original saliency map. The second image is the binarized saliency map. The third image is the result of salient object localized by maximum saliency region detection. The red rectangle is the detected result while the blue one is the ground-truth from [5]. The last one is our result via saliency density maximization. Better to see in color.	53
4.2	(a) Original image (b) Sparse saliency map by [1] which highlights edges and corners (c) Dense saliency map by [6] which highlights overall salient regions.	57

4.3	Detection of MSD, MRFS and MSR. The first row shows saliency maps using Hou's [1], Achanta's [6] and Bruce's [7] methods, respectively. The second row shows the localization results by MRFS with the exhaustive search in the three saliency maps. The third row shows results by MSR. The results in the third row are by MSD. Detected results are labeled with red lines. Blue rectangles are ground-truth from [5]. Better to see in color.	63
4.4	Precision, recall and F-measure for MRFS with four λ s {95%, 90%, 85%, 80%} on Hou's saliency map.	64
4.5	Performance comparisons among MSD, MRFS and MSR on three types of saliency maps.	65
4.6	Comparison of our MSD with other salient object detection results 1: Ma's saliency map and their salient object detection result [8]; 2: MRFS with $\lambda = 95\%$ on Itti's saliency map [2]; 3: MSD on Hou's saliency map; 4: MSD on Bruce's saliency map; 5: MSD on Achanta's saliency map; 6: MSD on the fused saliency map; 7: Supervised method from [5].	67
4.7	Evaluation C-value for MSD on (a) Hou's saliency map, (b) Bruce's saliency map, (c) Achanta's saliency map and (d) Combined saliency map. x-coordinate is C-value measured in unit $10^4 \times m \times n$ and y-coordinate is the corresponding averaged F-measure.	68
4.8	Evaluation C-value for three individual datasets which are separated groups from MSRA dataset on (a) Hou's saliency map, (b) Combined saliency map. x-coordinate is C-value measured in unit $10^4 \times m \times n$ and y-coordinate is the corresponding averaged F-measure.	69
4.9	More salient object detection results by MSD. The first row shows our results on Hou's saliency map [1]. The second row shows our results on Achanta's saliency map [6] and the results in the last row are based on Bruce's saliency map [7]. The red rectangle is the detected result while the blue one is the ground truth from [5]. Better to see in color.	70
4.10	The first row: sampled saliency maps of the table tennis video clip; The second row: corresponding detected bounding boxes for frame salient object in green and the human labeled ground truth in blue; The third row: intermediate states of constructing I_{map} ; The last row: the global bounding box with the target aspect ratio $r_0 = 9/16$ in red and the human labeled global ground-truth in blue.	72
4.11	(a) and (b) are frame salient objects (rectangles in green) detected by MSD without aspect ratio constraint. (c) and (b) are comparison of our selected salient sub-image regions (rectangles in red) between central cropping results (rectangles in yellow). Better to see in color.	76

4.12	Intermediate states of constructing I_{maps} for sports video 4×100 relay (in the top row), Simpson video (in the middle row) and the movie Graduate2 video (in the third row).	77
4.13	Comparison (a) our results in red and results from [9] in green. Green bounding box is moving which leads to dithering while our result completely contains the salient objects without shaking. (b) our results in magenta and results from [10] in red. Red bounding box is shaking which can be seen from its distance changes to our result which is better for this fixed camera scene. Better to see in color.	78
4.14	Video retargeting results by cropping salient sub-images for nine video shots which are displayed in each row. For each video, the first columns are the final selected region by our method; the second columns are the results by directly scaling original video and the last columns are the results by first-padding-then-scaling.	79
4.15	Performance of precision, recall and F-measure by our video salient object detection method on Ten-Video-Clips dataset with Chapter 3's saliency estimation method.	81
4.16	Multiple objects detection results on the balloon image from [2] and some natural images from [2, 5] with maximum salient object number $M = 5$ (first row) and $M = 2$ (second row) respectively. Better to see in color. .	84
5.1	The main idea of our method. Salient objects in a video can be detected by discovering a salient path which has the maximum accumulated saliency density. By adding the temporal coherence constraints to the most dense regions, our method can accurately detect the salient objects. Better to see in color.	87
5.2	Sample results of video saliency maps: original frames (the first row), context aware saliency map (the second row), motion saliency map (the third row) and our fused ones (the fourth row). Better to see in color. . .	89
5.3	The illustration of a window u in frame t and its 9-neighbors in frame $t-1$, and the pixel mapping by motion vectors is also shown.	93
5.4	Sample results by our method (third row) and MSD [11] (last row) on the 'Riding Horse' videos from UCF-sports dataset. The first and the second rows show the original frames and their corresponding video saliency map. The blue mask indicates the detected results while the orange ones are the ground truth. Without considering the motion coherence, MSD cannot consistently detect the salient object based on the saliency density alone). Better to see in color.	101

5.5	Precision, recall and F-measure comparisons our method with MSD and OPD on Ten-Video-Clips dataset.	102
5.6	Sample results by our method (second row) and OPD (last row) on the 'Kicking' videos from UCF-sports dataset. The blue mask indicates the detected results while the orange ones are the ground truth. Without considering the motion coherence, OPD cannot detect the small salient objects. Better to see in color.	102
5.7	Precision, recall and F-measure comparisons among our method, OPD with motion coherence and the original OPD without considering motion coherence on UCF-sports dataset.	104
5.8	Sample results by our method (first row) and the tracking method (last row) on the 'Swing-Side Angle' videos from UCF-sports dataset. The blue mask indicates the detected results while the orange ones are the ground truth. Since method in [12] tracks an object by a fixed size bounding box, it fails to detect the object with strong motions and big shape changes). Better to see in color.	105
5.9	F-measure comparisons between our method and tracking method [12] on the 'Swing-Side Angle' videos from UCF-sports dataset.	105
5.10	Precision, recall and F-measure comparisons our method with saliency maps from Chapter 3 and Chapter 5 on Ten-Video-Clips dataset.	107
5.11	Performance comparisons for our method by a fixed size window (single scale) and windows with multiple scales on the two employed video datasets.	108
5.12	Performance comparisons for our method by a fixed size window (second row) and windows with multiple scales (third row) on swing action videos from UCF-sport dataset and DO01-055 from Ten-Video-Clips dataset. The blue mask indicates the detection results for our method while the orange ones are the ground truth. Better to see in color.	109
5.13	Multiple salient object detection results by our method on 'Kicking' and 'Riding Horse' videos from UCF-sports. The blue mask indicates the detected results while the orange ones are the ground truth. Better to see in color.	110
6.1	Sample results of thematic video saliency discovery. The first row shows the original key-frames with the gaze points overlapped. Each green cross '+' represents one gaze point. The second row shows our detection results while the last row shows the results of co-saliency method [13]. The discovered video thematic salient region is rendered in red.	113

6.2	The main steps of our method has been illustrated. Details can be found from the text in Sec. 6.2.	115
6.3	Thematic video object discovery result comparisons with bottom-up saliency filtering (the third row) and without saliency filtering (the second row). Better to see in color.	116
6.4	Sample results of video saliency discovery. (a) and (b) show two videos of our eye tracking dataset. (c) shows one commercial video. The first rows of all three sub-figures show the original key-frame and the gaze points obtained by the eye tracking system are also rendered for (a) and (b). Each green cross ‘+’ represents one gaze point. The second rows show our result. The image saliency and co-saliency results are shown in the third and fourth rows respectively.	123
6.5	Thematic saliency discovery results from the video collection dataset. (a) shows several videos which contain the thematic salient object, i.e., “Starbucks Logo”. (b) shows the videos which do not contain the thematic salient object.	124
6.6	The comparison of thematic saliency discovery results with/without segment selection. (a) the results without segment selection. (b) the results with segment selection. The discovered video thematic salient region is rendered in red.	126
6.7	Comparisons of the proposed video saliency detection algorithm, the co-saliency algorithm [13], and the single image saliency algorithm [6]. (a) the eye tracker datasets. (b) the RSD dataset. (c) the commercial video dataset.	127

List of Tables

3.1	Notations of feature weights and saliency maps.	28
3.2	Performance comparisons for surprise model [3], image saliency model (CIOFM for short) [2] and the patch variance model [14] to ours on DB1 by KL-divergence and AUC.	41
3.3	Comparisons among Hou’s method [1], the proposed method with a universal dictionary and the proposed method with our learned dictionary on DB1 via KL-divergence and AUC.	42
3.4	Comparisons the performances of our frame-level spatial saliency, temporal consistency saliency and temporal difference saliency in a single scale form on DB2.	44
3.5	The number of randomly sampled patches vs. AUC performance.	47
3.6	The learned dictionary size vs. AUC performance.	48
4.1	Averaged time cost $\pm$ standard deviation by seconds for each image. . .	70
4.2	Precision, recall and F-measure for frame salient object and global salient sub-image detection in table tennis video clip.	74
4.3	Basic information for testing video shots.	75
5.1	Averaged F-measure (%) $\pm$ Standard Deviation for ten types of action videos in UCF-sports dataset.	103
5.2	Average time cost to process each video for ten types of action videos in UCF-sports dataset.	110
6.1	The information of four datasets.	120
6.2	The performance on video collection dataset.	125

Chapter 1

Introduction

In this chapter, we discuss the motivation to our research and what we aim to achieve through our investigations on the topic of visual saliency analysis and algorithms. We also provide the scope of our study and summarize the main contributions we have made.

1.1 Motivations and Objectives

Visual attention plays an important role in human visual system. It provides an efficient mechanism to allocate our sensory or limited computational resources to the most valuable information in the vast amount of incoming visual data. A lot of efforts have been devoted to methods on generating saliency maps to indicate where human focus on when seeing an image [2, 6, 7, 15, 16]. A saliency map with each pixel value ranging from $[0, 1]$ is used to show the salient positions in an image. Although many image saliency estimation methods have been proposed, few methods have been proposed to estimate video saliency. Some methods try to extend the center-surround based image saliency estimation methods to video saliency estimation [4, 17]. However, the selection of the

surround scale becomes more complex than it for image saliency. Besides, how to incorporate the temporal information into video saliency estimation becomes another issue. Some methods have been proposed to detect saliency from frame difference while others tend to detect the common parts in images as saliency [18, 19] and [20]. However, only using frame difference or common regions between images alone is not enough to estimate video saliency. This poses a problem to propose a novel saliency estimation method which can handle the scales of surrounds and well incorporate the spatio-temporal information among consecutive frames.

As the proliferation of computational saliency estimation methods in computer vision, salient object detection got broad interesting. A salient object is an object which draws attention in an image or a video clip. Accurate detection of salient object remains a challenge problem even though the salient positions have been identified in saliency maps. First of all, it is not uncommon that the obtained saliency map is noisy and incomplete. Due to the distraction from the clutter background, it is not easy to find the salient region and accurately crop it out. Moreover, there is no unique definition for a salient object. Some existing methods apply exhaustive search for the minimum region that covers a fixed amount of salient points, e.g. 95 % of the total salient points [5] as the target salient object. The major limitation is that it is difficult to predefine the amount of saliency the salient region should contains, as it depends on the size and shape of the salient object, as well as how cluttered the background is. Ideally, the salient region should be adapted to the shape of the salient object. It's even difficult to crop a salient region from videos. Exhaustive search becomes time consuming and impossible to apply to image sequences. Instead of cropping the salient object with a bounding box, some other methods rely on object segmentation to obtain an accurate object shape [6, 16]. This category of methods, however, it is not robust to the cluttered background and

usually requires prior knowledge to help an accurate segmentation. How to efficiently discovery salient objects in images and videos become an important problem. Moreover, accurate salient object detection is a fundamental problem to image/video retargeting, video summarization, video coding, etc.

Although salient object could be efficiently detected by our proposed method by maximizing saliency density (MSD), especially for image salient objects, the major limitations of this method on video salient object detection are that the saliency density is the sole factor to detect objects and the temporal correlations between salient objects are ignored. How to accurately detect salient objects based on saliency maps but adjusted by temporal coherence becomes an interesting problem.

Despite of many methods were proposed to detect salient object in the given saliency map, some methods were proposed to detect salient object along with the saliency estimation. Previous methods try to estimate the saliency of a patch or a region by comparing the difference to its neighborhoods or other object regions, e.g. objectness measurement in [21, 22]. Therefore, given a candidate object region, the salient object detection problem relies to measuring the difference and the saliency value of this region is approximately equal to the difference. Like the salient object could be detected via the saliency estimation, many thematic objects, which appear frequently and retain our impression in the whole video scene, could be obtained by our new thematic saliency estimation. Provided a video clip, finding such thematic objects, such as the bride and the groom in a wedding ceremony video, the birthday girl in a birthday party video, or a commercial product in an advertisement video, is of great interests as it can help to better understand and summarize the video content, which is critical to video annotation and retrieval applications.

1.2 Major Contributions of the Thesis

The major contributions of the thesis are the following:

1. a new video saliency estimation method is proposed to represent video saliency by spatio-temporally selected sparse bases which have high entropy gains, and a multiple level analysis of video saliency estimation is provided as well. Moreover, by taking both temporal consistency and temporal difference into consideration at pair-wise frame level, the obtained video saliency, which implicitly incorporates the spatial saliency into temporal consistency, is not only spatially important but also temporally consistent and novel, therefore is more accurate. Moreover, unlike image saliency estimation methods by employing a universal dictionary for content representation, a specific dictionary is learned from a given video, and the corresponded sparse bases can thus better characterize the given video content. This work has been published in Computer Vision and Pattern Recognition Workshops 2012 [23].
2. we first formulate the task of the unsupervised salient object detection as the maximum saliency density discovery problem, which well balance the size of the saliency object and the saliency it contains. It is also compatible to different types of saliency maps or a fused saliency map. Secondly, to obtain the global optimal solution of the saliency density maximization problem, we propose an efficient branch-and-bound search algorithm, which is based on the saliency density rather than the traditional classification scores. Derivation of the upper-bound estimation and the average convergent time show its efficiency on salient object detection in both images and videos. Moreover, we extend our salient video region detection method and apply it to video retargeting. All these works have been published in Proceeding of 10th

Asian Conference on Computer Vision [24], IEEE Conference on Multimedia Expo 2011 [25] and IEEE Transaction on Circuits and System for Video Technology [11].

3. we propose a novel method for salient video objects detection and localization as an optimal spatio-temporal path discovery problem. A globally optimal solution is also obtained by the proposed dynamic programming algorithm. By taking the temporal coherence of salient objects into consideration, the proposed algorithm obtains more accurate detection results. Moreover, without any prior knowledge of the salient objects, our method can automatically detect the salient objects across different scales and aspect ratios. Experimental results on two public datasets demonstrate the effectiveness of the proposed method on salient video object detection. This work has been published in ACM multimedia 2013 [26].
4. different from common salient object detection, we propose a new method to find the thematic objects that frequently appear at the salient positions in the video scenes. By representing all image segments in the video as the spatio-temporal context, we build an affinity graph among them, and discover the thematic object by a cohesive sub-graph mining method. Unlike individual image saliency or co-saliency analysis, our proposed video saliency fully incorporates the whole spatio-temporal video context. Experiments on our newly developed eye tracking dataset as well as other three datasets further validate the effectiveness of our method on thematic object detection. This work has been published in International Conference on Computer Vision Workshops 2013.

1.3 Organization of the Thesis

The rest of the thesis is organized as follows, which could also be found from Fig.1.1.

Figure 1.1: Outline of his thesis: arrows represent the dependencies between the various concepts that are introduced in the chapters.

Chapter 2 outlines the background information on visual saliency analysis, including the literature review for salient object detection and their corresponding applications in computer vision.

In Chapter 3, a new video saliency method is proposed to systematically estimate video saliency spatial-temporally. We make use of the over-complete bases learned from given video clips instead of using the universal ICA bases for common video content, therefore, the learned over-complete bases can better represent the given video content. Instead of simply combining the spatial saliency map and the temporal saliency map, our method achieves video saliency representation by the sparse features selected via measuring their responses entropy gain at multiple scales. Regarding to spatial-temporal feature selections, temporal correlation is fully considered by introducing temporal consistency and temporal difference into feature selections.

In Chapter 4, we formulate the task of the unsupervised salient object detection as the maximum saliency density discovery problem, which well balance the size of the saliency object and the saliency it contains. It is also compatible to different types of saliency maps or a fused saliency map. To obtain the global optimal solution of the saliency density maximization problem, we propose an efficient branch-and-bound search algorithm, which is based on the saliency density rather than the traditional classification scores. Derivation of the upper-bound estimation and the average convergent time show its efficiency on salient object detection in both images and videos.

In Chapter 5, we propose a novel method for salient video objects detection and localization as an optimal spatio-temporal path discovery problem. A globally optimal solution is also obtained by the proposed dynamic programming algorithm. By taking the temporal coherence of salient objects into consideration, the proposed algorithm obtains more accurate detection results. Moreover, without any prior knowledge of the salient objects, our method can automatically detect the salient objects across different scales and aspect ratios.

In Chapter 6, we present a new method to detect thematic video object, which draws human attention through a video and is another kind of salient object. To discover the thematic object, we formulate it as a cohesive sub-graph mining in the affinity graph of image segments in a video. By incorporating spatio-temporal context of the video into the graph building, the discovered thematic object is globally salient.

Finally, we conclude the work we have done and explore the further research in Chapter 7.

Chapter 2

Literature Review and Background

In this chapter, we first introduce the background of human visual saliency model and estimation methods, and then a literature review is presented for salient object detection methodologies. At last, some applications based on the salient object detection are also discussed.

2.1 Human Visual Saliency

The research of human visual saliency is originated from human visual attention. Before introducing visual saliency, we first present a brief literature interview of human visual attention. If there is no special description, visual saliency all refers to human visual saliency in this thesis.

2.1.1 Visual Attention

Attention is the cognitive process of selectively concentrating on one aspect of the environment while ignoring other things. Examples include listening carefully to what someone is saying while ignoring other conversations in a room (the cocktail party effect) or listening to a cell phone conversation while driving a car. After William James who is the first person to outline a theory of human attention [27], Broadbent proposed his filter theory of attention in an attempt to explain many of the existing experimental results [28]. The response selection theory of the attention was proposed by Deutsch, who indicated that a part of attention involves high level processing. In 1960s, Treisman proposed a series of models that combined early and late selection into a model known as Feature Integration Theory (FIT) [29].

Attention has also been referred to as the allocation of processing resources. It provides a mechanism for selection of particular aspects of a scene for subsequent processing while eliminating interference from competing visual event. Attention focuses processing on a selected region of the visual field that need not coincide with the center of fixation. In biological vision, people do not usually look at every object in the visual field but concentrate on some particular aspects of a scene. Research on finding what places and why these places people put their attention to are called visual attention. Psychologists have investigated visual attention for many decades using psychophysical experiments, such as visual search tasks, with carefully controlled stimuli. Detailed explanation of experimental visual saliency analysis is presented in Section 2.1.2.

Moreover, the deployment of visual attention is believed to be driven by visual saliency mechanisms, which is a fundamental, yet hard to define, property of vision systems, that had been known to exist for a number of elementary attributes of visual stimuli, including color, orientation, depth, and motion, among others. Visual saliency has been

extensively studied in psychology, neuroscience and computer vision literatures over last several decays. With the proliferation of eye-tracking systems over the last two decades, a number of computational models attempting to account for the data and addressing the question of what attracts attention were also proposed. A short summary of the state-of-the-art computational visual saliency models is presented in Section 2.1.3.

2.1.2 Saliency Analysis by Experiment

There is no common sense on that what attention usually is. But most would agree that attention is a selective process which acts to focus sparse computational resources onto relevant aspects of sensory inputs.

When we talk about attention, it is usually discussed from 'overt' and 'covert' attention. Overt attention is to direct sense organs towards a stimulus source. Covert attention is to mentally focus on a possible stimulus. Currently, the visual covert attention is thought of a mechanism for quickly scanning interesting locations. The shift between different locations is linked to eye movement which is a slower saccade to that location. In other words, eye movement is typically divided into fixations (a certain position that the eye gaze pauses) and saccades (the shift between different positions). More information about the principles of eye movement detection and gaze tracking methods could be found from a review paper in [30]. The series of fixations and saccades is called a scan path. Scan paths are useful for analyzing cognitive intent, interest, and saliency. Previous researchers found that most information from the eye is available during a fixation, but not during a saccade.

Fixation or visual fixation is the maintaining of the visual gaze on a single location. Humans typically alternate saccades and visual fixations. An example is shown

in Fig. 2.1 [31]. Eye tracking is the process of measuring either the point of gaze or

Figure 2.1: An example of fixations and saccades over text. This is the typical pattern of eye movement during reading. The eyes never move smoothly over still text.

the motion of an eye relative to the head. An eye tracker is a device for measuring eye positions and eye movement. Eye tracking experiment is always used to generate the ground truth data to measure the performance of saliency computational model on prediction human eye movement. Therefore, a good saliency models should achieve good performance in predicting human fixations in viewing images. With the development of affordable and easy-to-use modern eye-tracking systems, the locations that people fixate when they perform certain tasks can be explicitly recorded and can provide insight into how people allocate their attention when viewing complex natural scenes. The proliferation of eye-tracking data over the last two decades has led to a number of computational models attempting to account for the data and addressing the question of what attracts attention.

2.1.3 Computational Visual Saliency Models

During last two decades, visual saliency detection and saliency map generation aiming to find out what attracts humans' attention got broad interesting in computer vision espe-

cially for object detection/recognition from different scenes. There are many categorization methods for computational visual saliency models, such as the bottom-up method vs. the top-down method, the image saliency vs. the video saliency, the edge/corner based method [1, 7, 32] vs. the global region based method [5, 16, 6, 33], etc. More details can be found from [34]. In this Section, we briefly introduce the related work from the image and the video saliency estimation, respectively.

2.1.3.1 Image Saliency

Image saliency estimation methods can be broadly grouped into two categories: information theory based and contrast based.

At first, in information theory based methods, visual saliency is represented by rarely used features and the features employed are obtained from different methods. Discriminative image saliency is obtained by learning salient features through mutual information maximization [35]. Despite different information criteria are adopted to estimate image saliency, most of them are based on the universally learned features [7, 1, 36]. To obtain the universal features, independent components analysis (ICA) is broadly used to train a complete dictionary [37]. With the ICA sparse representation, Bruce et al. proposed to first obtain the probability of each basis by kernel density estimation and then approximate the patch saliency by summing the Shannon's self-information of the probabilities [7]. Hou et al. select the features based on their responses to entropy gain across the entire image and estimate the patch saliency as the reweighed basis responses [1]. Zhang et al. also use sparse representations by ICA and propose to estimate image saliency based on a Bayesian framework [38].

Saliency is explained as contrast and parts standing out from the rest of the image. The center-surround scheme is implemented by performing Difference of Gaussian filters

on multiple feature maps in [2]. Patch saliency in an image is set to the reconstruction error with its surrounding patches as the dictionary in [4]. A modified patch saliency is built by the reconstruction error weighted by sparse coding length and the dictionary is again composed by surrounding patches in [17]. Another way similar to contrast is to detect image saliency by sparsity pursuit, in which an image is decomposed into a low-rank matrix plus the sparse noise. The low-rank matrix is considered as the repeated background while the sparse noise indicates the unique locations which are salient [39, 40].

2.1.3.2 Video Saliency

Considering video saliency mainly different a static image saliency in extra temporal information among frames, various strategies are proposed to well incorporate it, such as pairwise image relationship and multiple frame correlations.

Pairwise image relationship is recently explored to detect the so-called co-saliency. Methods in [41, 42] aim to detect the generalized temporal difference such as the novel signal or the newly changed positions as saliency. While methods in [18, 19] and [20] tend to detect the common objects in the pairwise images as the saliency. Common features are extracted by KL-divergence distance to represent co-saliency between pairwise frames [18]. By multilevel over-segmentation, similar segments are found and the similarity between them is estimated as the co-saliency in the image pairs [19]. A co-saliency model is built to generate image regions that are similar to each other across images, and meanwhile retain their distinctness within each image [20].

General ways to implement multiple frame correlations are summarized as follows. At first, similar to image saliency estimation, ICA bases learned from 2D natural image patches are employed to represent each frame while temporal relationship is implemented by the basis relationship among consecutive frames [1, 36]. Second, ICA and similar 3-D

filters or kernels are extended and performed on 3D spatial-temporal cubes to better incorporate temporal information [38, 43, 44, 45]. Still based on the 3D ICA sparse representation, a local regression kernel to estimate the local contrast is proposed in [44]. Third, video saliency maps are obtained by exploring different fusion schemes to integrate static saliency maps and dynamic saliency maps simultaneously in [46, 47, 48]. Considering the complex semantic information involved in videos, learning based methods are employed to get high level guidance for video saliency estimation and fusion in [49, 50, 51]. Fourth, systemic models are built to detect video saliency. A surprise model by calculating the distance between the prior and the posterior probability of the observed data is first proposed in [52]. To treat the video sequence as the dynamic systems, video saliency estimation is modeled and solved by graphical models in [53, 54]. A center-surround scheme considered spatio-temporally is proposed in [55].

2.2 Salient Object Detection and Its Applications

Generally, there are two ways to detect salient objects: (1) given saliency maps, salient object detection could be directly performed on the maps and the key techniques lay onto efficiently finding the candidate salient regions; (2) together with various definition of an object, salient object could be detected through the saliency estimation procedure.

2.2.1 Salient Object Detection Based on Saliency Map

2.2.1.1 Salient Object Detection in Images

Given the saliency map, the simplest way to obtain salient object is to threshold saliency map into binary mask. Methods to threshold saliency map are intensively discussed

in [56, 6, 15, 57, 58]. These methods depend on the selection of the threshold. In order to accurately detect salient objects from saliency maps, image segmentation is combined with the saliency map in [32, 16, 6]. However, the performance heavily relies on the accuracy of image segmentation results. Some heuristic methods [59, 60] are proposed to improve the performance of salient object detection. However, accurate detection of the salient object boundary is not always necessary. The other category of salient object detection is performed by finding the bounding box of the object. For example, a fuzzy growing algorithm to locate the attended area is proposed in [8]. Exhaustive search is adopted in [5] to find the smallest rectangle containing at least 95% of salient pixels for a learned saliency map model. Liu et al. [56] noticed the disadvantages of exhaustive search and proposed to use the dynamic threshold and the greedy algorithm to improve the search efficiency. However, their methods still rely on thresholds. In [16], the search of the rectangular sub-window is speeded up by applying the efficient sub-window search (ESS). ESS is a recently proposed branch-and-bound search algorithm for sliding window search on object recognition [61]. Salient object detection with bounding boxes has many applications in image analysis, such displaying images on a small device, or browsing image collection [16, 56, 62].

2.2.1.2 Salient Object Detection in Videos

Salient object detected in videos can be performed by finding a rectangle bounding box in individual frame like the ways in images [63, 64]. In [64], a single salient object is detected by energy minimization and labeled by a bounding box in each frame. However, salient objects on consecutive frames are probably detected with different sizes, thus they cannot be simply cropped out and applied for displaying purpose. In order to consistently present the regions of interest detected from videos, an active window allowing slight

orientation change is optimized in each frame to contain most informative pixels [65]. Liu et al. [66] found out salient regions from videos by minimizing penalties of information loss. Deselaers et al. [10] presented a supervised learning method to crop sub-windows based on the manually labeled salient regions. Both [66] and [10] have limitations due to virtual camera motions. The principled solution proposed in [67] allows the salient region of different sizes and considers the motion smoothness of the consecutive frames. However, it still has the risk of introducing virtual camera motions.

2.2.2 Salient Object Detection Along with Saliency Map Estimation

Although there is no unique definition for a salient object, we all agree that a salient object should be an object and salient in an image or a video. Previous methods try to find salient regions which distinguish themselves from background or other object regions, such as objectness measurement in [21, 22]. Therefore, given the candidate object region, the salient object detection problem lies onto finding how different it is from its context and the saliency value of the region is approximately equal to the difference. Based on the scales of surroundings, methods for saliency estimation or salient object detection could be categorized into three groups: local/global saliency within one image, image or frame saliency w.r.t. simple context and video saliency estimation.

Various methods have been proposed for local/global image saliency. Itti's center-surround feature extraction-integration method [2] is the representative work of local image saliency. Although, their features are obtained across multiple scales, this type of method is still limited to a local region [4]. In order to add the global effects to the saliency estimation, features characterizing the global effects are extracted from images

in [16] [68]. A global center-surround technique is implemented by using Difference-of-Gaussian (DoG) band pass image filters to estimate image saliency uniformly in [6]. In [69], image saliency is obtained as salient object and its context through globally patch comparisons within the image. In [70], image saliency estimation is modeled as anomaly detection with regard to different context, such as a patch saliency could be the reconstruction error by the patch dictionary. Similar idea was also used in [70] for image saliency which is obtained by comparing the difference of the target image to the image dictionary.

Recently, co-saliency as a new concept was proposed based on the pairwise image relationships. Methods in [41, 42] aim to detect difference e.g. the novel signal or the newly changed positions as saliency. The method proposed in [71] tend to detect the saliency in two images.

Compared with images, it is more difficult to define video saliency as there are more factors to influence the video saliency. Each frame as an image has its own saliency map by using different features such as color, intensity, etc.. If pairwise relationships among frames are considered, frame-level saliency can also be estimated as co-saliency [41]. Unluckily, few of them apply the whole video sequence as a spatio-temporal context for saliency estimation. Video saliency estimated from multiple frame context is also a recently explored idea [72]. However, most of them start from finding a wise combination scheme for different salient clues [73, 49]. The method proposed in [52] defines the video saliency as the surprise and a supervised method is used in [50] to obtain the video scene saliency. Different from them, to find the video saliency or a salient object from the understanding of its themes is still untouched.

Some high-level knowledge are also learned and involved to guide the salient object detection in images and videos [5, 74]. All these methods are performed with the default

assumption that there must be at least one salient object in each image. Wang et al. proposed, before salient object detection in each web image, a judgment whether there are any objects in images should be performed in advance [75]. There is no method to detect salient objects which can represent the video content and dont appear in every frames.

2.2.3 Applications of Salient Object Detection

Various applications were proposed for salient object detection either in images or in videos. In the following sections, we introduce two of them: image/video retargeting and video summarization.

2.2.3.1 Video Retargeting

Video retargeting is one of the most popular application for salient object detection in videos. Video retargeting aims to adjust frame resolution for better display on various sizes of screens. To keep and select salient regions from videos is one of the methods. Liu et al. [64] proposed to use energy minimization to detect a single salient object, which was labeled by a rectangular in a video segment. Hazem et al. [76] used multi-scale cropping based attention region selection method on surveillance video to select a particular region of the scene. For cropping/scaling based video retargeting, Wolf et al. [77] applied non-homogeneous transformation on different regions based on their saliency information. Yin et al. [78] incorporated user preference into image adaptation by iterative image retrieval. Liu et al. [66] found out the optimal window and its scale factors to the target screen from each frame by minimizing penalties of information loss. Deselaers et al. [10] presented a supervised learning method to crop windows based on the

manually labeled salient regions. Both [66] and [10] have limitations due to the conflict between large transformation from the cropping window to the target screen size and small virtual camera movements. To overcome the limitations of the above methods, we aim to automatically find an intermediate cropping window between detected salient regions and the target screen. With the auxiliary of this intermediate cropping window, less virtual camera motion is introduced among frames and more salient information is preserved.

2.2.3.2 Video Summarization

Video summarization targets to best keep the main idea of the video via brief video content representation. The most representative one is the group of Microsoft Research Asian led by Zhang Hong Jiang [79, 80]. They proposed many attention models from different channels such as visual, aural and linguistic attention. Visual attention model includes motion attention model, static salient model, face attention model and camera motion model. Aural attention model includes aural saliency, speech attention model and music attention model. All these attention saliency are fused by nonlinear fusion scheme to be an attention curve. The curvature of this curve is considered as the key frame. Motion attention is based on the assumption that motion with high intensity attracts more human attention. Hence, the motion intensity is calculated macro-block by macro-block for each frame. Spatial coherence and temporal coherence are added to make up the situations when only the motion intensity cannot work well. Static attention model is modeled by contrast based saliency map generation since they own all the saliency attention to appearance and contrast. Face attention model is built because they think dominant face is usually attractive. Faces are weighted differently based on the positions they appeared. That aims to show that human attention is always in the

center of screen. Camera attention model is built because the previous motion model can not handle camera motion and the camera motion is indicative on human visual attention.

Another work which is similar to their work is [80]. They built the model from the experience not from the principle of human vision system. That means that there is still some redundancy in the video summarization if we summarize video by user attention model. How to get a better video summarization is still a open problem.

Chapter 3

Spatio-Temporal Enhanced Sparse Feature Selection for Video Saliency Estimation

3.1 Introduction

Although many image saliency estimation methods have been proposed, it is not trivial to extend them to estimate saliency in videos. For example, in local contrast models [4, 17], the saliency of an image patch is estimated based on the reconstruction error by its surrounding patches. However, only comparing the center patch with the neighborhood patches may not be suitable as mentioned in [69, 68]. Taking a *LV* bag for instance, although each individual *LV* logo can be salient itself, they are not so attractive when the bag shows repeated *LV* logo pattern. The scale of the neighborhood causes problems. It is particularly true in video saliency estimation since multiple level patch comparisons are involved such as across pairwise frames and among multiple frames in the whole

video. Therefore, it is more rational to estimate video saliency from a global view of patch comparison or feature selection.

Moreover, to incorporate the temporal information into video saliency estimation, some proposed to explore various temporal correlations [1, 36, 18]. Method proposed in [36] aims to detect saliency from frame difference while [1, 18] tends to detect the common parts in images as saliency. However, only using frame difference or common regions between images alone is not enough to estimate video saliency. For example, a journalist who dominates a news video clip could be less attractive than an intruder, and a journalist who stands still is generally more interesting than pedestrians passing by. Therefore, a scheme which fully considers temporal correlations is desired. Finally, video saliency estimation targets to identify salient locations which are spatially important and temporally novel. It is natural to combine the spatial saliency and the temporal saliency, after obtaining them respectively [47]. However, it is difficult to select a fixed combination form or coefficients.

To address the aforementioned problems, we propose a novel saliency method: spatio-temporal enhanced sparse feature selection for video saliency estimation, in which the video saliency is represented by spatio-temporally selected sparse bases that have high entropy gains. Based on the sparse coding framework, we first learn a dictionary on a given video and each basis from the dictionary can be treated as a feature which could encode the content of the given video. Considering that how much information a basis provided is related to how many patches can be represented by this feature, spatial saliency estimation is analyzed based on the sampled patches from two levels: frame level and clip level. Given a frame, the frequency of a basis being used to reconstruct all patches from the frame indicates how common the basis is. Therefore, those bases that are rarely used provide more information and can be selected to estimate the frame-level saliency.

In other words, instead of measuring the patch saliency by the local reconstruction error, we resort to select informative sparse bases for the video saliency estimation. With regard to the clip-level saliency, it mainly differs from frame-level saliency into the number of patches used for feature selection. Patches from all frames in the video clip are used for the clip-level feature selection. At both frame level and clip level, since patches from different frames are treated equally and no temporal correlation is involved, we refer them to the *Spatial Saliency* estimation.

To obtain our spatio-temporal saliency, we rely on the *Temporal Coherence* at pairwise frame level. Instead of estimating video saliency either from frame difference or from common parts detection, temporal coherence considers two aspects: temporal consistency and temporal difference. Temporal consistency targets to select features repeatedly contributed to spatial saliency in consecutive frames while temporal difference chooses features which lead to scene changes and incorporate temporal novelty. We represent our video saliency by selecting sparse bases which are not only temporally consistent but also temporally novel. Features corresponded to different types of temporal correlations are first selected and then multiplied together. Instead of simply combining the spatial and the temporal saliency maps, we can obtain our video saliency by summing the responses of all fused features. It is also worth noting that the spatial saliency has been enhanced into our temporal consistency and the target spatio-temporal saliency can just rely on the saliency estimated at pairwise frame level. Since sparse bases are learned from all video patches and the proposed video saliency is represented by the selected features, it thus avoids problems that exist in local contrast based methods. Last but not least, instead of learning a universal dictionary from natural images [1, 36], we learn a dictionary for each video to better characterize the given video content. Fig. 3.1 illustrates the benefits of our video saliency estimation method. The headline is detected as the spatial saliency

(the second column) while the talking person is highlighted as the temporal difference saliency (the third column). The fixation points shifting between the spatial saliency and the temporal difference regions indicate that the desired video saliency should comprise both of them, and the fusing results in the forth column show the capability of our method to detect them all.

Figure 3.1: Fixation points of five people by different colors are overlaid on one video clip from CRCNS-ORIG dataset (the first column). Our video saliency maps by sparse feature selection that comprise both of the spatial saliency map (the second column) and the temporal difference saliency map (the third column) are shown in the forth column. Results from Hou’s method [1] are also provided for comparison in the last column. Better to see in color.

3.2 Video Saliency by Sparse Features Selection

In this Section, the general idea of the proposed method is illustrated in Fig. 3.2. We first learn a dictionary by the employed dictionary learning method. Given the learned dictionary, two levels of spatial saliency estimation methods are discussed consequently.

Figure 3.2: Illustration of the proposed video saliency estimation. Feature selections and the saliency maps built by the selected features at each step are also shown, e.g. spatial saliency map, temporal consistency saliency map and temporal difference saliency map. Here, SMap is for short of Saliency Map.

Based on the spatial saliency estimation results, we present our proposed spatio-temporal saliency estimation method, in which temporal coherence is considered from two aspects: temporal consistency and temporal difference. At last, to integrate the selected features which characterize different temporal correlations, a fusion scheme is proposed, and the proposed video saliency is obtained by summing all the responses from the fused features. Some notations we used for feature weights and saliency maps for temporal consistency, temporal difference, their fused one's, the frame level and the clip level are list in table 3.1.

Table 3.1: Notations of feature weights and saliency maps.

	temp.consistency	temp.difference	fused one	frame-level	clip-level
Weight	$\mathbf{W}_t^c$	$\mathbf{W}_t^d$	$\mathbf{W}_t$	$\mathbf{W}_t^s$	$\mathbf{W}_t^g$
SMap	S_t^c	S_t^d	S_t	S_t^s	S_t^g

3.2.1 Dictionary Learning

Instead of using the universal dictionary, we learn a specific dictionary for each video. RGB images are first converted into L*a*b* color space and then M non-overlapped image patches ($K = 16 \times 16 \times 3$ is the dimension of the patch feature) are randomly sampled from each frame in a video clip as $X = [x_1, x_2, \dots, x_M] \in \mathbb{R}^{K \times M}$. After subtracting patch mean from each patch, by employing dictionary learning method in [81], we obtain a dictionary $D = [d_1, \dots, d_N] \in \mathbb{R}^{K \times N}$ with each element $d_n \in \mathbb{R}^K$ ($n = 1, 2, \dots, N$) a basis:

$$\begin{aligned} D^*, A^* &= \arg \min_{D, A} \|X - DA\|_F^2 \\ \text{s.t.} \quad \sum_k D_{kn}^2 &\leq \delta, \quad n = 1, 2, \dots, N. \end{aligned} \quad (3.1)$$

Here $\|X\|_F^2 = \sum_k \sum_m X_{km}^2$, δ is a small positive constant and $A \in \mathbb{R}^{N \times M}$ is the coefficient matrix for all patches. Denote the m_{th} column vector of A by $A_m \in \mathbb{R}^N$ which is a new representation of the m_{th} patch by the obtained N bases. In other words, the content of the given video could be encoded by the sparse bases. Therefore, the more frequently the basis used, the more common appearance the patches share and the less information the bases provide. Specifically, by measuring how often a basis has been used, we can select informative ones to represent the expected video saliency. In the following Sections, we first introduce the spatial saliency estimation at the frame level and the clip level, and then our proposed spatio-temporal saliency estimation method and a fusion scheme are provided consequently.

3.2.2 Spatial Saliency Estimation

Given the learned dictionary, since the importance of a basis is varying with the number of patches used, based on where the patches are sampled, we discuss the spatial saliency estimation at the frame level and the clip level, respectively.

3.2.2.1 The frame-level saliency

Denotes all the patches extracted from frame t as $X^t \in \mathbb{R}^{K \times M_t}$, where M_t is the total number of patches in frames t . We can obtain the sparse coding representation $A^t \in \mathbb{R}^{N \times M_t}$ via D for frame t by solving $X^t = DA^t$ with the constraints of high sparseness and low reconstruction error. Based on the new representation, the probability or the activity ratio of the basis d_n within the frame t can be defined:

$$p_n = \frac{\sum_m^{M_t} |A_{n,m}^t|}{\sum_n \sum_m^{M_t} |A_{n,m}^t|}, \quad (3.2)$$

where $A_{n,m}^t$ is the (n, m) element of A^t . We take the absolute value of A from Eq. 3.2 to indicate the quantity of the n_{th} basis used to represent the m_{th} patch. That is to say, if p_n is small enough, the corresponding basis d_n is rarely used but a slight change to it would potentially lead to huge information transmitted. Thus, d_n is significant for saliency detection. With this definition, the visual information passed by this frame is jointly encoded by all these bases, and the optimal encoding strategy can be obtained by maximizing the entropy $H(p) = -\sum p_n \log p_n$. The entropy gain defined as Eq. 3.3 can measure the bias of each basis being used to maximize $H(p)$, and a similar idea was also used in [1]. The large value of the entropy gain means that the high chance to get large entropy by using a certain basis. Therefore, we use the entropy gain as the weight to select the basis:

$$\frac{\partial H(p)}{\partial p_n} = -H(p) - p_n - \log p_n - p_n \log p_n. \quad (3.3)$$

The weights for all bases contributed to the spatial saliency in frame t can be calculated as:

$$\mathbf{W}_t^s = \left[\frac{\partial H(p)}{\partial p_1}, \dots, \frac{\partial H(p)}{\partial p_N} \right]^T. \quad (3.4)$$

Therefore, the saliency value of a patch located at (x, y) in frame t can be calculated as the summation of weighted responses to all bases as:

$$S_t^s(x, y) = \sum_{n=1}^N w_{n,t}^s \times |A_{n,m}^t|, \quad (3.5)$$

where (x, y) is the center of the m -th patch and $w_{n,t}^s = \frac{\partial H(p)}{\partial p_n}$ is the weight of the n_{th} basis in $\mathbf{W}_t^s$ and $A_{n,m}^t$ is the coefficient of the m_{th} patch in frame t to the n_{th} basis.

3.2.2.2 The clip-level saliency

After obtaining the frame-level spatial saliency, it is natural to ask that it is possible to calculate the probability of the basis d_n in a video clip instead of a frame. For the complementary of this work, we introduce the clip-level saliency which is obtained by all the patches in a video. Therefore, Eq. 3.2 can be rewritten as:

$$\tilde{p}_n = \frac{\sum_m |A_{n,m}|}{\sum_n \sum_m |A_{n,m}|}, \quad (3.6)$$

where $A_{n,m}$ is the (n, m) element of $A \in \mathbb{R}^{N \times M}$. Since all the patches in the video clip are involved in deciding the importance of bases, we call the selected features by the global sparse features and the saliency maps built by the globally selected features as the global saliency maps. Similar to frame-level spatial saliency estimation, we set $w_n^g = \frac{\partial H(p)}{\partial p_n}$ as one element from the global feature weight $\mathbf{W}^g \in \mathbb{R}^N$. Larger value of w_n^g indicates that the basis d_n is more informative and intends to draw more attention. Thus, we can obtain our clip level spatial saliency map as:

$$S^g(x, y, t) = \sum_{n=1}^N w_n^g \times |A_{n,m}|, \quad (3.7)$$

where $S^g(x, y, t)$ is the global saliency score for the m_{th} patch located at (x, y, t) and $A_{n,m}$ is the coefficient of this patch to n_{th} basis.

Compared to saliency estimated at the frame level, saliency estimated at clip level mainly benefits from more patches being used to estimate the probability of a basis, but it still suffers from ignoring the temporal correlations among patches. The performance comparisons between the frame-level spatial saliency and the clip-level results are provided in Sec. 3.4.1. In the following subsection, we present the proposed video saliency estimation method which considers two types of temporal correlations among pairwise frames but can incorporate the spatial saliency at frame level implicitly. It is also worth mentioning that we just present the possibility to estimate clip-level saliency but did not use it into our video saliency estimation framework.

3.2.3 The Proposed Spatio-temporal Saliency

In this Section, we present the proposed spatio-temporal saliency method, which is just relying on temporal coherence at pairwise frame level. Moreover, without specifically combining the spatial saliency map, our method could implicitly incorporate the spatial saliency estimation results.

3.2.3.1 Temporal consistency

As one type of temporal correlations to estimate video saliency, temporal consistency targets to enhance the saliency or the salient feature which is consistently appeared in consecutive frames. There are methods to find co-occurring bases for pairwise image saliency estimation by measuring the KL-divergence between the probabilities of sparse bases as shown in Eq. 3.2 [18]. Only the bases which decrease the KL-divergence are the features of interest. The main disadvantage of this method lies into that the selected features may not be the features to each image saliency. Therefore, we first select the

features which are key to each image saliency and then among these selected features, we re-select some co-occurring features. Denote the weight of bases in two consecutive frames t and $t-1$ by $\mathbf{W}_t^s$ and $\mathbf{W}_{t-1}^s$. We update the weights of the co-occurring features by:

$$w_{n,t}^c = \begin{cases} w_{n,t}^s + w_{n,t-1}^s & \text{if } w_{n,t}^s \geq 0 \text{ and } w_{n,t-1}^s \geq 0, \\ w_{n,t}^s & \text{Otherwise.} \end{cases} \quad (3.8)$$

In other words, features repeatedly contributed to consecutive image saliency are the features of interest and their roles to video saliency estimation should be enhanced. Specifically, the weights corresponded to those features should be enlarged. Denote the weights of N bases by $\mathbf{W}_t^c = [w_{1,t}^c, w_{2,t}^c, \dots, w_{N,t}^c]^T \in \mathbb{R}^N$. The saliency value for the m_{th} patch located at (x, y) in the t_{th} temporal consistency saliency map is calculated as:

$$S_t^c(x, y) = \frac{1}{Z} \sum_{n=1}^N w_{n,t}^c \times |A_{n,m}^t|, \quad (3.9)$$

where $Z = \sum_{n=1}^N w_{n,t}^c$ is the normalization factor. Features selected in this way tend to detect the repeated image saliency as the video saliency. However, video saliency may not always be the static image saliency. In the following Section, we introduce a feature selection method to pick out the ones leading to large temporal novelty.

3.2.3.2 Temporal difference

There are cases that all foreground objects are static and only one object in the background is moving in a video clip. This moving object draws our attention while we may not notice this object in any single frame from this video. Hence, objects with obvious difference in the consecutive frames are considered equally important to our video saliency estimation. In previous Section, we obtain the sparse representation of each video frame. In order to measure human attention drawn by new information added, we describe the

m_{th} patch by the difference vector $\mathbf{S}_m^d \in \mathbb{R}^N$ between the sparse representations of the m_{th} patch A_m^t and its match $A_{m^*}^{t-1}$ in previous frame as:

$$\mathbf{S}_m^d = |A_m^t - A_{m^*}^{t-1}|, \quad (3.10)$$

where $A_{m^*}^{t-1} \in \mathbb{R}^N$ is the sparse representation of matched patch m^* in frame $t-1$ for the m_{th} patch. Here, a match patch means the most similar patch from the previous frame. We calculate the saliency value of the m_{th} patch located at (x, y) in t_{th} temporal difference based saliency map as the L_1 norm of $\mathbf{S}_m^d$:

$$S_t^d(x, y) = \|\mathbf{S}_m^d\|_1. \quad (3.11)$$

Based on the assumption that a basis frequently related to big frame changes is important, we can obtain the basis selection by setting their weights as $\mathbf{W}_t^d = [w_{1,t}^d, w_{2,t}^d, \dots, w_{N,t}^d]^T \in \mathbb{R}^N$, where each element $w_{n,t}^d$ is defined as:

$$w_{n,t}^d = \frac{\sum_{m=1}^{M_t} S_{m,n}^d}{\sum_n \sum_m S_{m,n}^d}, \quad (3.12)$$

where $S_{m,n}^d$ is the n_{th} element of vector $\mathbf{S}_m^d$. In Fig. 3.3, we show the temporal consistency saliency maps and the temporal difference saliency maps. From this figure, we can see the temporal difference based saliency maps highlight the differences between the consecutive frames while the temporal consistency saliency maps pay more attention to the common parts.

3.2.3.3 Fusion strategy

Considering the difficulty to use a fixed form to fuse different saliency maps (i.e. S^c and S^d), we obtain our video saliency map by fusing the bases which have large temporal consistency weight $\mathbf{W}_t^c$ and the temporal difference weight $\mathbf{W}_t^d$ simultaneously. That is:

$$\mathbf{W}_t = \mathbf{W}_t^c \cdot \mathbf{W}_t^d, \quad (3.13)$$

Figure 3.3: Sample results of temporal consistency saliency maps (second row) and temporal difference saliency maps (third row). Temporal consistency saliency maps highlight the shared saliency (e.g. the bodies of the tap and the fox) while temporal difference maps focus on the changes (e.g. the dropping water and the moving legs).

where $\cdot$ is the dot product and $\mathbf{W}_t \in \mathbb{R}^N$. During the implementation, we set $\mathbf{W}_t$ equal to the spatial saliency weight $\mathbf{W}_t^s$ when the previous frame is blank while the current one is not. Till now, our final video saliency map for the m_{th} patch located at (x, y) can be calculated as:

$$S_t(x, y) = \sum_n \frac{w_{n,t}}{\sum_{n=1}^N w_{n,t}} \times |A_{n,m}^t|, \quad (3.14)$$

where $w_{n,t}$ is the n -th element of $\mathbf{W}_t$. Note that the role of spatial saliency map S_t^s to S_t has already been implemented in S_t^c and thus our obtained saliency map is not only temporally novel but also spatially important.

A systematic overview of the algorithm in the proposed method is shown in Alg. 1. Some saliency estimation results obtained by the fused features are shown in Fig. 3.4. From this figure we can see our detection results correctly capture the spatio-temporally interesting parts (e.g. not only the falling water drops but also the body of the tap, not only the running soccer players but also the news titles, etc.).

Figure 3.4: Sample results of our video saliency detection in employed video datasets. The last video clip shows psychological pattern displayed in a video.

Algorithm 1: Sparse Feature Selection for Video Saliency Estimation

input : The learned dictionary D , the video sequence X .**output** : The video saliency map S .Obtain $\mathbf{W}_1^s$ via Eq. 3.4 and set $t = 2$;**while** $t \leq T$ **do**

- Get the sparse representation of frame t as A^t ;
- Select spatial salient features $\mathbf{W}_t^s$ via Eq. 3.4;
- Update $\mathbf{W}_t^c \leftarrow \{\mathbf{W}_{t-1}^s, \mathbf{W}_t^s\}$ via Eq. 3.8;
- Compute absolute temporal difference S_t^d via Eq. 3.10;
- Calculate temporal feature weight $\mathbf{W}_t^d$ via Eq. 3.12;
- Fuse $\mathbf{W}_t \leftarrow \{\mathbf{W}_t^c, \mathbf{W}_t^d\}$ via Eq. 3.13;
- Calculate the final video saliency map S_t via Eq. 3.14.
- $t = t + 1$.

Output the video saliency maps S .

3.3 Experimental Results and Analysis

3.3.1 Experimental Setup

3.3.1.1 Datasets

We validate our method on two public datasets: DB1 [14] and DB2 [82]. DB1 is also known as the CRCNS-ORIG dataset. It contains 50 video clips including indoor and outdoor scenes and the total time to play all videos is around 25 minutes. Eye tracking results are collected from eight subjects aged 23-32 with normal or corrected-to-normal vision. In all, there is total 12,211 saccades record. DB2 contains 10 short video clips of 5 to 10 seconds each. Every video includes one major object which moves across the whole video and draws human attention. Segmented objects are provided for all frames as the ground-truth for the binarized video saliency validation as [64] did.

3.3.1.2 Evaluation Metrics

To evaluate the accuracy of our method, we use both receiver operating characteristic (ROC) and Kullback-Leibler (KL) divergence as quantitative metrics. For both metrics, a higher value indicates better performance.

The ROC curve tells how well our saliency map matches fixated locations in the eye tracking experiments. By varying the threshold value to generate the binary saliency map, each spot in the ROC curve indicates the true-positive-rate on the y axis against the false-positive-rate value on the x-axis. For the convenience of our evaluation, the area under the ROC curve, i.e., denoted by AUC, is used to evaluate the performance of various algorithms.

A good saliency map should predict locations that human observers will gaze at. To compare different saliency models, we employ the KL-divergence to indicate how well the saliency map distinguishes between the distributions of saliency values at fixated locations and random locations in the obtained saliency maps. Saliency models which better predict human scan paths exhibit higher distances from random ones, as observers non-uniformly gaze towards a minority of regions with highest model responses. Details can be found from [3].

3.3.1.3 Implementation Details

Considering the scale of the object influences human attention a lot, we implement our method in a multi-scale way. Three different resolutions (i.e. 1, 1/2 and 1/4 of the original frame size) are employed and the proposed method is performed at every scale. The final saliency map is the summation result of three scales saliency maps by upscaling them into the original frame size. In the following sections, all experiments are performed on

the multi-scale color based video saliency estimation, unless specific description is given for single scale or gray-level frames.

3.3.2 Result Comparisons

In this Section, we compare the proposed method with five state-of-the-art saliency estimation methods.

3.3.2.1 Comparisons on DB1

In order to show the benefits by incorporating two kinds of temporal relationships, we compare our method to [1] and [36], which estimated video saliency based on the temporal coherence and the temporal difference. Since both methods only provides video saliency estimation results on one video (i.e. beverly03) in DB1, we also present the result comparisons restricted into this video with four other methods [2, 3, 1, 36]. Since they use the same evaluation measure (i.e. KL-divergence), we directly cite the results reported in [1] and [36]. Due to the unpublished source code of the video saliency estimation in [36], Fig. 3.5 only shows four saliency distributions and KL-divergence scores. From the KL ranking results, we can see our method obtains the best performance among [3, 1, 2]. We take the numerical results directly from [36]. Compared to our method, the KL score in [36] has been increased from 0.6927 to 0.7082. The results validate that by taking both the temporal coherence and the temporal difference, we get more accurate video saliency estimation.

More experimental results for all videos on DB1 are provided for [2] and [3]. Table 3.2 shows the corresponded AUC and KL-divergence results. It is obvious to see that, by incorporating the temporal information, our method outperforms them on video saliency

(a) model [2]

(b) model [3]

(c) model [1]

(d) Ours

Figure 3.5: Video saliency detection comparisons between models [2, 3, 1] and ours on "beverly03" from CRCNS-ORIG dataset via KL-divergence.

Table 3.2: Performance comparisons for surprise model [3], image saliency model (CIOFM for short) [2] and the patch variance model [14] to ours on DB1 by KL-divergence and AUC.

	Surprise	CIOFM	Variance	Ours
KL	0.221 ± 0.006	0.185 ± 0.006	0.122 ± 0.004	0.238 ± 0.007
AUC	0.679 ± 0.003	0.665 ± 0.003	0.625 ± 0.002	0.682 ± 0.002

estimation. Moreover, in order to show the advantages of our method on video saliency detection, we also compare our method to the surprise model [3] which also considered temporal correlations in video saliency estimation. We obtain the positive results once again in Table 3.2. All results of [3] and [2] are obtained by directly running the code provided in [14].

3.3.2.2 Comparisons on DB2

In order to show that the proposed sparse feature selection method outperforms the local contrast based methods which are based the sparse coding reconstruction error, we compare our method with [4]. From results in Fig. 3.6, we can see that our method obtains more completed boundaries than [4] even though multiple resolution is used in both methods. The reasons probably lie into that our video saliency is represented by the statistically selected features rather than the local reconstruction errors. Moreover, in order to further validate that our learned dictionary can improve video saliency estimation, we compare our performance with [1], in which a universally learned dictionary is used to estimate both the image saliency and the video saliency by selecting the ICA bases. Detection results of our method on DB2 are shown in Fig. 3.6. Quantitative results are also provided in Table 3.3, which can demonstrate that our learned dictionary can get better saliency representation than the universal features did.

Figure 3.6: Video saliency detection results of two video clips on DB2. The first row shows the original key-frames and the second row shows our results. Video saliency estimated from Hou’s method [1] and Li’s method [4] are shown in the third and the fourth rows, respectively.

Table 3.3: Comparisons among Hou’s method [1], the proposed method with a universal dictionary and the proposed method with our learned dictionary on DB1 via KL-divergence and AUC.

	Hou’s Method[1]	Ours-Universal-Dictionary	Ours-Learned-Dictionary
KL	0.159 ± 0.005	0.167 ± 0.006	0.238 ± 0.007
AUC	0.647 ± 0.002	0.651 ± 0.002	0.682 ± 0.002

To quantitatively compare our method with methods [1] and [4] on DB2, Fig. 3.7 shows the ROC curves for these three methods. The corresponding AUCs for our method, [1] and [4] are 0.87167, 0.7674 and 0.8289, respectively, which further validate that the proposed method outperforms the other two methods on this dataset.

Figure 3.7: The ROC comparisons of our method with Hou’s method [1] (ICL for short) and Li’s method [4] (CRC for short) on B2.

3.4 Discussion

In this Section, based on DB2, some discussions related to our method are provided.

3.4.1 The Clip-level Saliency vs. the Frame-level Saliency

In order to show the difference between the clip-level spatial saliency and the frame-level spatial saliency, sampled spatial saliency maps in both levels are shown in Fig. 3.8. As can be seen, for videos with complex content (the first row), features selected globally can characterize the saliency but it is not so discriminative to differ it from its background while the frame based spatial saliency map characterizes interesting parts in each frame. It is probably because such a large scale of patches involved in calculating the entropy gain dominates a single patch’s contribution to the saliency estimation.

Figure 3.8: Sample results of spatial saliency maps at clip level (the second row) and frame level (the third row). As can be seen, for a video with complex content (first video clip), features selected at clip level can characterize the saliency but they are not so discriminative and only the salient global structures are highlighted. For a video clip with simple background (second video clip), both methods perform similarly well.

Table 3.4: Comparisons the performances of our frame-level spatial saliency, temporal consistency saliency and temporal difference saliency in a single scale form on DB2.

	Spatial Saliency	Temp. Consistency	Temp. Difference	Fused One
KL	0.204 ± 0.024	0.211 ± 0.027	0.131 ± 0.015	0.233 ± 0.022
AUC	0.661 ± 0.007	0.663 ± 0.009	0.600 ± 0.005	0.668 ± 0.007

3.4.2 Spatial Saliency vs. Temporal Saliency

In order to show which saliency (i.e. frame-level spatial saliency, temporal consistency and temporal difference) in our method plays a more important role to video saliency estimation, we compare these saliency maps on DB2.

Fig. 3.9 compares the performance of our multi-scale saliency estimation method on DB2. From this figure we can see that temporal consistency saliency map provides comparable results with the spatial saliency map, while both of them outperform the temporal difference based saliency map. This is probably caused by the simplicity of using the temporal patch difference.

Table 3.4 shows our single scale results for four different saliency maps on DB2. Our

Figure 3.9: The ROC curves of our frame-level spatial saliency map, temporal correspondence saliency map and temporal difference based saliency map on DB2.

temporal consistency saliency map outperforms image saliency map. However, assisted by temporal difference, the fused saliency map generates even better results than the temporal consistency map does.

3.4.3 Multi-scale Saliency vs. Single-scale Saliency

Comparison results are also provided in Fig. 3.11 for our multi-scale and single-scale methods on DB2. From the ROC curves, we can see that multi-scale method outperforms the single-scale based method, which further validates the assumption that saliency occurs into any scale is the desired video saliency. Sampled results for multi-scale and single-scale saliency maps are shown in Fig. 3.10.

Figure 3.10: Sample results of single-scale based saliency maps (second row), multi-scale based saliency maps (third row) and original key frames in the first row.

3.4.4 Parameters to Dictionary Learning

3.4.4.1 Number of patches.

In order to avoid using a huge number of image patches to learn a dictionary for a video clip, we first uniformly sample key-frames from a video clip (e.g. one frame per two seconds). Then, a fixed number of patches are sampled from each key frame randomly. We use the sampled patches to train the dictionary. As the redundancy of many repeated scenes in video frames, this kind of down sample could ensure that each frame patch can be represented by the learned dictionary. Fig. 3.12 and Tabel 3.5 show the ROC curves and the AUC values with different patch numbers on DB2. We can see that the performance of our method varies a little when the number of sampled patches is around 10000. To balance the computational cost and the performance, in all our experiments, we randomly sample 10000 patches from each video to learn the dictionary.

Figure 3.11: The ROC performance comparisons for our multi-scale and single-scale saliency map estimation on DB1.

Table 3.5: The number of randomly sampled patches vs. AUC performance.

No. of Patches	10000	15000	20000
AUC	0.7844	0.7913	0.8089

3.4.4.2 Dictionary size.

Since the performance of our method depends on the sparse coding dictionary learning, we further discuss the influence of different dictionary sizes. To maintain an over-complete dictionary such that each patch can be well represented, the dictionary size should be larger than the patch dimension. Considering the computational complexity, we only test the dictionary size which is around four times of patch dimension. Fig. 3.13 and Tabel 3.6 show the ROC curves and the AUC values for our method with previously assigned various dictionary sizes. We can see that our method is not sensitive to dictionary size as it obtains comparable results with dictionaries of various sizes. Therefore, in all

Figure 3.12: The ROC performance comparisons of our method with different number of randomly sampled patches to learn a dictionary of size 1200 for each video clip on DB1.

Table 3.6: The learned dictionary size vs. AUC performance.

No. of Bases	1200	2400	3000
AUC	0.7844	0.7835	0.7937

experiments, we fix the dictionary size with 1200.

3.4.4.3 Patch features.

Performances of using different patch features are also compared on DB2. Two features are employed to represent each patch: the intensity value (i.e. $K = 256$) and the $L^*a^*b^*$ color feature (i.e. $K = 768$). Fig. 3.14 shows the corresponded ROC curves. Unsurprisingly, it shows that $L^*a^*b^*$ color feature achieves better performance than using the intensity alone.

Figure 3.13: The ROC performance comparisons of our method with different dictionary sizes on DB1. For each video, the dictionary is learned from 10000 randomly sampled patches.

3.5 Conclusion

We propose a novel video saliency estimation method, which is obtained by fusing spatio-temporally selected sparse bases. By leveraging the sparse coding framework, we can better characterize a video content by a dictionary individually learned on the given video. Based on the learned dictionary, we analyze video saliency estimation at multiple levels (i.e. frame level, pairwise frame level and clip level). Unlike the spatial saliency estimated at the frame level and the clip level, we propose to obtain the spatio-temporal saliency at pairwise frame level by selecting temporally consistent and temporally novel bases for representation. Our fusing scheme which is via multiplying all selected features together to represent the target video saliency avoids the difficulties to choose a combination form for individual spatial saliency map and the temporal saliency map fusion. Since our temporal consistency enhanced the spatial saliency, it also ensures the obtained

Figure 3.14: The ROC performance comparisons of our method with $L^*a^*b^*$ color feature and intensity feature on DB1.

video saliency is not only spatially important but also temporally consistent and novel. Our method also fully incorporates the temporal correlations and achieves a better content representation by our learned dictionary. Extensive experiments on two public video datasets demonstrate that our method outperforms the state-of-the-art saliency estimation methods and works effectively on a wide range of video genres. Considering multiple frame coherence into sparse feature selection in the proposed method will be one of our future works.

Chapter 4

Saliency Density Maximization for Salient Objects Discovery

4.1 Introduction

In Chapter 3, we have introduced a video saliency estimation method. A lot of work have also been proposed to generate the saliency map of an image or a video frame [1, 5, 6, 7, 32, 16, 44, 38, 33], where the intensity of the map indicates how salient each pixel is. Although different saliency models have been generated, accurate detection of the salient object in a saliency map remains a challenging problem. It is not uncommon that the obtained saliency map is noisy and incomplete. As an example shown in Fig. 4.1, only several salient parts of the flower are highlighted in the saliency map, while the rest are missing. Due to the distraction from the cluttered background, it is not easy to accurately locate the salient object and crop it out. To handle this problem, some existing methods generate saliency maps by supervised learning methods and find the minimum rectangular region that covers a fixed amount of salient points, e.g. 95 % of

the total salient points [5, 58]. The major limitation is that it is troublesome to train the model or difficult to predefine the amount of saliency the salient object should contain, as it depends on the size and the shape of the salient object, as well as how cluttered the background is. Such a detection method thus cannot be adapted to various sizes of the salient objects. Instead of cropping the salient object with a bounding box, some other methods rely on object segmentation to obtain an accurate object shape. This category of methods, however, is not robust to the cluttered background and usually requires prior knowledge to help an accurate segmentation [6, 16].

To address the above mentioned problems, we propose a novel unsupervised method to discover salient objects from various saliency maps of images or videos. Given a saliency map of spatial size $m \times n$, our goal is to detect a sub-image of smaller size $m' \times n'$ that well covers the salient object. To enable a robust detection, given the saliency map of an image, we propose to locate the object by finding a bounding box of the maximum saliency density (MSD).

Figure 4.1: Our salient object detection result using the saliency map proposed in [1]. The first image is the original saliency map. The second image is the binarized saliency map. The third image is the result of salient object localized by maximum saliency region detection. The red rectangle is the detected result while the blue one is the ground-truth from [5]. The last one is our result via saliency density maximization. Better to see in color.

As a new formulation of salient object detection, it balances the size of the object

and the saliency it should contain, thus can be adapted to the various sizes of the salient object. Moreover, it can tolerate the noise and incompleteness in the saliency map due to the cluttered background and the imperfectness of the saliency map. For example, as shown in Fig. 4.1, even though the salient pixels distribute sparsely, the detected bounding box with the highest saliency density accurately crops the salient object out. The proposed method does not require any prior knowledge of the salient object and is an unsupervised method. To avoid an exhaustive search of all possible bounding boxes, a branch-and-bound search algorithm is proposed to efficiently find the optimal one with maximum density rather than largest classification score [61]. Experimental results on a public dataset of five thousand images show that the performance of our salient object discovery is comparable to the state-of-the-art learning based salient object detection. The testing on four different types of saliency maps ([1, 6, 7] and a fused saliency map) show our method can perform equally well.

Besides object detection in images, the proposed method can be easily extended to salient object detection in videos as well. The main difference between salient object detection in images and videos is that the salient object may move in videos. Given a sequence of saliency maps for a video shot, instead of identifying the salient object at each frame, we search for a sub-image region which covers not only salient objects of the current frame but also its previous peers'. Therefore, the detected sub-image contains most of the saliency of the video shot and keeps the temporal smoothness of the selected regions. We define the sub-image as the salient object for a video shot, which has the maximum saliency density on the global saliency map (name as importance map I_{map} , see Section 4A for more details) of all the frames. With this definition and detection method for video salient objects, the obtained sub-image can be cropped and applied to video retargeting. Such a cropping-based video retargeting does not distort the salient objects

as seam carving, or introduce dithering under complex camera motions as in [9, 10]. The evaluations on a few video shots show that the proposed method extended in videos provides promising results even on challenging videos with dynamic and complex camera motions.

4.2 Salient Object Detection in Images

Given a single image I and its associated saliency map S , where $S(x, y)$ indicates the saliency value of the pixel at (x, y) , our goal is to accurately locate the salient object, i.e. to locate a salient sub-image $W \subseteq I$ with maximum saliency density. We first review some existing methods, then propose our new approach.

4.2.1 Existing Schemes

4.2.1.1 Minimum Rectangle with Fixed Saliency (MRFS)

Previous approaches treated a salient object as a smallest rectangle containing at least a fixed percentage of salient pixels (e.g. 95 %) from saliency maps [56, 6] or binary saliency maps [5] and the rectangle is obtained by exhaustively searching. As in [5], the binary saliency map is adopted and the formulation of salient object detection by MRFS can be written as in Eq. 4.1:

$$W^* = \arg \min_{W \subseteq I} A(h(W)) \quad (4.1)$$

$$h(W) = \{W | \frac{\sum_{(x,y) \in W} S_b(x, y)}{\sum_{(x,y) \in I} S_b(x, y)} \geq \lambda\},$$

where $A(h(W))$ measures the area size of $h(W)$ and S_b is the binary image of S . We set $S_b(x, y)$ to 1 when $S(x, y) \geq \tau$ and $S_b(x, y)$ to 0 when $S(x, y) < \tau$. Here, τ is the threshold;

W is the sub-window of the whole image region I and λ is the fixed percentage. The brute force method works, however, it is not time efficient and λ is heuristically decided.

4.2.1.2 Maximum Saliency Region (MSR)

Other approaches proposed that salient object detection can be solved as object localization by an efficient sub-window search [16]. Pixels on saliency map are scored by a binary classifier (salient and non-salient) and the salient object is the sub-window with maximum classification score. Based on the maximum sub-array problem [61, 83], the sub-window can be found by an efficient sub-window search. Therefore, the idea of salient object detection in [16] can be formulated as in Eq. 4.2:

$$W^* = \arg \max_{W \subseteq I} h(W) \quad (4.2)$$

$$h(W) = \sum_{S_b(x,y) \in W} S_b(x,y),$$

where $S_b(x,y)$ is obtained in a similar way in Eq. 4.1 with a slight difference that $S_b(x,y) = -1$ when $S(x,y) < \tau$. From Eq. 4.2, the salient object is located with the region W^* that contains the maximum of saliency. We call this method as the maximum saliency region. However, there are two major limitations of this method: (1) it highly relies on the selection of threshold τ , which is difficult to optimize; (2) when the binary saliency map is sparse, it intends to detect a small region.

4.2.2 Our New Approach

Different from aforementioned methods, we present our new formulation and solution for the salient object discovery problem in this Section.

Figure 4.2: (a) Original image (b) Sparse saliency map by [1] which highlights edges and corners (c) Dense saliency map by [6] which highlights overall salient regions.

4.2.2.1 Maximum Saliency Density

Since different saliency map generation methods emphasize different aspects, the obtained saliency map could be sparse or dense, as shown in Fig. 4.2.

Sparse saliency map accurately detects the salient parts of the object but the boundary of the salient object is not well defined. Dense saliency map represents the salient object completely but some cluttered background is also included in the detection result. Despite different types of saliency maps, we notice that the average density of the region of salient object is much larger than that of any other regions on the saliency map.

This motivates us to propose the salient object detection with maximum saliency density from the raw saliency map S . As a result, there is no need to select the threshold τ and the fraction ratio λ . Moreover, it balances the size of the salient object when the saliency map is sparse. To handle different types of saliency maps, we formulate our objective function $f(W)$ as:

$$f(W) = \frac{\sum_{(x,y) \in W} S(x,y)}{\sum_{(x,y) \in I} S(x,y)} + \frac{\sum_{(x,y) \in W} S(x,y)}{C + A(W)}. \quad (4.3)$$

Here C is a positive constant to balance $A(W)$ which is the area of (W) . The first term in $f(W)$ prefers that W contains more salient points, while the second term ensures that the detected region W is of high quality in terms of the saliency density. Therefore, by

finding the optimal bounding box W^* that maximizing the two terms together in $f(W)$ as in Eq. 4.4, we balance the size of the salient object and the saliency it contains. We call our new formulation as the maximum saliency density (MSD).

$$W^* = \arg \max_{W \subseteq I} f(W). \quad (4.4)$$

4.2.2.2 Constrained Maximum Saliency Density

Depending on applications, we would have prior knowledge or requirement on the aspect ratio or the size of the salient object. For example, video retargeting requires the selected region has the same aspect ratio as the target display, to avoid visual distortion. Therefore, we further present an aspect ratio preserved solution to find the global optimal W^* .

$$\begin{aligned} W^* &= \arg \max_{W \subseteq I} f(W) \\ f(W) &= \frac{\sum_{(x,y) \in W} S(x,y)}{\sum_{(x,y) \in I} S(x,y)} + \frac{\sum_{(x,y) \in W} S(x,y)}{C + A(W)}, \\ \text{s.t. } & r_0 - \Delta r \leq r \leq r_0 + \Delta r \end{aligned} \quad (4.5)$$

where r_0 is the target aspect ratio and r is the aspect ratio of region W . Δr is an offset allowed r to deviate slightly from r_0 . This is based on the assumption that we allow slightly distortion if important information can be included in the detected region.

4.2.2.3 Detection Algorithm

Exhaustive search of W^* from either Eq. 4.4 or Eq. 4.5 is time consuming. $W^* = [T, B, L, R]$ contains four parameters, where T , B , L , R are the top, bottom, left, and right positions of W^* , respectively. Suppose the frame is of size $m \times n$, the original hypotheses space is $[0, n - 1] \times [0, n - 1] \times [0, m - 1] \times [0, m - 1]$, where we need to

pick up T , B , L , R from each dimension respectively. To solve this combinatorial optimization problem, an exhaustive search is of complexity $O(m^2n^2)$. A branch-and-bound search method is proposed in [61] to accelerate the search by recursively partitioning the parameter space and pruning the sub-space based on the calculated upper bound until it reaches the optimal solution. Such a branch-and-bound search can lead to the exact solution as the exhaustive search, while with a practical complexity of only $O(mn)$. The details of the branch-and-bound search can be found in [61].

Despite its efficiency, the original branch-and-bound only works with both positive and negative values. However, in our case, the saliency map only contains positive elements and we do not want to deliberately introduce negative pixels. Therefore we need to develop our own branch-and-bound search algorithm. Considering the efficiency of the branch-and-bound searching method depends on the upper bound estimation, we derive the upper bound of our $f(W)$ first. Denote the set of regions by $\mathbb{W} = \{W_1, \dots, W_i\}$, where each $W_i \subseteq I$. Suppose there exists two regions W_{min} ($W_{min} \in \mathbb{W}$) and W_{max} ($W_{max} \in \mathbb{W}$), such that for any ($W \in \mathbb{W}$), $W_{min} \subseteq W \subseteq W_{max}$. Given the set $\mathbb{W}$, we denote by $\hat{f}(\mathbb{W})$ the upper bound estimation of the best solution that can be found from $\mathbb{W}$. In other words, we have $\hat{f}(\mathbb{W}) \geq f(W)$, $\forall W \in \mathbb{W}$, using W_{min} and W_{max} , the upper bound of Eq. 4.4 can be estimated as:

$$\hat{f}(\mathbb{W}) = \frac{\sum_{(x,y) \in W_{max}} S(x,y)}{\sum_{(x,y) \in I} S(x,y)} + \frac{\sum_{(x,y) \in W_{max}} S(x,y)}{C + Area(W_{min})}. \quad (4.6)$$

As a similar scenario, the upper bound for Eq. 4.5 is derived as Eq. 4.7. Mathematical notations here follow Eq. 4.6.

$$\begin{aligned} \hat{f}(\mathbb{W}) &= \frac{\sum_{(x,y) \in W_{max}} S(x,y)}{\sum_{(x,y) \in S} S(x,y)} + \frac{\sum_{(x,y) \in W_{max}} S(x,y)}{C + A(W_{min})}, \\ \text{s.t. } r_{min} - \Delta r &\leq r_0 \leq r_{max} + \Delta r \end{aligned} \quad (4.7)$$

where $r_{\min} = \text{Width}(W_{\min})/\text{Height}(W_{\max})$ and $r_{\max} = \text{Width}(W_{\max})/\text{Height}(W_{\min})$. The expected aspect ratio r_0 for region W^* is also bounded by $W_{\min}$ and $W_{\max}$. Thus, $f(W^*)$ is the optimal solution according to the set $\mathbb{W}$.

Algorithm 2: Aspect Ratio Constrained MSD Salient Object Detection Algorithm

Data: Image saliency map $S \subseteq R^{m \times n}$ and upper-bound function $\hat{f}(\mathbb{W})$ as Eq. 4.7

Result: $W^* = \arg \max_{W \subseteq S} f(W)$

begin

 Set P as an empty priority queue

$\mathbb{W} \leftarrow [0, n-1] \times [0, n-1] \times [0, m-1] \times [0, m-1]$

while $\mathbb{W}$ contains multiple windows **do**

$\mathbb{W}_1 \cup \mathbb{W}_2 \leftarrow \mathbb{W}$ and $\mathbb{W}_1 \cap \mathbb{W}_2 = \emptyset$

for $i \leftarrow 1$ **to** 2 **do**

 Find $W_i^{\min}$ and $W_i^{\max}$ from $\mathbb{W}_i$

 Calculate $r_{\min}$ and $r_{\max}$ for $\mathbb{W}_i$

if $(r_{\min} - \Delta r \leq r_0 \leq r_{\max} + \Delta r)$ **then**

 Push $(\mathbb{W}_i, \hat{f}(\mathbb{W}_i))$ into P

 Retrieve the top state $\mathbb{W}$ from P

Based on the upper bound estimation, our salient object detection algorithm is proposed. Here, only the constrained MSD salient object detection algorithm is given and shown in algorithm 2. Detection procedure without constraint is easily obtained by removing the *if* sentences in the algorithm.

4.3 Salient Objects Detection in Videos

To detect salient objects in videos, videos are first separated into shots with the consideration that salient objects are moving smoothly within a shot. Given a video shot $\{I_1, \dots, I_t\}$ (t is the frame number of the video shot), its corresponding saliency map sequence $\{S_1, \dots, S_t\}$ is extracted. After getting the spatio-temporal saliency map, three

steps are taken to detect a salient sub-image W across all frames. Firstly, salient object is obtained by MSD salient object detection algorithm at each frame. An importance map is then constructed by the frame salient object and the global optimal salient sub-image for a video shot is finally obtained by searching on the importance map.

4.3.1 Salient Object Detected in Each Frame

Similar to salient object detected in images, various kinds of video saliency map estimation methods could be used in our method. Given the saliency map sequence generated by a specific method, for each frame, a salient object W_t^* is obtained by employing our maximum saliency density (MSD) salient object detection algorithm. Without any constraints, the result W_t^* tightly bounds the salient object for the current frame. Moreover, by detecting W_t^* , visual saliency caused by cluttered background is partially removed in each frame. It thus helps the global salient object detection in the video shot.

4.3.2 Importance Map and Globally Optimal Salient Sub-image

Salient object W_t^* detected in each frame cannot be treated as the final cropped region for each frame, as it does not incorporate any motion smoothness among different W_t^* . Here, W_t^* is actually a local optimal solution, which labels the salient object in an individual frame I_t . For a video shot, a global optimal W^* is related to W_t^* but different from W_t^* . To find the importance for other points with regard to the whole video shot, we generate an importance map I_{map} based on the frequency of a point $S(x, y, t)$ falling into W_t^* . The large value of $I_{map}(x, y)$ indicates that the point (x, y) attracts our attention in the whole video sequence frequently. Hence, I_{map} serves as a mean filter to filter out outlier salient

points that only appear in a few frames while not significant for the whole video:

$$I_{map}(x, y) = \frac{1}{N} \sum_t g(S(x, y, t)) \quad (4.8)$$

$$g(S(x, y, t)) = \begin{cases} 1 & S(x, y, t) \in W_t^* \\ 0 & S(x, y, t) \notin W_t^*. \end{cases}$$

Here N is the total number of frames and normalizes all values in the important map into $[0, 1]$. The constrained MSD algorithm as described in algorithm 2 can be used to search W^* in I_{map} which is the globally optimal salient sub-image for a video shot.

4.4 Experimental Results

In this Section, the proposed method is evaluated with extensive experiments from two aspects: salient objects detected from images and videos, respectively.

4.4.1 Salient Object Detection in Images

4.4.1.1 Database

In order to evaluate the results, the benchmark dataset in [5] is used to test our algorithm. The dataset provides 5000 images with the average spatial resolution of 300×400 or 400×300 . Each image contains a salient object and each salient object is labeled by nine users by drawing a bounding box around it. Since different users have different preference of saliency, only the point voted more than four times is considered as the salient point. The averaged saliency map S_g is then obtained from the user annotation. To obtain the better performance of each saliency model especially for those sensitive to the image scales, we resize all the images and its corresponding ground truth masks

into $3/4$ of the original size. That means the spatial resolution of all images is set to $m \times n = 225 \times 300$ or $m \times n = 300 \times 225$.

Figure 4.3: Detection of MSD, MRFS and MSR. The first row shows saliency maps using Hou’s [1], Achanta’s [6] and Bruce’s [7] methods, respectively. The second row shows the localization results by MRFS with the exhaustive search in the three saliency maps. The third row shows results by MSR. The results in the third row are by MSD. Detected results are labeled with red lines. Blue rectangles are ground-truth from [5]. Better to see in color.

4.4.1.2 Comparison of MSD with MRFS

First of all, we compare the proposed MSD with minimum rectangle with fixed saliency (MRFS). The parameter λ is set to 95% as [5] suggested. Fig. 4.3 shows the results obtained by MRFS and our method in the second and the last rows, respectively. As 95% is a fixed parameter and not determined by the content of the saliency map, only parts of the salient object are detected on Hou’s saliency map. It is mainly because most of salient points are the boundaries of the salient object in [1]. On contrary, the detected results of MRFS include a large part of the nonsalient object area on Achanta’s and Bruce’s saliency maps. While in our method, more salient object area is included owing to the first term in Eq. 4.3 on Hou’s saliency map, and small salient area away from main

Figure 4.4: Precision, recall and F-measure for MRFS with four λ s {95%, 90%, 85%, 80%} on Hou's saliency map.

salient object is dropped under the constraint of the saliency density on Achanta's and Bruce's saliency maps. Therefore, our result is more accurate than MRFS. Precision, recall and F-measure on Fig. 4.5 further validate our claim. In order to evaluate λ , we test four different λ values as shown in Fig. 4.4 shows. Precision is improved while recall is reduced as λ becomes smaller and it is difficult to choose an optimal λ .

4.4.1.3 Comparison of MSD with MSR

To compare MSD with MSR, we test both methods on three different saliency maps. The threshold τ is obtained by Otsu [84] for all saliency maps in MSR. The parameter C is set to be 0.9, 0.03 and 0.24 of the size of $m \times n$ for Hou's, Bruce's and Achanta's saliency maps respectively in MSD through parameter evaluation. Fig. 4.5 shows the comparison results. The average of precision, recall and F-measure are reported on each group. On

(a) Hou's saliency map [1]

(b) Bruce's saliency map [7]

(c) Achanta's saliency map [6]

Figure 4.5: Performance comparisons among MSD, MRFS and MSR on three types of saliency maps.

Hou’s saliency map, edges/corners are detected as salient parts. By using MSR, very small region is bounded while larger salient regions are detected by MSD. F-measure and recall are significantly improved by MSD. For the other two saliency maps, MSD also outperforms MSR. The results on three different types of saliency maps show that MSD improves F-measure, and at the same time, keeps the high precision rate.

4.4.1.4 Evaluation on Different Saliency Maps

Since the salient object detection result is based on the saliency map, the more accurate the saliency detection is, the better performance the object detection method obtains. It is worth noting that a single salient object in [5] is obtained through supervised learning. Both [32] and [16] have prior knowledge about the region size provided by image segmentation. Thus, they are not directly comparable with the proposed method. However, even without any prior knowledge of the salient object, MSD on Bruce’s saliency map outperforms Ma’s method [8] which directly uses detected salient object (F-measure 61%) and search result on Itti’s saliency map [2] which finds the minimum rectangle containing at least 95% salient points by the exhaustive search (F-measure 69%). For MSD on Hou’s saliency map, it obtains comparable result compared with the exhaustive search on Itti’s saliency map. Though our result on Achanta’s saliency map is not as good as the result on [2], the precision is still higher than searching result on Itti’s and Ma’s saliency maps.

Various saliency maps estimation methods are proposed recently. Evaluating the efficiency of them itself is an interesting problem. A small range comparison is performed as shown in Fig. 4.6. Since different bottom-up saliency map generation method has different advantages and disadvantages, to minimize the influence to the saliency object detection result, four saliency maps from [7, 1, 15, 6] are fused together. Each saliency

Figure 4.6: Comparison of our MSD with other salient object detection results 1: Ma’s saliency map and their salient object detection result [8]; 2: MRFS with $\lambda = 95\%$ on Itti’s saliency map [2]; 3: MSD on Hou’s saliency map; 4: MSD on Bruce’s saliency map; 5: MSD on Achanta’s saliency map; 6: MSD on the fused saliency map; 7: Supervised method from [5].

map is normalized into $[0, 1]$ and the combined saliency map is obtained by integrating them together, then normalizing the summation into $[0, 1]$. Different weights are tested and the ratio $2 : 1 : 1 : 2$ for saliency maps [7, 1, 15, 6] respectively, which gets the best performance, is selected. As shown in Fig. 4.6 for our method 6, after combining four saliency maps together, F-measure is further improved to 75.84% compared to the best individual saliency map result from Bruce’s saliency map 73.06%. The performance is close to the learning based salient object detection results in [5] which requires training and has much higher computation cost.

4.4.1.5 Parameter Evaluation

Two groups of experiments are performed to evaluate the influence of the only parameter C in the proposed method. Firstly, we present the optimal C selections against different saliency maps. When C is small, the method is sensitive to the density change and prone to converge to a region with high average density but relative smaller size. When

Figure 4.7: Evaluation C-value for MSD on (a) Hou's saliency map, (b) Bruce's saliency map, (c) Achanta's saliency map and (d) Combined saliency map. x-coordinate is C-value measured in unit $10^4 \times m \times n$ and y-coordinate is the corresponding averaged F-measure.

a large value of C is selected, the density term becomes trivial in the objective function $f(W)$ and the whole algorithm converges to a larger region with a lower average density. Hence, the parameter C plays the role of balancing the size of salient object and its value is also closely related to the image size. Therefore, the ratio to the image size $m \times n$ can be decided to represent C value. The relationship between C value and F-measure is shown in Fig. 4.7. In Fig. 4.7(a), within the range of $[0.5, 1.3]$, the F-measure obtains its best performance ranging from $[68\%, 68.9\%]$. Overall, it is not sensitive to the selection of C . Similarly, when C is in the range $[0.026, 0.045]$, the F-measure is above 72% in Fig. 4.7(b); when C in the range $[0.21, 0.36]$, the F-measure is above 63.5% in Fig. 4.7(c).

From these results, we can see that the region based saliency map has a smaller optimal C than edge/corner based methods (Fig. 4.3 shows Bruce’s saliency map is denser than Achanta’s). That further indicates that the density term in Eq. 4.3 is important when the salient points are densely distributed on the saliency map.

Figure 4.8: Evaluation C-value for three individual datasets which are separated groups from MSRA dataset on (a) Hou’s saliency map, (b) Combined saliency map. x-coordinate is C-value measured in unit $10^4 \times m \times n$ and y-coordinate is the corresponding averaged F-measure.

Secondly, for a particular saliency map, we show that the selection of parameter C is not sensitive to different datasets. MSRA dataset is separated into three non-overlapped groups, each of which can be treated as an individual dataset. On each dataset, we get a group of C values and its corresponding F-measure. We test on two saliency maps and more saliency models can also be adopted. The comparisons between the overall results (solid red curve) from Fig. 4.7 and three groups’ (dot blue curves) are shown in Fig. 4.8. Results in Fig. 4.8 further validate that different datasets under the same saliency map get the optimal performance around the same C and within one dataset it is not very sensitive for the parameter C to get the optimal F-measure.

Table 4.1: Averaged time cost $\pm$ standard deviation by seconds for each image.

SMap	[1]	[7]	[6]
MRFS [5]	22.235 $\pm$ 5.85	23.3587 $\pm$ 3.47	22.1464 $\pm$ 5.75
MSR [16]	0.0039 $\pm$ 0.0073	0.0056 $\pm$ 0.0027	0.0048 $\pm$ 0.0036
MSD	0.0113 $\pm$ 0.011	3.4718 $\pm$ 2.55	0.0351 $\pm$ 0.058

4.4.1.6 Efficiency

Tabel 4.1 shows the average computational time and the standard deviation tested on 5000 images for MRFS by the exhaustive search, MSR and MSD based on Hou’s, Bruce’s and Achanta’s saliency maps. From Table 4.1, we can see that our method is much faster as compared to MRFS by the exhaustive search and is comparable to MSR search. The algorithm is tested on a Duo Core desktop of 2.66GHz, implemented with C++.

Figure 4.9: More salient object detection results by MSD. The first row shows our results on Hou’s saliency map [1]. The second row shows our results on Achanta’s saliency map [6] and the results in the last row are based on Bruce’s saliency map [7]. The red rectangle is the detected result while the blue one is the ground truth from [5]. Better to see in color.

4.4.2 Salient Object Detection in Videos

In order to demonstrate the good performance of our method, we conduct our video salient object detection experiments on two kinds of video saliency maps. The first one is a simple combination method and the other is based on the saliency maps generated via the proposed video saliency map estimation method in Chapter 3. Other video saliency estimation methods can be used in our method as well.

4.4.2.1 Salient Video Object Detection on the Combined Saliency Maps

In this Section, we first introduce the way to generate the simple video saliency map, then a simulation experiment is conducted to show the general idea of our video salient object detection. Next, experiments on real videos and an application of our method to video retargeting are presented accordingly.

1. Spatio-temporal Saliency Map Extraction

To generate the spatio-temporal saliency map $S(x, y, t)$ for the frame I_t , we define the weighted linear combination of the spatial saliency map S_I and the temporal saliency map S_M as the video saliency map:

$$S(x, y, t) = (1 - \omega) \times S_I(x, y, t) + \omega \times S_M(x, y, t), \quad (4.9)$$

where S_I is the spatial saliency map proposed in [6] and S_M is the frame difference to characterize the motion saliency. Here, $\omega \in [0, 1]$ is the weight to balance two terms and we set $\omega = 0.8$ to emphasize the motion saliency. Based on the assumption that the fixation of humans is always in the image center, fused saliency map is weighted by a near flat Gaussian $G(x - x_0, y - y_0, \sigma^2)$ as the attention priori:

$$S(x, y, t) = S(x, y, t) \times \exp\left(-\frac{(x - x_0)^2 + (y - y_0)^2}{2\sigma^2}\right). \quad (4.10)$$

Here, σ^2 is taken larger enough to make the priori value decay slowly when it approaches the frame boundary. (x_0, y_0) is the image center. Semantic features such as face [77, 85], visual concepts [86] could also be added to improve the saliency detection.

2. Simulation on a Video Shot with Human Labeled Ground-truth

Due to lack of the ground-truth for salient object detection from videos, it is difficult to measure the efficiency of the proposed method quantitatively. We manually labeled a simple table tennis video shot and simulate our method on it to better illustrate our ideas and the efficiency on salient object detection in videos. In the table tennis video clip, the player is the dominant and salient object while the ball and the shadow of the player are neglected because of either the small size or low contrast.

Figure 4.10: The first row: sampled saliency maps of the table tennis video clip; The second row: corresponding detected bounding boxes for frame salient object in green and the human labeled ground truth in blue; The third row: intermediate states of constructing I_{map} ; The last row: the global bounding box with the target aspect ratio $r_0 = 9/16$ in red and the human labeled global ground-truth in blue.

Our aims here are two folds: detect the player per frame and select the best region to represent the whole video. In each frame, the ground-truth of the frame salient object is the blue bounding box of the player shown in the second row of Fig. 4.10. Since the player is moving within a limited range, the best region selected for displaying the whole video

is the largest place the salient object can reach. This region is set to be the ground-truth of the global salient object detection. However, there is no aspect ratio constraint to this global ground-truth because it is barely based on the motion range of the salient object. When the target screen has aspect ratio constraint different from the largest place the salient object can reach, the size of the global ground-truth should be adjusted. To fairly and easily compare, we choose $\Delta r = 0$ and $r_0 = 9/16$ which is close to the aspect ratio of the ground-truth and no further adjustment is needed.

Sample frames and their corresponding detection results are shown in Fig. 4.10. The sampled saliency maps in the first row of Fig. 4.10 show that the player keeps changing its shape and position as it moves. The green bounding boxes in the second row of Fig. 4.10 are our object detection results in single frame, which are compared to the blue rectangles (ground-truth of frame salient objects). That further validates that MSD method can adaptively detect the salient object regardless of its shape and position changes. The red bounding boxes in the third row of Fig. 4.10 are our global salient sub-image detection results constrained with aspect ratio 9/16. We can see that our results are close to the ground-truth with the assigned aspect ratio. The slight difference is because the predefined region is the largest region the salient object can reach while our method is to detect the region the salient object frequently appears. Specifically, given the binary mask of our detected region S_d in red and the ground-truth binary mask S_g in blue, we calculate precision and recall as in Eq. 5.4.1:

$$\begin{aligned} precision &= \frac{\sum S_g \times S_d}{\sum S_d}, recall = \frac{\sum S_g \times S_d}{\sum S_g} \\ F - measure &= \frac{(1 + \alpha) \times precision \times recall}{\alpha \times precision + recall}, \end{aligned} \quad (4.11)$$

where α is a positive constant which weights the precision over recall while calculates F-measure. We take $\alpha = 0.5$ as suggested in [16, 5].

Table 4.2: Precision, recall and F-measure for frame salient object and global salient sub-image detection in table tennis video clip.

$r_0 = 9 : 16$	<i>Frame</i>	<i>Global</i>
<i>precision</i>	73.69%	82.30%
<i>recall</i>	93.36%	86.67%
<i>F - measure</i>	79.26%	83.71%

The performance on precision, recall and F-measure for our method in the table tennis video is given in table 4.2, from which we can see our detection method can correctly locate the salient object either in a single frame or in a video shot.

3. Experiments on Real Videos and its Application

In the implementation of salient object detected in real videos, we first process the source video to identify shot boundaries and the detection is performed within each shot. Eleven high definition video shots with not common 4 : 3 aspect ratio style are tested. Table 4.3 shows their detailed information. All cinema movies are provided by [87]. Two sport videos downloaded from Youtube are 2009 NBA All Star Game and 4x100m Relay World Record in Beijing Olympic Game. The cartoon sequence was used for video retargeting in [10]. In order to show the effectiveness of the proposed method, we apply the frame salient object detected results and the global sub-image detected results to video retargeting. Video retargeting aims to change the resolution of video sources while faithfully convey the original information. Cropping-based video retargeting is employed to compare our approach with two baseline methods on adapting a video for better viewing on a small screen. Our results are obtained by detecting the global salient sub-image from the video shot first, then cropping the region out and finally resizing it into the target size 320×240 which is a typical resolution for today's mobile devices. Results of two baseline methods are obtained by directly scaling the original videos into

Table 4.3: Basic information for testing video shots.

Name	Size	r_0	No. of frames
Movie_Bigfish	448×240	1.87 : 1	156
Movie_Graduate1	560×240	2.33 : 1	460
Movie_Graduate2	560×240	2.33 : 1	1201
Movie_Graduate3	560×240	2.33 : 1	109
Movie_Bufferfly Effect	404×244	1.66 : 1	91
Movie_Mission to Mars	560×240	2.33 : 1	496
Movie_Being John	448×240	1.87 : 1	141
Movie_Forrest Gump	544×240	2.27 : 1	54
Cartoon_Simpson	480×212	2.26 : 1	141
Sports_2009 NBA	640×340	1.88 : 1	585
Sports_4x100m Relay	558×480	1.16 : 1	349

the expected size or by padding the videos into expected aspect ratio then scaling.

First, experimental results of salient object detection per frame are shown in Fig. 4.11(a). In practice, there are usually multiple salient objects existing in one scene. Here, we find them simultaneously with arbitrary shape of bounding boxes to fit their sizes.

After obtaining the frame salient objects sequentially, I_{map} is built for each shot as shown in Fig. 4.12. Non-salient object points are filtered out from I_{map} and the global salient sub-image are obtained by the Algorithm 2. We set the desired aspect ratio $r_0 = 4/3$ and $\Delta r = 0.1$ which is flexible to cover more saliency with little sacrifice of aspect ratio. Red bounding boxes in Fig. 4.11(b) and Fig. 4.11(c) contain most of the saliency of I_{map} which is equivalent to capture most of the saliency of the video shot. Results for central cropping in yellow show the adaptiveness of our method. Detection results on Sports 4x100m Relay demonstrate the robustness of our method on videos which have complex camera motions.

Comparisons between [10, 9] are performed to show the efficiency of our method on

(a)

(b)

(c)

Figure 4.11: (a) and (b) are frame salient objects (rectangles in green) detected by MSD without aspect ratio constraint. (c) and (b) are comparison of our selected salient sub-image regions (rectangles in red) between central cropping results (rectangles in yellow). Better to see in color.

Figure 4.12: Intermediate states of constructing I_{maps} for sports video 4×100 relay (in the top row), Simpson video (in the middle row) and the movie Graduate2 video (in the third row).

avoiding virtual camera motion. We use the position of bounding box changes to indicate the virtual camera motion involved. All comparisons are performed on the original video results post by [10, 9]. Green bounding boxes in Fig. 4.13(a) are cropping markers for video retargeting in [9]. Since the source video is a news video without camera motion, slight changing of the cropping markers will lead to obvious virtual camera motions which are unfavorable. While our results with red bounding boxes in Fig. 4.13(a) capture most of the interesting region without dithering. By slightly scaling down the detected region into the target size, we solve the problem of big bounding boxes to the expected resolution while more saliency is preserved. In order to make a distinction between the detection results in [10] and ours, cropping markers of our method is changed from red into magenta. The same problem of shaking red bounding boxes in Fig. 4.13(b) for [10] does not exist in our results.

More results for video retargeting are shown in Fig. 4.14. By detecting the salient

Figure 4.13: Comparison (a) our results in red and results from [9] in green. Green bounding box is moving which leads to dithering while our result completely contains the salient objects without shaking. (b) our results in magenta and results from [10] in red. Red bounding box is shaking which can be seen from its distance changes to our result which is better for this fixed camera scene. Better to see in color.

Figure 4.14: Video retargeting results by cropping salient sub-images for nine video shots which are displayed in each row. For each video, the first columns are the final selected region by our method; the second columns are the results by directly scaling original video and the last columns are the results by first-padding-then-scaling.

sub-image with desired aspect ratio, our method with scaling has less visual distortion compared to the directly scaling method. By dropping the nonsalient regions, it preserves most of salient information for the interested area, such as the text in the car frame and the brand of the can in the drinking frame, compared to the results of the first-padding-then-scaling method in the third and sixth column. By globally cropping one salient sub-image out, the proposed method keeps the temporal coherence and never introduces any virtual camera motions. It works well for sport videos with complex camera motions which are very difficult to handle by other video retargeting methods. In the third column of Fig. 4.14, we can see unimportant regions such as logos *all star* in the top right of the Sports 2009 NBA video and *channel 7* in the top right of the 4x100m Relay video are removed by our method. The direct scaling and first-padding-then-scaling methods have slight difference for sports videos in that the aspect ratios of them are very close to the target $4/3$. We set C in the range as shown in Fig. 4.7 (d) and $\sigma^2 = 10^5$ for Eq. 4.10.

4.4.2.2 Salient Video Object Detection on Saliency Maps by Chapter 3

In order to show the effectiveness of the proposed video salient object detection method and the video saliency maps generated by Chapter 3, we pick the *Ten-Video-Clips* [82] dataset which has been used in Chapter 3 to conduct our experiments. This dataset contains 10 short video clips of 5 to 10 seconds each. Every video includes one major object which moves across the whole video and draws human attention. The ground truths are provided in the form of segmented object masks. For fair comparisons, the minimum bounding box which includes the segmented object is used for evaluation. Similar to Section 4.4.2.1, precision, recall and F-measure are calculated and used to obtain the quantitative results for our method. For a video clip, precision, recall and F-measure are calculated in each frame, and then averaged as the overall measurement

for the video. Since this dataset targets to evaluate the performance of single object detection, the global sub-image detection which is for video retargeting is not included here. Fig 4.15 shows the performance of our detection method for each video on the

Figure 4.15: Performance of precision, recall and F-measure by our video salient object detection method on Ten-Video-Clips dataset with Chapter 3’s saliency estimation method.

Ten-Video-Clips dataset. From these results we can see that our method gets good performance for most of the videos and the averaged precision, recall and F-measure for all videos is up to 82.62%, 71.14% and 78.89%, respectively.

4.5 Discussion

In this section, we first discuss the connections between saliency map and salient objects. Then, a further extension of the proposed method is presented for multiple salient object detection in images. Some limitations of our method are also discussed afterwards.

4.5.1 From Saliency Map to Salient Objects

Most bottom-up saliency estimation models aim to predict human visual attention. One representative work on computational visual attention is Itti’s model [2], in which saliency map is obtained by computing center-surround difference among multi-scale low-level visual features and the shift of attention to next attentional region is modeled by interplaying Winner-take-all and inhibition-of-return (IOR). It is believed that interesting objects on the visual field have specific visual properties that make them different from their surroundings. Therefore, salient objects can be detected in the saliency map.

Although salient objects can be detected via saliency analysis [16, 6], some previous object detection methods, on the other hand, can not rely on bottom-up saliency analysis and are usually based on supervised learning in order to make the detection robust. Such detection methods usually focus on a specific class of objects, e.g. face detection. Compared to specific object detection, in our object discovery, no explicit learning is required and we do not constrain the specific type of the objects, in order to perform general object detection. As will be explained below, our proposed salient object discovery method can detect multiple salient objects simultaneously, regardless of their categories.

4.5.2 Multiple Salient Objects Detection

Our proposed method can be easily extend to detect multiple objects in the same image. To implement multiple salient objects detection, we can iteratively detect salient object one by one from the saliency map, which is also biological plausible according to the spotlight theory [88] on visual search. The primary idea of the spotlight theory is to describe human attention as a movable spotlight which is directed towards intended targets and focusing on each target serially. Our strategy of multiple salient object

detection is shown in the Algorithm 3.

Algorithm 3: Multiple Salient Objects Detection Algorithm by MSD

Data: Image saliency map $S \subseteq R^{m \times n}$ and upper-bound function $\hat{f}(\mathbb{W})$ as Eq. 4.7

Result: $W_i^* = \arg \max_{W \subseteq S} f(W), i = 1 \cdots M$

begin

 Initialize $density \leftarrow \infty$

 Obtain θ using Eq. 4.12

for $i \leftarrow 1$ **to** M **do**

 Perform Algorithm 2 with S and $\hat{f}(\mathbb{W})$

if $D(W) \leq \theta$ **then**

$\perp$ break

 Calculate $D(W)$ using Eq. 4.13

$S \leftarrow \{S(x, y) = 0 | (x, y) \in W\}$

$W_i^* \leftarrow W$

Assuming there are maximum M salient objects in an image, each time we detect a salient object with the proposed MSD algorithm and update the saliency map by setting the detected salient object region into zero. The detection stops when the density of the salient object region is less than a predefined threshold θ which is set proportionally to the average density of the saliency map as:

$$\theta = C_2 \times \frac{\sum_{(x,y) \in I} S(x, y)}{Width(S) \times Height(S)}, \quad (4.12)$$

where C_2 is a positive constant for all testing images. The density of the detected salient object is

$$D(W) = \frac{\sum_{(x,y) \in W} S(x, y)}{Width(W) \times Height(W)}, \quad (4.13)$$

where W is the bounding box of the detected salient object.

Previous work [89] proposed that the selection of M is related to the resolution of image and $M = 3$ is usually enough for 320×240 . For the MSRA dataset with resized size 300×225 , we choose $M = 2$. Since the MSRA dataset is mainly for single salient object

detection, several images with multiple objects from [2] are selected and tested. For the balloon image, we set $M = 5$ and the results are shown in the first row of Fig. 4.16. The second row in Fig. 4.16 shows detection results of other testing images with $M = 2$. To improve the efficiency of multiple object detection, the advanced branch-and-bound search method for top-K objects [90] can be used in the future.

Figure 4.16: Multiple objects detection results on the balloon image from [2] and some natural images from [2, 5] with maximum salient object number $M = 5$ (first row) and $M = 2$ (second row) respectively. Better to see in color.

4.5.3 Limitations of the Proposed Method

Our salient object detection is based on the observation that the saliency density of the salient object region is always denser than other places. Therefore, the proposed method has the limitation in separating multiple salient objects that appear close to each other. The same issue also exists in [49], where only disjointed salient objects can be detected successfully while the detection fails to distinguish different objects if they are bound together. The main difference between our method and [49] is that our method is an unsupervised method which does not rely on the training data. Our future work includes the parameter learning and the localization of multiple salient objects combined with the semantic cues in images and videos.

4.6 Conclusion

After obtaining the saliency estimation results: saliency maps of images or videos, how to accurately locate the salient object in images and videos is not a trivial problem. A novel method is proposed to detect the salient object from saliency maps of an image or a video shot. The task of the salient object detection is formulated as an optimization problem to find a salient sub-image with maximum saliency density. The global optimal solution is efficiently obtained by the proposed branch-and-bound search algorithm. A constrained salient object region detection in terms of its size and aspect ratio is also proposed, which can be applied to image/video retargeting. Without a prior knowledge of the salient object, the proposed method can adapt to different sizes and shapes of the objects, and is less sensitive to the cluttered background. The global salient sub-image detected by our method in a video shot can preserve most of the interested regions from the video content with no visual distortion. Different types of saliency maps are also tested to validate the generality of the proposed method. The experiments on a public dataset of five thousand images have shown the promising results of our method on images. Extensive experimental results on real video shots demonstrated the efficiency and robustness of our algorithm for salient object detection in video sequences.

Chapter 5

Salient Region Detection via Optimal Path Discovery

5.1 Introduction

Our previous method in Chapter 4 locates the salient object by finding a bounding box of the maximum saliency density (MSD) based on the observation that the average density of the salient object region is much larger than that of any other regions on saliency maps. The major limitations of this method on salient video object detection are that the saliency density is the sole factor to detect objects and the temporal correlations between salient objects are ignored. Some examples can be seen in Fig. 5.1. A referee who is breaking into the view is wrongly detected as the salient object by the saliency alone. On the contrary, the football player and the flying ball are the concerns of audience. How to accurately detect salient objects based on saliency maps but adjusted by temporal coherence becomes an interesting problem.

Despite of many methods proposed to detect salient objects, few methods focus on

Figure 5.1: The main idea of our method. Salient objects in a video can be detected by discovering a salient path which has the maximum accumulated saliency density. By adding the temporal coherence constraints to the most dense regions, our method can accurately detect the salient objects. Better to see in color.

locating salient objects in videos. There are methods proposed to locate a specific object in videos, such as the slide window search [91] and the optimal path discovery (OPD) [92], but neither of them targets to find generic objects without clear category labels. Unlike a specific object detection and localization in which a discriminative score map can be obtained in advance for each object belonging to a category, salient objects only have a confidence map which shows how salient each pixel is. Therefore, it is more difficult to not only detect but also locate the salient objects. Moreover, instead of using a bounding box to locate objects, OPD targets to find a path with various sizes of windows as the nodes to indicate the existence of an object or an event. However, they did not consider the temporal correlations among each object regions.

To address the aforementioned problems, we propose a novel unsupervised method to discover salient objects from a given video clip. Our salient object location is formulated as an optimal salient path discovery problem. Suppose that any 2D sliding window in

a frame is a node and has a local detection score (i.e. saliency density) assigned by the estimated saliency map. By fusing these local evidences and connecting nodes, we build a spatio-temporal salient path. After adjusting the score by adding the temporal coherence constraints among nodes, the accumulated score for all nodes along the path shows the likelihood of the occurrences of salient objects. Thus, the path which has the maximum accumulated score corresponds to the target salient objects. Our salient object discovery method can maximize the likelihood of the occurrences of the salient objects and rectify the detection results by enhancing the temporal coherence among salient object regions. Regarding to the solution, a dynamic programming algorithm is proposed and a global optimal solution is obtained afterward. Different from OPD, our solution is not performed on a discriminative score map and the temporal correlations among nodes in the path are taken into consideration. Last but not least, in order to better characterize the role of motion in video saliency estimation, we propose our own way to estimate motion contrast. Extensive experiments on hundreds of videos from two public datasets demonstrate the effectiveness of the proposed method on salient video object detections.

5.2 Discriminative Map for Salient Video Object

5.2.1 Video Saliency Estimation

Since the saliency map estimation method proposed in Chapter 3 is time consuming and the simple combination method in Chapter 4 did not pay much attention to the motion cues, we use motion contrast instead of the magnitudes of motion vectors to represent motion saliency. We combine image saliency S_I and motion contrast S_M to obtain the

video saliency map M for a specific frame as Eq.5.1 shows:

$$M = N(S_M) + N(S_I), \quad (5.1)$$

where $N()$ is the min-max normalization operator, which normalizes the saliency map via finding the minimum and the maximum values of a frame individually. This equation is also plausible in that various saliency clues (i.e. the static image saliency and the motion contrast) are equally important. We adopt S_I from [69] as a static part of our video saliency estimation in that context is considered important to salient object detection in videos.

There is no standard method to represent the dynamic saliency. Here, our motion saliency is calculated as the motion contrast, which is obtained via removing the global motion from the estimated motion vectors. Motion estimation is performed by the pyramidal implementation of the Lucas-Kanade feature tracker algorithm for all frames [93]. The obtained motion vectors describe the velocity of a pixel point moving from frame t to the corresponding position in frame $t-1$.

Figure 5.2: Sample results of video saliency maps: original frames (the first row), context aware saliency map (the second row), motion saliency map (the third row) and our fused ones (the fourth row). Better to see in color.

Specifically, based on the obtained dense optical flow V [93], each motion vector

$v \in V$ is projected into x-axis and y-axis as v_x and v_y . Denote the motion vector sets after projection by V_x ($v_x \in V_x$) and V_y ($v_y \in V_y$). We can obtain the direction of the global motion along x-axis, which is decided by the major direction and should be either x-axis positive or x-axis negative, as shown in Eq. 5.2.

$$\begin{aligned} pV_x &= \{v_x | v_x \in V_x \text{ and } v_x > 0\} \cup \{0 | v_x \in V_x \text{ and } v_x \leq 0\}, \\ nV_x &= \{v_x | v_x \in V_x \text{ and } v_x < 0\} \cup \{0 | v_x \in V_x \text{ and } v_x \geq 0\}, \\ G_x &= \begin{cases} pV_x & \text{if } |pV_x| > |nV_x|, \\ nV_x & \text{otherwise,} \end{cases} \end{aligned} \quad (5.2)$$

where $|\cdot|$ is to calculate the number of nonzero elements in a set. Similarly, we can obtain the global motion affected on y-axis. After removing the global motion from the original motion vector set at both axes, the local motion of objects can be obtained and the motion saliency can be written as:

$$S_M = \sqrt{(V_x - G_x)^2 + (V_y - G_y)^2}. \quad (5.3)$$

The sampled results of our video saliency maps are shown in Fig. 5.2. As can be seen, our results (the fourth row) well balance the static saliency (the second row) and the motion saliency (the third row), when neither of them could get better saliency estimation results individually.

5.2.2 What's Salient Object

After obtaining a saliency map M , we know that each value of a pixel on the map shows the likelihood of the pixel belonging to be salient. For our salient object discovery task, a sliding window can be used to search every spatial position u ($u=(x,y)$) in a frame or a spatial-temporal position (u,t) in a video. Denote the window by $W(u,t)$ which is centered at (u,t) . Suppose the area of the window and the summed saliency within the

window are $A(W)$ and $S(W) = \sum_{(x,y) \in W} M(x,y)$, respectively. The saliency density of $W(u,t)$ can be calculated as:

$$d(u,t) = \frac{S(W)}{A(W)}. \quad (5.4)$$

We assume that the salient object should correspond to a location of large saliency density $d(u,t)$ and the window which has largest saliency density can be the candidate of the salient object.

However, a windows W of fixed size cannot accurately detect salient object with varying sizes. By gradually changing the scale and the aspect ratio of the window, the salient object can be cropped by the adjustment of the window scales. Given a scale level n_s , an aspect ratio level n_r and the initialized height h_0 , the height and the width of $W(u,t)$ can be decided:

$$\begin{aligned} \text{height} &= h_0 - \nabla h \times n_s, \\ \text{width} &= \text{height} \times (r_0 - \nabla r \times n_r), \end{aligned}$$

where ∇h and ∇r are the steps to change the scale and the aspect ratio. The numbers of the scales and the aspect ratios are N_s ($n_s \in [1, N_s]$) and N_r ($n_r \in [1, N_r]$). We start the detection at a large scale and decrease it gradually to avoid the detection stopping at a small window with large density. The parameters such as the number of scales (N_s and N_r), the steps to change the height ∇h and the aspect ratio ∇r are empirically set during the experiments.

5.3 Salient Path Discovery

Given a sequence of saliency maps $M = \{M_1, \dots, M_T\}$ where T is the number of frames in a video, our goal is to detect and locate salient objects by selecting a salient spatio-temporal path that can track the salient object in the video. By connecting suitable

candidate windows in a video, a spatio-temporal path denoted by p can be obtained and the accumulative score of all nodes along the path shows the evidence of the occurrences of salient object.

More formally, the salient video object detection problem, becomes to find the optimal path p^* with the maximum accumulated saliency density:

$$p^* = \operatorname{argmax}_{p \in \operatorname{path}(\mathcal{G})}(D(p)), \quad (5.5)$$

where $D(p) = \sum_{u,t} d(u,t)$ is the summed saliency density along a path p which starts from the start location (x_s, y_s, t_s) to the end location (x_e, y_e, t_e) and $\operatorname{path}(\mathcal{G})$ is a set of all possible paths, where each path is a smooth spatio-temporal trajectory of arbitrary length.

From Eq. 5.5, we know that salient objects appearing consecutively in a video can be found by maximizing the accumulative saliency density. However, each salient object is decided by its saliency only and the temporal coherence as one of the most important characteristics of salient objects in videos is ignored. The way to obtain the correlations between two salient objects in consecutive frames is introduced in the following sections.

5.3.1 Temporal Coherence

To accurately detect salient objects one by one in a video, the confidence of an individual frame is not enough. We know that the coherence between two salient objects could make the detection more robust. In order to represent the temporal coherence between two salient objects in consecutive frames, we employed the optical flow estimation for video saliency estimation in Section 5.2.1. With the estimated motion vectors, each frame has a pixel point mapping from current frame to the next frame by the obtained motion vectors. Based on the obtained motion vectors, each pixel in the current frame t

can be mapped to a new position in frame $t-1$. We can use the number of matched pixels between two windows as the similarity measurement.

To simplify the description, in this part, we only use the center location to represent a window, such as u , v , u_j , u_k and v_i . All of them identify different windows under a specific scale and a fixed aspect ratio. For a window u in frame t , we target to find the window which has the strongest similarity among its 9-neighbors (i.e. $v_1, \dots, v_9$) in frame $t-1$. Fig. 5.3 illustrates the 9-neighbors of a window temporally and the mapping between two consecutive frames by motion vectors. The similarity between u and v_i ($i = 1, \dots, 9$) is measured by the number of pixels which are from u but located in v_i after the mapping. In other words, each pixel in u has a mapping in frame $t-1$ but the location should be in one of the 9-neighbors of u . Therefore, the neighbor which has the majority of the pixels in u is the most similar window to u . For simplification, a searching window is also called a node and the similarity between two nodes is called an edge weight in the following paragraphs.

Figure 5.3: The illustration of a window u in frame t and its 9-neighbors in frame $t-1$, and the pixel mapping by motion vectors is also shown.

Suppose the total number of pixels in u is N , and the number of pixels from u and mapped into v_i is N_i . We can measure the similarity between u and v_i as:

$$w_{u,t}(v_i, t-1) = \frac{N_i}{N}, \quad (5.6)$$

where $\sum_{i=1}^9 N_i = N$. The large value of $w_{u,t}(v_i, t-1)$ means that more pixels in v_i are from u . The value of $w_{u,t}(v_i, t-1)$ is non-negative and within $[0, 1]$. Since both the optical flow estimation and the similarity calculation are backwards, for the first frame, we set $w_{u,1} = 0, \forall u$.

5.3.2 Fusing Temporal Coherence to Saliency Density

By considering the temporal consistency of an object, we make the discovery algorithm more robust to handle the occlusions. However, we know that each searching window is represented by the saliency density, which is not a ratio. How to well incorporate the saliency density of a node and the edge weight is another issue. We show that a better performance can be achieved by multiplying them together. That is:

$$d'(u, t) = d(u, t) \times w_{u,t}(v, t-1), \quad v \in N(u). \quad (5.7)$$

We provide some analysis to show the feasibility of our fusion strategy. Assume we are in a more general circumstance: the saliency map becomes a classification score map in which the value could be positive and negative. The classification score within a window u_k is d_k and another window u_j has the score d_j . For each of u_k and u_j , we can find the most similar window in previous frame and obtain the edge weight as w_k and w_j . Without loss of generality, we set $S_k > d_j$ and $0 \leq w_k < w_j \leq 1$. In order to incorporate the temporal consistency into the classification score, we consider three cases: **1.** $d_k > 0 \geq d_j$; **2.** $0 \geq d_k > d_j$ and **3.** $d_k > d_j > 0$.

1. When $d_k > 0 \geq d_j$, it means that u_j does not belong to the specified class while u_k does. For this case, we do not want that our new method by considering the temporal consistency makes the object which does not belong to the specified class be detected. Based on the previous assumption, we have $d_k \times w_k > 0 \geq d_j \times w_k$. Moreover, based on $0 < w_k < w_j$, we can obtain $d_j \times w_k > d_j \times w_j$. Therefore, $d_k \times w_k > 0 > d_j \times w_j$, and we did not change the decision to select the region u_k as the candidate.

2. When $0 \geq d_k > d_j$, it means that both the regions does not belong to the specified class. Based on the previous assumption, we have $0 \geq d_k \times w_k > d_j \times w_k$ and $d_j \times w_k > d_j \times w_j$. Therefore, $0 \geq d_k \times w_k > d_j \times w_j$. The candidate region still lies to u_k .

3. When $d_k > d_j > 0$, it means that u_k is more confident than u_j to be the specified class. However, the large value of d_k does not grantee the true category of object detected. We prefer u_k if it satisfies $d_k \times w_k > d_j \times w_j > 0$. That's $\frac{d_k}{d_j} > \frac{w_j}{w_k} > 1$. Otherwise, we change the candidate region to u_j when $d_j \times w_j > d_k \times w_k > 0$. That's $\frac{w_j}{w_k} > \frac{d_k}{d_j} > 1$.

To summary, our temporal coherence only influences the decision when the candidate regions are potentially belong to the target class, and our salient object detection problem fits to the third case when the node score can only be a positive value. In other words, our fusion strategy and detection scheme can be extended for other applications, such as action recognition, event detection, etc.

Back to our case, since $\frac{w_j}{w_k} \in [0, +\infty)$, the changes to the saliency density could be huge (e.g. $d_j \times \frac{w_j}{w_k} \gg d_j$), we use the exponential function to modulate the value of w_k into $[1, e]$, such that $\frac{w_j}{w_k}$ ranges from $\frac{1}{e}$ to e and Eq. 5.6 can be rewrote as:

$$w_{u,t}(v_i, t-1) = \exp\left(\frac{N_i}{N}\right). \quad (5.8)$$

It is worth mentioning that the performance varies slightly when the base of the

modulation function changes to other values, such as 2, 10.

By modulating the weight, we can avoid the problem caused by multiplying 0 to making the saliency density meaningless. Moreover, by taking the multiplication, especially for classification problems with positive and negative node values, the prediction of a node cannot be changed but the decision has been adjusted.

5.3.3 The Proposed Method

After solving the fusion problem for the edge weight and the node confidence, we do a spatial-temporal search to find an optimal path for salient object in a video. The main idea of the proposed method can be described as in Algorithm 4.

Algorithm 4: Optimal Salient Path Discovery Algorithm

Data: Discriminative score maps and the estimated optical flow; **Initialization:**

$$w_{u,1} = 1.0, \forall u;$$

Result: $P(u, t)$: the best path record for tracing back;

$A(u, t)$: the accumulated scores of $P(u, t)$ leads to (u, t) ;

l^* : the ending location of the best path;

A^* : the accumulated scores of the best path;

begin

Set $A(u, 1) = d(u, 1), \forall u$; $p = null, \forall (u, t)$;

$A^* = -\infty$; $l^* = null$;

for $t \leftarrow 2$ **to** T **do**

for $u \in [1 \cdots w] \times [1 \cdots h]$ **do**

Calculate $w_{u,t}$ as Eq.5.8 did;

$v' \leftarrow \arg \max_{v \in N(u)} A(v, t-1) + d(u, t) \times w_{u,t}(v, t-1)$;

$A(u, t) \leftarrow A(v', t-1) + d(u, t) \times w_{u,t}(v', t-1)$;

$P(u, t) \leftarrow (v', t-1)$;

if $A(u, t) \geq A^*$ **then**

$A^* \leftarrow A(u, t)$;

$l^* \leftarrow (u, t)$;

The whole searching procedure starts from the first frame, and the information is

propagated from the current frame to the next one. At first frame $t = 1$, the position and the shape of the salient object are decided by the window which has the largest saliency density at a certain place. As the procedure comes to the next frame $t = 2$, at every point (u, t) , we search by a window $W(u, t)$, which has fixed size and shape decided by previously assigned scale level and aspect ratio level. Meanwhile, the similarities of $W(u, t)$ to its nine neighbors in frame $t-1$ can be obtained. The possible positions for the salient object moving from previous frame to the current position are among its nine temporal neighbors. In other words, the confidence of the salient object trajectory from $(v, t-1)$ to (u, t) is decided by $A(v, t-1) + d(u, t) \times w_{u,t}(v, t-1)$. When the neighbor which has the strongest relationship to the current window has been selected, the accumulative score $A(u, t)$ of the path is updated accordingly. Among the newly updated $A(u, t)$ in frame t , the largest one is labeled in $P(u, t)$ as the detected salient object in the current frame, and the search continues until the end of frames.

Regarding to the computational complexity, as can be seen in Alg. 4, the extra computation caused on node similarity calculation is linear to the cost of the whole searching process. Therefore, we have $O(WHTN_sN_r)$ for the whole algorithm, where W , H are the width and the height of a frame, and T is the number of frames in a video. It is worth mentioning that our method is able to detect multiple salient video objects. We can perform our method on the revised saliency maps by wiping the previously detected object regions out and detect the multiple salient video objects one by one.

In order to show that a path discovered by the Algorithm 4 is a global optimal solution, we present a proof in the following paragraphs.

Define $Q(t) \triangleq "A(u, t)"$ as the maximum accumulated score for the best path leading to the spatial-temporal position $(u, t)"$. We will prove that $Q(t)$ is the global optimal solution $\forall t \in [1, T]$. We initialize $A(u, 1) = d(u, 1), \forall u$, hence $Q(1)$ is true. Assume

that $\mathbb{Q}(t-1)$ is true, we show that $\mathbb{Q}(t)$ is also true. If a node u at frame t has m directly connected neighbors, then there are m possible paths leading to it. These paths include m paths going through its neighbors with an accumulated scores of $A(v, t-1) + w_{u,t}(v, t-1) \times d(u, t)$, $\forall v \in N(u)$.

Given a position (u, t) and a searching window $W(u, t)$, its saliency density $d(u, t)$ is fixed as $C \in [0, 1]$. We show that by maximizing $A(u, t)$ to select the best path leading to (u, t) is equal to maximize $\mathbb{P}(u, t)$, where $\mathbb{P}(u, t) \triangleq \max_{v \in N(u)} \{A(v, t-1) + w_{u,t}(v, t-1) + d(u, t)\}$:

$$\begin{aligned}
& \max_{v \in N(u)} \{A(v, t-1) + w_{u,t}(v, t-1) \times d(u, t)\} \\
& \Leftrightarrow \max_{v \in N(u)} \{A(v, t-1) + w_{u,t}(v, t-1) \times C\} \\
& \Leftrightarrow \max_{v \in N(u)} \{A(v, t-1) + w_{u,t}(v, t-1)\} \\
& \Leftrightarrow \max_{v \in N(u)} \{A(v, t-1) + w_{u,t}(v, t-1)\} + C \\
& \Leftrightarrow \max_{v \in N(u)} \{A(v, t-1) + w_{u,t}(v, t-1) + d(u, t)\} \\
& = \mathbb{P}(u, t)
\end{aligned} \tag{5.9}$$

It is worth mentioning that we use $\mathbb{Q}(t)$ instead of $\mathbb{P}(u, t)$ because it is difficult to balance the values of $w_{u,t}(v, t-1)$ and $d(u, t)$ when the summation is performed. However, we can borrow $\mathbb{P}(u, t)$ for our proof. It is easy to show that the best path across $(v, t-1)$ to (u, t) satisfies the following equation and the inequalities:

$$\begin{aligned}
v_0 &= \arg \max_{v \in N(u)} \{A(v, t-1) + w_{u,t}(v, t-1) + d(u, t)\} \\
&\Rightarrow A(v_0, t-1) + w_{u,t}(v_0, t-1) + d(u, t) \\
&\geq A(v, t-1) + w_{u,t}(v, t-1) + d(u, t), \forall v \in N(u)
\end{aligned} \tag{5.10}$$

Hence, the best v , selected from u 's 9-neighbors can make $\mathbb{P}(u, t)$ to be the maximum.

According to the previous derivation, we have:

$$\begin{aligned}
\mathbb{P}(u, t) &= A(v_0, t-1) + w_{u,t}(v_0, t-1) + d(u, t) \\
&\Leftrightarrow A(v_0, t-1) + w_{u,t}(v_0, t-1) \times d(u, t) \\
&= \max_{v \in N(u)} \{A(v, t-1) + w_{u,t}(v, t-1) \times d(u, t)\} \\
&\geq A(v, t-1) + w_{u,t}(v, t-1) \times d(u, t), \quad \forall v \in N(u)
\end{aligned} \tag{5.11}$$

Therefore, we can update $A(u, t)$ as:

$$A(u, t) = A(v_0, t-1) + w_{u,t}(v_0, t-1) \times d(u, t). \tag{5.12}$$

From the inequality 5.11 and Eq. 5.12, we have shown that $A(u, t)$ is always the best accumulated score compared to m paths that can lead to (u, t) . This confirms that $\mathbb{Q}(t)$ is true.

5.4 Experiments

In this Section, we first evaluate the proposed method by comparing it with three state-of-the-art methods on two public datasets on the

5.4.1 Experimental Setting

Datasets: We conduct experiments on two public datasets: *UCF-Sport* [94] and *Ten-Video-Clips* [82]. UCF-Sport dataset consists of 150 video sequences of 10 different action classes. Since, almost every video clip features one action performed by one or several people, the participants who draw our interests across the whole video can be discovered by our method. Meanwhile, provided the ground truth of each object with human labeled bounding box, the averaged precision, recall and F-measure of our method could be

evaluated. Ten-Video-Clips dataset contains 10 short video clips of 5 to 10 seconds each. Every video clip focuses on one major object in the natural scene. The ground truths are provided in the form of ten sets of segmented object masks. For fair comparisons, the minimum bounding box which includes the segmented object is used for evaluation.

Evaluation Metrics: Given the ground truth S_g which is the mask of the salient object and the binary detected salient object region S_d , we evaluate the performance of our method based on precision, recall and F-measure:

$$\begin{aligned} \text{precision} &= \frac{\|S_g \cap S_d\|}{\|S_d\|}, \\ \text{recall} &= \frac{\|S_g \cap S_d\|}{\|S_g\|}, \\ \text{F-measure} &= \frac{(1 + \alpha) \times \text{precision} \times \text{recall}}{\alpha \times \text{precision} + \text{recall}}. \end{aligned} \quad (5.13)$$

We set $\alpha = 0.3$ as suggested in [64]. For a video clip, precision, recall and F-measure are calculated in each frame, and then summed up and averaged as the overall measurement for the video.

5.4.2 Experimental Results

In order to demonstrate the superiority of the proposed method on video salient object detection, we compare our method with OPD [92] and MSD [11], which are two object detection methods. Due to [64] is a supervised method and there is no quantitative results provided in [82], we do not compare it. Moreover, to distinguish our method to tracking based method, performance comparison also were performed between our method to [12]. Experiments are also performed to show that the proposed method can adapt to various video saliency map estimation methods. For simplicity, the saliency map estimation method proposed in Chapter 3 is used, and the video salient object detection results are compared with the performance by the same method with the newly proposed

video saliency estimation method used in this Chapter.

Figure 5.4: Sample results by our method (third row) and MSD [11] (last row) on the 'Riding Horse' videos from UCF-sports dataset. The first and the second rows show the original frames and their corresponding video saliency map. The blue mask indicates the detected results while the orange ones are the ground truth. Without considering the motion coherence, MSD cannot consistently detect the salient object based on the saliency density alone). Better to see in color.

5.4.2.1 Comparisons with MSD

Comparison our method with MSD is to show the superiority of the proposed method on salient video object detection and the advantages to take temporal coherence among salient objects into consideration. For fair comparisons, we use the parameters which get the best performance as [11] did. Sampled detection results on the 'Riding Horse' videos from UCF-sports dataset are shown in Fig. 5.4. Since MSD locates the salient object for each frame and the decision is made only based on the saliency density in the current frame, when the salient object does not have the largest saliency density, some wrongly detection results appear. On contrary, with the help of temporal consistency, our method gets more robust detection results. From the objective results on the Ten-Video-Clips dataset in Fig.5.5, we can see that the performance of our method outperforms MSD

method.

Figure 5.5: Precision, recall and F-measure comparisons our method with MSD and OPD on Ten-Video-Clips dataset.

Figure 5.6: Sample results by our method (second row) and OPD (last row) on the 'Kicking' videos from UCF-sports dataset. The blue mask indicates the detected results while the orange ones are the ground truth. Without considering the motion coherence, OPD cannot detect the small salient objects. Better to see in color.

5.4.2.2 Comparisons with OPD

The experiments between our method and OPD target to show that the superiority to take temporal coherence among salient objects into consideration. Since OPD is only for discriminative maps, we first find a threshold for each saliency map by Otsu [84], and

then OPD is performed on the saliency map after subtracting the threshold. The same parameters are used in both our method and OPD.

Table 5.1: Averaged F-measure (%) $\pm$ Standard Deviation for ten types of action videos in UCF-sports dataset.

	Ride	Run	Kick	SwingSide	lift
Ours	71.04 $\pm$ 0.10	61.12 $\pm$ 0.28	64.11 $\pm$ 0.22	54.47 $\pm$ 0.18	88.48 $\pm$ 0.02
OPD[92]	68.77 $\pm$ 0.23	55.57 $\pm$ 0.10	64.10 $\pm$ 0.23	37.29 $\pm$ 0.15	87.63 $\pm$ 0.01
MSD[11]	56.52 $\pm$ 0.20	53.55 $\pm$ 0.32	60.02 $\pm$ 25.83	34.13 $\pm$ 0.10	83.86 $\pm$ 0.02
	Skate	Diving	Golf	SwingBench	Walk
Ours	46.33 $\pm$ 0.35	69.76 $\pm$ 0.13	62.88 $\pm$ 0.26	59.06 $\pm$ 0.18	54.17 $\pm$ 0.26
OPD[92]	42.41 $\pm$ 0.34	68.62 $\pm$ 0.10	56.32 $\pm$ 0.26	58.98 $\pm$ 0.20	50.67 $\pm$ 0.22
MSD[11]	40.69 $\pm$ 0.34	61.50 $\pm$ 0.13	52.22 $\pm$ 0.23	58.62 $\pm$ 0.19	45.74 $\pm$ 0.20

Sampled detection results on two video clips from the employed datasets are shown in Fig. 5.6. Based on the same saliency map and the same parameters for the initial scale and the aspect ratio, OPD ignores the small salient objects compared to our results in the third row. It is probably because each saliency map is thresholded independently and no temporal relationships among detected regions are considered.

To quantitatively compare MSD and OPD with our method on UCF-sports dataset, we present the averaged F-measure in Table 5.1. The result comparisons are performed and categorized based on the video type on action. For simplicity, results from 'kick side' and 'kick front' are summed into one category 'kick'. Similarly, 'golf swing side', 'golf swing front' and 'golf swing back' are combined as 'Golf'. Our method averagely improves the F-measure by 1% to 6% compared to OPD and MSD for all action categories. It also demonstrates that the proposed method can handle various categories of salient object detection. The salient object could be a person riding horse, a golf player swing his/her golf club, a runner, a skater, etc. Without any prior knowledge, any object performing a specific action and drawing our interests across the video can be accurately detected.

In order to show the benefit from adding the temporal coherence into detection, we improve OPD method by fusing the correlations between pairwise searching windows into judgement. In other word, the confidence of a searching window containing a salient object depends on not only the positive-negative classification score but also the temporal coherence of it to its nine neighbors in previous frame. The performance comparisons has been shown in Fig. 5.7 for our method, improved OPD method with motion coherence and the original OPD without considering the motion coherence. From Fig. 5.7, the advantages of the proposed method and the improved OPD with the temporal coherence can be seen clearly.

Figure 5.7: Precision, recall and F-measure comparisons among our method, OPD with motion coherence and the original OPD without considering motion coherence on UCF-sports dataset.

5.4.2.3 Salient Object Detection vs. Tracking Method [12]

To further discriminate our method from the object tracking method, we compare our detection results by the tracking method [12] on the ‘Swing-Side Angle’ videos from UCF-

Figure 5.8: Sample results by our method (first row) and the tracking method (last row) on the ‘Swing-Side Angle’ videos from UCF-sports dataset. The blue mask indicates the detected results while the orange ones are the ground truth. Since method in [12] tracks an object by a fixed size bounding box, it fails to detect the object with strong motions and big shape changes). Better to see in color.

Figure 5.9: F-measure comparisons between our method and tracking method [12] on the ‘Swing-Side Angle’ videos from UCF-sports dataset.

sports video dataset. Fig. 5.8 shows the sample results obtained by our method and the method in [12] for comparison. We adjust the parameters for achieve better tracking results based on the codes provided by the author. In order to obtain the same result representation as the detection, we fill the tracking bounding box with different colors for better view. From this figure, we can see that the tracking method can accurately locate the object at the first several frames with the manually assigned bounding box. However, when the object moves swiftly or has big shape changes, the tracking method failed to track them. On the contrary, our method achieved robust detection results. We also present the quantitative comparisons between our method and the tracking method [12] in Fig. 5.9. It further demonstrates the superiority of our method on salient object detection on F-measure from 42.95% to 54.47%.

5.4.2.4 Adaption to Various Saliency Maps

Similar to our previously proposed salient object detection method in Chapter 4, our optimal path discovery based salient object detection can adapt to various saliency maps as well. To show the superiority of the method from this point of view, we compare the performance by saliency map generation methods in this Chapter and the one in Chapter 3.

From figure 5.10, we can see that comparative results are obtained from both kinds of saliency maps on Ten-Video-Clips dataset. However, regarding to the time complexity, the saliency map estimation method proposed in this chapter is more effective compared to the one proposed in Chapter 3. Therefore, we use the simple saliency estimation method proposed in this chapter for all our experiments.

Figure 5.10: Precision, recall and F-measure comparisons our method with saliency maps from Chapter 3 and Chapter 5 on Ten-Video-Clips dataset.

5.4.2.5 Multiple Scale vs. Single Scale

In order to display the ability of our method on handling deformable object detection, we perform the comparisons between the results by using multiple and single aspect ratio and scale respectively for our method on both two datasets. Sample results on the ‘Swing-Side Angle’ videos from UCF-sports video dataset and on the ‘DO01-055’ video from Ten-Video-Clips are also shown in Fig. 5.12. From this figure we can see that the salient objects in both selected videos have a high level of articulated body movements, and our method with multiple scales and aspect ratios better capture them compared to the single scale method. To further demonstrate the advantages of our method on deformable salient object detection, we also present the F-measure curves for each dataset, and the performance improvement by using multiple scale windows for detection is clearly displayed.

Figure 5.11: Performance comparisons for our method by a fixed size window (single scale) and windows with multiple scales on the two employed video datasets.

5.4.2.6 Multiple Salient Objects Detection

We also show that our method is able to detect multiple salient objects. We can perform our method on the revised saliency maps by wiping the previously detected object regions out, then the multiple salient video objects can be detected by the proposed algorithm one by one. Since each video in the employed two datasets only focuses on one action, only several of videos contain multiple salient objects. We sample some multiple salient object detection results in Fig. 5.13. As can be seen, our method can correctly detect all salient objects in both 'Kicking' and 'Riding Horse' videos.

5.4.2.7 Computational Complexity

All our experiments are implemented by C++ on Dell Server T710 Intel(R) Xeon(R). Each video is searched by a window with averaged 3 scales and 3 aspect ratios. Table 5.2 shows the averaged time cost to process one video for our method, OPD and MSD, respectively. Due to the calculation of the motion coherence in each scale, our method is slower than MSD and OPD. However, we achieve better detection accuracy than both of them.

Figure 5.12: Performance comparisons for our method by a fixed size window (second row) and windows with multiple scales (third row) on swing action videos from UCF-sport dataset and DO01-055 from Ten-Video-Clips dataset. The blue mask indicates the detection results for our method while the orange ones are the ground truth. Better to see in color.

Figure 5.13: Multiple salient object detection results by our method on 'Kicking' and 'Riding Horse' videos from UCF-sports. The blue mask indicates the detected results while the orange ones are the ground truth. Better to see in color.

Table 5.2: Average time cost to process each video for ten types of action videos in UCF-sports dataset.

	Ours	MSD	OPD
Time Cost (Sec.)	145.211	1.121	1.326

5.5 Conclusion

As a significant improvement of the video salient object detection method in chapter 4, in this chapter, we first formulate video salient object detection as an optimal salient path discovery problem and propose a dynamic programming algorithm to solve it efficiently. By adding the constraint of temporal coherence in salient path discovery, our method obtains superior performance compared to two state-of-the-art methods. Moreover, without any prior knowledge of the salient objects, our method can automatically detect various salient objects across different scales and aspect ratios. Extensive experiments on two public datasets demonstrate the effectiveness of our approach.

Chapter 6

Thematic Object Discovery via Subgraph Mining

6.1 Introduction

In Chapter 3, Chapter 4 and Chapter 5, we have discussed methods to estimate video saliency and to locate the salient objects in images or videos. Apart from salient objects in videos, there is another category of objects which attract human attentions and represent the theme of a video content. We name it as the thematic objects. It is not a trivial problem to find thematic objects in videos. First of all, the same thematic object may vary significantly in different video frames due to pose and scale variations, such as the bride and the groom in Fig. 6.1. It thus brings extra challenges as we have no prior knowledge of the thematic objects. Moreover, for videos captured with moving cameras, the motion cue is of limited help to segment the moving object. Thus we have to rely on other cues to detect the thematic object. Last but not least, unlike salient object detection in images or videos, not every frame contains the thematic object. For example, a wedding

Figure 6.1: Sample results of thematic video saliency discovery. The first row shows the original key-frames with the gaze points overlapped. Each green cross '+' represents one gaze point. The second row shows our detection results while the last row shows the results of co-saliency method [13]. The discovered video thematic salient region is rendered in red.

car could be the salient object in a frame or a specific video shot, but it may not be the thematic object given the whole wedding video. Thus, it requires the help of the video contexts to extract the thematic object. Despite these challenges, the thematic object must be the common object that is frequently highlighted in various video scenes and should be visually salient. As a result, we can leverage the characteristics of thematic objects to find them.

We propose to discover thematic video object by taking the whole spatio-temporal video context into consideration. Given a video clip, we first select the key frames via sampling every 10 frames, and these key frames are decomposed into image segments.

After selecting salient image segments, the similarity relationships among all potential segments are presented in an affinity graph. Then object discovery is formulated as sub-graph mining problem on the affinity graph of image segments. By discovering the cohesive sub-graph, i.e., a group of cohesive image segments of high saliency and similarity, we can discover the image segments associated with thematic object, followed by the localization of the thematic object. Fig. 6.1 shows our results and the comparison with the eye tracking result, as well as the co-segmentation detection method in [13] which only emphasizes the common parts between consecutive frames.

To evaluate our object discovery result, we create a video dataset with the eye tracking results as the ground truths. We also test our method on three public video datasets for video object detection. Our proposed method show promising results when compared with other image and video saliency based object detection methods. Moreover, we also compare our cohesive sub-graph mining methods with existing sub-graph mining methods and it shows that our method can find better optimization result without comprising the computational cost.

6.2 The Proposed Method

The spatio-temporal occurrences of objects are very important information provided in videos. Usually, dominant objects across the whole video always draw our attention and the instances of these objects typically share similar appearance. Inspired by this, we propose a new video saliency detection method which considers the spatial-temporal context. The key steps of our method are shown in Fig. 6.2. We first sample key-frames from each video and each key-frame is segmented multiple times to obtain the segments. Candidate segments are selected by image saliency and then described by the appearance

features. After that the pairwise relations between selected segments are characterized by an affinity graph. By cohesive sub-graph mining algorithm, segments which have strong pair-wise affinity relationships are obtained as the thematic video saliency.

Figure 6.2: The main steps of our method has been illustrated. Details can be found from the text in Sec. 6.2.

6.2.1 Salient Segment Selection

To obtain the segments, we perform a superpixel segmentation method [95] per key-frame, with the expectation that some segments could correspond to object regions while some may fail to agree with object boundaries. Several types of bag-of-features histograms are used to describe each segment and the details can be found in Sec. 6.3.2.

In order to speed up the processing and reduce the size of segment set, candidate segments are first selected from multiple segmentation results. Image saliency and motion saliency, which is stimulus driven, aim to identify important positions in images or videos.

Figure 6.3: Thematic video object discovery result comparisons with bottom-up saliency filtering (the third row) and without saliency filtering (the second row). Better to see in color.

However, they also partially coincide to our long term interest (e.g., the bride in the wedding video). We employ them as a weighting prior to filter out some unimportant segments (e.g., background segments). For each segment, its weight can be calculated as the average saliency fell into this segment. Here the saliency is the linear combination of spatial saliency S_I [6], motion saliency S_M estimated by pixel-wise optical flow, and face mask S_{face} [96]. Each saliency map is first normalized into $[0, 1]$. Then the overall saliency is obtained as $S_{map} = w_1 * S_I + w_2 * S_M + w_3 * S_{face}$, where all the weights are experimentally set as $w_1 = 0.3, w_2 = 0.2, w_3 = 0.5$ in our experiments. We add face detector to emphasize human being's particular interests to human face. Similar to [6, 13], a segment is kept as a salient object region if its average saliency is larger than its saliency density. Fig. 6.3 shows the benefits of saliency filtering for segment selection. After the segment selection, more thematic objects rather than the background have been correctly detected.

It is worthy to note that it is possible to use other types of image/motion saliency

fusion schemes [73, 49]. Moreover, even if some segments of the thematic objects are missed by saliency weighting, our method can still discover them considering their frequent occurrences in the whole video. After the selection, all the remain segments are noted as Π .

6.2.2 Segment Affinity Graph Model

In our scheme, we first estimate the pairwise segment affinity by considering the spatio-temporal context information and then build the affinity graph to represent the affinity relationships of all the segments.

6.2.2.1 Pairwise Affinity Estimation

Before introducing the affinity estimation, we explain the support segment set firstly. For a specific segment S_i in a key-frame I_m , we select segments which have similar appearance as S_i from all other key-frames. All these selected segments are the support segment set of S_i which is defined as $\{\Psi_{mi}\}$, where $\Psi_{mi} = \{S_{mi}\}$ represents all support segment of S_i in key-frame I_m , and S_{mi} is one support segment in key-frame I_m . To reduce the size of the support segment set, we only select one support segment which has a smaller distance with S_i . In other words, the size of Ψ_{mi} is set to be 1. The support segment set formulation integrates not only the spatial information of the salient object, but also the temporal trajectory information. Therefore, by comparing the support segment sets of two segments, we can obtain the affinity relationship of these two segments. If two support sets have an intersection of large size, then the two segments have high affinity relationship. Otherwise, these two segments have week affinity relationship. Based on

the Jaccard similarity coefficient, the affinity value of two segments is defined as:

$$A_{i,j} = \begin{cases} \frac{|\{\Psi_{mi}\} \cap \{\Psi_{mj}\}|}{|\{\Psi_{mi}\} \cup \{\Psi_{mj}\}|} & \text{if } |\{\Psi_{mi}\} \cap \{\Psi_{mj}\}| > 0 \\ \tau & \text{else} \end{cases}, \quad (6.1)$$

where τ is a negative value and $|\cdot|$ represents the cardinality of one set. If S_i and S_j have strong affinity, the value of $A_{i,j}$ is positive and vice versa. If the support segment sets of S_i and S_j do not have any intersection, $A_{i,j}$ is set to be a negative constant τ .

6.2.2.2 Building the Cohesive Sub-graph

Obviously, the segments belonging to the same thematic object share the similar appearance and have strong mutual affinity relationship. Moreover, they tend to have weak affinity relationship with the segments from the background or other objects. Therefore, the thematic saliency could be obtained by the cohesive sub-graph and this sub-graph can be represented by its vertices set Ω . Each element in Ω is the segment share the same thematic saliency. In other words, the thematic saliency can be represented as the spatio-temporal collocated segment group $\Omega \subseteq \Pi$, where all segments $S_i \in \Omega$ share the same thematic saliency. We define the affinity potential function for the sub-graph as Ω as $f(\Omega) = \sum_{S_i, S_j \in \Omega} A_{i,j}$ and find the cohesive sub-graph as follows:

$$\Omega^* = \arg \max_{\Omega \subseteq \Pi} f(\Omega). \quad (6.2)$$

In other words, the sub-graph with the largest affinity score is the maximum cohesive sub-graph and thematic segments as a sub-graph can be discovered sequentially with erasing the ones belonging to the previously found cohesive sub-graph.

Suppose the affinity matrix for all segments is A . We can convert the subset optimization problem in Eq. 6.2 to a binary optimization problem. Given a sub-graph Ω , let $\mathbf{x} = \{x_i\}_{i=1}^N$ with $x_i \in \{-1, 1\}$ as a indicator vector. When $x_i = 1$, segment S_i belongs

to Ω , and vice versa. As $\mathbf{x}$ and Ω correspond to each other, we can rewrite Eq. 6.2 as:

$$\begin{aligned} \mathbf{x}^* = \arg \max_{\mathbf{x}} \quad & f(\mathbf{x}) = \frac{1}{4}(\mathbf{1} + \mathbf{x})^T A(\mathbf{1} + \mathbf{x}), \\ \text{s.t.} \quad & x_i \in \{-1, 1\}, i = 1, \dots, N, \end{aligned} \quad (6.3)$$

Here, our objective function is $f(\mathbf{x}) = \frac{1}{4}(\mathbf{1} + \mathbf{x})^T A(\mathbf{1} + \mathbf{x})$ and the optimization of Eq. 6.3 is a binary quadratic programming (BQP) problem. Inspired by the solution in [97], we employ their method and solve the proposed problem efficiently.

6.3 Experiments

In this section, we first evaluate our method on the collected video dataset with the ground truth provided by the eye tracking system. Then, we test our method with two other public video datasets [50] [72]. Based on these datasets, we compare the proposed approach with the state-of-the-art saliency methods [13] [6]. After that, we show the scalability of our method by testing a large self-collected video dataset which includes more than one hundred videos. Finally, to show the advantages of the proposed trust region based sub-graph mining solution, we compare it with two other recently proposed methods [98] [99].

6.3.1 Datasets

Eye tracking dataset

To show our thematic video saliency detection is biologically plausible, we designed an experiment by using the eye tracker to record the gaze data. Five video sequences are collected from Youtube and the average length of video is about 40 seconds. All these videos have one or multiple thematic objects such as Steve Jobs' bibliography, or Prince

Table 6.1: The information of four datasets.

Dataset	Video No.	Average Length (Sec.)
Eye tracker	5	40
RSD saliency	13	20
Commercial	10	35
Video collection	104	31

William with his bride Kate. Ten participants are with normal sight or corrected normal sight. Five of them are males and the other half are females. All of them were asked to freely view the videos with the task to find the thematic objects. With this setting, we record the gaze information for total about 5000 video key-frames. The gaze points are used to obtain the ground-truth for the thematic saliency in a video. In specific, the ground truth is selected as the minimum size bounding box which includes all the gaze points. We will release all the videos and the corresponding ground truth in our website.

RSD saliency dataset [50]

This dataset contains videos from six genres: documentary, ad, cartoon, news, movie and surveillance. Twenty-three subjects are assigned to manually label the salient regions with one or multiple rectangles. Since this dataset is built with different aims from ours, we only choose ad and news videos to evaluate whether our algorithm can be used to effectively discover the thematic saliency of the videos.

Commercial video dataset [72]

To evaluate the proposed approach on thematic saliency estimation, we further test our method on ten video sequences from the commercial video dataset [72]. The manually labeled ground truth is provided. All of the sequences are commercial advertisements, where the length of videos ranges from 30 to 40 seconds.

Video collection dataset

In this dataset, we collect 104 videos from diverse categories such as news, commercials, documentary videos, etc.. The length of the video in this dataset ranges from 20 to 40 seconds. For different genre, we have several video clips share the same themes, which will be used in our method to implement the video thematic saliency detection across different video clips. Consequently, the scalability of our method on handling large scale data is obtained. Table 6.1 summarizes the four employed datasets.

6.3.2 Experimental Setting

To obtain the segment representation for the videos, key-frames are sampled at two frames per second in each video. Then every key-frame is segmented multiple times using normalized cut [100], with different number of segments K . In our implementation, we segment each key-frame into $K = 3, 5, 7, 9, 11$ and 13 segments, respectively. We perform normalized cut in both original key-frames as well as the down-sampled key-frames of half size of the original key-frames. After selection, we use several types of bag-of-features histograms to describe each segment: SIFT Histograms, Texton Histograms, Color Histograms, and pyramid of HOG [101]. SIFT features are first extracted from every key-frame and all the obtained features are then quantized into 1000 visual words by the k -means clustering for each video clip. For TH, we use a filter bank with 18 bar and edge filters. We quantize them to 400 textons via k -means. For CH, we use Lab color space, with 23 bins per channel. For pHOG, we use 3 pyramid levels with 8 bins. The concatenation of four types of feature histograms is used to describe each segment. The similarity of two segments is measured by using the normalized cross correlation (NCC) criterion.

With the obtained saliency map S_{map} , we select the segments which have large average saliency value. Specifically, we rank the segments by their average saliency values and then select the top 30% of them. To suppress segments which do not share the same appearance, the negative value τ is set to be -0.05. To deal with the constraint $\psi(x_i) = 0$, the penalty parameter $\beta > 0$ is set to be 10. Other parameters of the trust region methods are set similar with the modern trust region methods [102]. We fixed all these parameters during the experiments.

6.3.3 Performance Evaluation

In order to objectively measure the performance of our method on the employed datasets, we employ two criteria: $precision = \frac{|GT \cap DR|}{|DR|}$ and $recall = \frac{|GT \cap DR|}{|GT|}$, where DR and GT are the discovered bounding boxes and the bounding boxes of ground truth, respectively. To calculate the F -measure, $precision$ and $recall$ values for one video, the F -measure, $precision$ and $recall$ values are first calculated for each key-frame and then the average value of all key-frames is used to evaluate the whole video. As the co-saliency and image-saliency methods do not provide the results in the form of region, we select the region containing the same number of pixels with high saliency value as the proposed method obtained in each key-frame saliency map. For those key-frames that the thematic saliency is lost by our method but having image saliency/co-saliency, we select the top 30% salient pixels from image saliency/co-saliency map as their detection results.

Fig. 6.4(a) and (b) show the sample results of our method on the eye tracking dataset. In the video sequences, the thematic salient objects are subject to variations caused by partial occlusions, scales, viewpoints and lighting condition changes. But we still can see that our results get the best overlap with the gaze points. Moreover, the complete thematic salient objects or a large part of regions of them are detected by our method

Figure 6.4: Sample results of video saliency discovery. (a) and (b) show two videos of our eye tracking dataset. (c) shows one commercial video. The first rows of all three sub-figures show the original key-frame and the gaze points obtained by the eye tracking system are also rendered for (a) and (b). Each green cross ‘+’ represents one gaze point. The second rows show our result. The image saliency and co-saliency results are shown in the third and fourth rows respectively.

while only parts of the salient objects are detected by the co-saliency [13] and image-saliency methods [6]. From the last two columns in Fig. 6.4(a) and (b), we also can see that our method correctly avoid the false alarm when there is no thematic salient object occurring in some key frames. On contrary, the co-saliency and image-saliency algorithms are not so discriminative due to the lack of spatio-temporal context information. The *F-measure*, *precision* and *recall* curves for the five video sequences in Fig. 6.7(a) objectively validate that the obtained thematic video saliency corresponds to human’s task-driven attention.

In addition, we also evaluate the proposed approach on two other video datasets: RSD saliency dataset [50] and commercial video dataset [72]. Both datasets provide the manually labeled ground truth by bounding-box. In Fig. 6.4(c), some sample results of thematic saliency discovery on commercial datasets are presented. Fig. 6.7(b) and Fig. 6.7(c) show the *F-measure*, *precision* and *recall* curves on the two datasets. These results illustrate that the proposed approach gets excellent performance on discovering thematic salient objects from video sequences.

Figure 6.5: Thematic saliency discovery results from the video collection dataset. (a) shows several videos which contain the thematic salient object, i.e., “Starbuck Logo”. (b) shows the videos which do not contain the thematic salient object.

Table 6.2: The performance on video collection dataset.

Time for clip-level	Time for collection-level	Correct Rate
50 Sec.	20 Sec.	93%

To show the ability of our method on handling a large number of video clips simultaneously, we test on the video collection dataset. As a huge number of image segments will be obtained in this dataset, how to process them is a challenging problem. However, our method is free of this issue. A hierarchical implementation can be used to keep the affinity graph with a reasonable size. We detect the thematic object in each video clip firstly. After obtaining all clip-level thematic objects, we detect the collection-level video thematic saliency by using the sub-graph mining again with the clip-level thematic objects as graph nodes. Table 6.2 summarizes the performance of our method. After segmentation and feature extraction, it costs less than one minute to process one video clip averagely. Specifically, it overall requires 20 seconds for 104 videos on the collection-level and 50 seconds for every video on the clip-level. Meanwhile, the correct detection rate of our method is also high (about 93%). For example, we can find 14 out of 15 “Starbuck” videos. Fig. 6.5(a) shows the detected collection-level video thematic saliency for starbuck logo. Fig. 6.5(b) shows video clips which are processed simultaneously but do not contain the thematic saliency. In addition, the thematic saliency discovery results with/without segment selection are also compared in Fig. 6.6. It can be seen that a lot of background regions are detected as saliency when the image saliency S_{map} is not used. That’s because the highly repeated background, instead of the Broom and Bride, becomes the thematic saliency in this video.

Figure 6.6: The comparison of thematic saliency discovery results with/without segment selection. (a) the results without segment selection. (b) the results with segment selection. The discovered video thematic salient region is rendered in red.

6.4 Discussion and Conclusion

6.4.1 Discussions

The role of segmentation The segmentation provides the spatial constraint for the thematic saliency object. However, our method does not require the precise segmentation of the thematic objects because local features are used to estimate the similarity between segments. Even if some segments of the thematic objects are missed either by segmentation or saliency weighting, our method can still discover them considering their frequent occurrences in the whole video. After obtaining the segments, it is still challenge to find the thematic objects, as we do not know the shapes, appearances, scales, and locations of the thematic objects in advance. Therefore, it relies on the graph mining method to find the thematic objects and detect the thematic saliency. Last but not least, we use image segments to bundle local features, but there could be other ways to provide the spatial constraint for matching. For example, we can use the image patches selected at random to replace the image segments.

Figure 6.7: Comparisons of the proposed video saliency detection algorithm, the co-saliency algorithm [13], and the single image saliency algorithm [6]. (a) the eye tracker datasets. (b) the RSD dataset. (c) the commercial video dataset.

The case of long videos For a long video clip, which may contain many thematic objects, also could be similarly detected by a hierarchical implementation of our method as in Sec. 6.3.3. We can partition the video into shots/short clips and detect the thematic object in each shot/clip firstly. After obtaining all shot/clip thematic objects, the video thematic saliency can be obtained by using the sub-graph mining again with the obtained shot/clip thematic objects as graph nodes. To handle multiple thematic salient object detection, we can discover them one by one from the affinity graph by erasing the corresponding nodes belonging to previously detected ones.

6.4.2 Conclusion

Thematic saliency discovery in videos is a interesting problem which differs itself from traditional saliency detection methods by finding important positions in an image or a video. It targets to discover the saliency regarding to a video theme. It is also a challenging problem due to visual pattern variations of the thematic object and the loss of the prior knowledge of the thematic object. By representing the relations of all key-frame segments in a video as an affinity graph, we formulate the thematic saliency discovery problem as a cohesive sub-graph mining problem. A new algorithm has been proposed to obtain the cohesive sub-graph efficiently, which outperforms existing sub-graph mining methods. Our approach has the ability to identify the thematic saliency and accurately locate its regions in the cluttered and dynamic video scenes. Experimental evaluations on challenge video datasets demonstrate the effectiveness and the efficiency of the proposed method.

Chapter 7

Conclusion and Future Work

In this dissertation, we have investigated the problem of saliency estimation, salient object detection in images/videos and its applications in computer vision. Conclusion and future work will be discussed in this Chapter.

7.1 Conclusion

We summarize our research work in the following.

1. we propose a novel video saliency estimation method which is obtained by fusing spatio-temporally selected sparse features. By learning an over-complete dictionary for a given video, every patch in this video can be better represented. By selecting the learned sparse features corresponding to different scale of entropy gains, we propose to estimate spatial saliency from clip-level and frame-level, respectively. To fully consider the temporal correlations, we select sparse features which are contributed to temporal information from two aspects: temporal consistency aims to detect the common saliency among consecutive spatial saliency maps while tem-

poral difference ensures that the spatial common saliency is also temporal novel in the dynamic scenes. Instead of combining the spatial and the temporal saliency maps directly, we fuse the selected sparse features on different scales and obtain the overall patch saliency by summation of all the final selected feature responses. Since the proposed method can accurately represent each frame by the learned dictionary and incorporate the temporal information both consistently and novelly, our method outperforms the state-of-the-art methods which use a generalized dictionary or partial temporal coherence. Experimental results on two public video datasets further demonstrate the effectiveness of our method.

2. given saliency estimation results, we propose a novel method to efficiently detect salient objects from images and videos. Salient object detection is first formulated with the saliency density. A branch-and-bound search algorithm is developed to optimize the newly formulated problem globally. Without a prior knowledge of the salient object, our method can adapt to different sizes and shapes of the object, and is less sensitive to the cluttered background. The experiments on a public dataset show that our method greatly improves the existing baseline methods on the measurements of precision, recall and F-measure. Our method gains comparable performance compared to learning based salient object detection results with a high time efficiency. Tests on different saliency maps indicate our method works well with different types of saliency maps.
3. instead of detecting salient object in individual frames separately, we propose to detect and track salient object simultaneously by finding a spatio-temporal path of the highest saliency density in the video. As salient video objects usually appear in consecutive frames, leveraging the motion coherence of videos can detect salient object more robustly. Without any prior knowledge of the salient objects, our

method can automatically detect the salient objects of different shapes and sizes, and is able to handle noisy saliency maps and moving cameras. Experimental results on two public datasets demonstrate the effectiveness of the proposed method on salient video object detection.

4. similar to salient object detection, thematic object detection, which aims to find the salient objects representing the theme of the video content, is a challenging problem due to the possibly large visual pattern variations of the thematic object and the prohibitive computational cost to explore the candidate set without *a priori* knowledge of the thematic object. By representing the relations of all image segments in the video as an affinity graph, we formulate the thematic object discovery problem as a novel cohesive sub-graph mining problem. A novel algorithm is proposed to extract the cohesive sub-graph efficiently, which outperforms existing sub-graph mining methods. Our approach has the ability to identify the thematic saliency which is able to accurately locate the thematic object regions in the cluttered and dynamic video scenes. Experiments on challenge video datasets show that our method is efficient, robust and accurate.

To summary, this thesis has identified and addressed the problems in the area of saliency estimation and salient object detection in images and videos. It has shown that it is possible to predict human attentions by a sparse feature selection and weighting scheme when watching a video. It has demonstrated the efficiency of the proposed saliency density maximizing method on salient object detection. It also has shown the superiority by taking temporal coherence into consideration to detect salient object in a optimal path discovery framework. It also provided a new concept on thematic object detection which is salient globally by a sub-graph mining algorithm. Even though there are still some practical limitations to overcome, this thesis has provided a potential of further

improving on image or video salient object detection.

7.2 Further Research

In the buildup to this thesis, our research has been focused on video saliency estimation and salient object detection in images/videos. We looked for novel and feasible solutions for these problems, and inevitably have left something to be desired and more related topics to be explored.

Saliency estimation is still an open problem, especially video saliency. Although lots of saliency estimation methods have been proposed and we also developed several methods to generate video saliency maps which also got excellent performance on salient object detection and thematic object detection, there is still no unique definition to visual saliency and it is also a very interesting problem to validate the hypotheses from which various saliency methods have been proposed.

We have shown that the proposed methods got excellent performance on detecting salient object of any category. However, it is common that the expected salient object just appears in several frames across a video sequence. Since the class label of salient objects is ignored in most of the existing salient object detection solutions, salient objects from other classes and different from the target salient objects will be detected inevitably. We may focus on our future research to detect salient objects from a specific class, e.g. salient car, salient people, such that some irrelevant salient objects can be filtered out and better detection results could be achieved.

Accurate salient object detection is a fundamental problem to many applications in computer vision. This thesis has shown its advantages to video retargeting. Next, we may shift to explore other applications, such as saliency based video summarization,

salient action detection, action recognition and abnormal detection. Take salient action detection for example. With numerous of surveillance videos invaded into our life, efficient detection an important action or a salient event from a cluttered scene is a very challenge problem. Successful salient action or salient event detection could greatly reduce the response time to the emergence and save a lot of human power on many aspects.

Bibliography

- [1] X. Hou and L. Zhang, “Dynamic visual attention: Searching for coding length increments,” in *Advances in Neural Information Processing Systems (NIPS)*, pp. 681–688, 2008.
- [2] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [3] L. Itti and P. F. Baldi, “Bayesian surprise attracts human attention,” *Vision Research*, vol. 49, no. 10, pp. 1295–1306, 2009.
- [4] Y. Li, Y. Zhou, L. Xu, X. Yang, and J. Yang, “Incremental sparse saliency detection,” in *Proc. of International Conference on Image Processing (ICIP)*, pp. 3093–3096, 2009.
- [5] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, “Learning to detect a salient object,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2007.
- [6] R. Achanta, S. Hemami, F. Estraday, and S. Susstrunk, “Frequency-tuned salient region detection,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1597–1604, 2009.
- [7] N. Bruce and J. Tsotsos, “Saliency based on information maximization,” in *Neural Information Processing Systems (NIPS)*, pp. 155–162, 2005.
- [8] Y.-F. Ma and H.-J. Zhang, “Contrast-based image attention analysis by using fuzzy growing,” in *Proc. ACM Multimedia (MM)*, pp. 374–381, 2003.
- [9] T. Lu, Z. Yuan, Y. Huang, D. Wu, and H. Yu, “Video retargeting with nonlinear spatial-temporal saliency fusion,” in *Proc. of International Conference on Image Processing (ICIP)*, pp. 1801–1804, 2010.

-
- [10] T. Deselaers, P. Dreuw, and H. Ney, “Pan, zoom, scan time-coherent, trained automatic video cropping,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.
 - [11] Y. Luo, J. Yuan, P. Xue, and Q. Tian, “Saliency density maximization for efficient visual objects discovery,” *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 21, no. 12, pp. 1822–1834, 2011.
 - [12] X. Jia, H. Lu, and M.-H. Yang, “Visual tracking via adaptive structural local sparse appearance model,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1822–1829, 2012.
 - [13] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, “From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2129–2136, 2011.
 - [14] L. Itti, “CRCNS data sharing: Eye movements during free-viewing of natural videos,” in *Collaborative Research in Computational Neuroscience Annual Meeting, Los Angeles, California*, 2008.
 - [15] X. D. Hou and L. Q. Zhang, “Saliency detection: A spectral residual approach,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2007.
 - [16] R. Valenti, N. Sebe, and T. Gevers, “Image saliency by isocentric curvedness and color,” in *International Conference on Computer Vision (ICCV)*, pp. 2185–2192, 2009.
 - [17] B. Han, H. Zhu, and Y. Ding, “Bottom-up saliency based on weighted sparse coding residual,” in *Proc. ACM International conference on Multimedia (ICME)*, pp. 1117–1120, 2011.
 - [18] H.-T. Chen, “Preattentive co-saliency detection,” in *Proc. of International Conference on Image Processing (ICIP)*, pp. 1117–1120, 2010.
 - [19] H. Li and K. N. Ngan, “A co-saliency model of image pairs,” *IEEE Trans. on Image Processing (TIP)*, vol. 110, no. 3, pp. 346 – 359, 2011.
 - [20] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, “From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2129 – 2136, 2011.

-
- [21] B. Alexe, T. Deselaers, and V. Ferrari, “What is an object,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 73–80, 2010.
 - [22] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, “Fusing generic objectness and visual saliency for salient object detection,” in *International Conf. on Computer Vision (ICCV)*, pp. 914–921, 2011.
 - [23] Y. Luo, J. Yuan, and Q. Tian, “Spatio-temporal enhanced sparse feature selection for video saliency estimation,” in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 33–38, 2012.
 - [24] Y. Luo, J. Yuan, P. Xue, and Q. Tian, “Saliency density maximization for object detection and localization,” in *Proc. of 10th Asian Conference on Computer Vision (ACCV)*, pp. 396–408, 2010.
 - [25] Y. Luo, J. Yuan, P. Xue, and Q. Tian, “Salient region detection and its application to video retargeting,” in *IEEE Conference on Multimedia Expo (ICME)*, pp. 1–6, 2011.
 - [26] Y. Luo, J. Yuan, and Q. Tian, “Salient object detection in videos by optimal spatial-temporal path discovery,” in *ACM multimedia (MM)*, 2013.
 - [27] W. James, *The principles of psychology*, vol. 1. New York: Henry Holt, 1890.
 - [28] D. E. Broadbent, *Perception and Communication*. London: Pergamon, 1958.
 - [29] A. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.
 - [30] D. W. Hansen and Q. Ji, “In the eye of the beholder: A survey of models for eyes and gaze,” *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, pp. 478–500, 2010.
 - [31] Wikipedia, “Reading fixations saccades.jpg,” 2007.
 - [32] Z. S. Wang and B. X. Li, “A two-stage approach to saliency detection in images,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 965–968, 2008.
 - [33] Z. Liu, Y. Xue, L. Shen, and Z. Zhang, “Nonparametric saliency detection using kernel density estimation,” in *Proc. of International Conference on Image Processing (ICIP)*, pp. 253–256, 2010.

-
- [34] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Proc. of European Conference on Computer Vision (ECCV)*, vol. 7573, pp. 414–429, 2012.
 - [35] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 31, no. 6, pp. 989–1005, 2009.
 - [36] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2368–2375, 2010.
 - [37] A. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research*, vol. 37, pp. 3327–3338, 1997.
 - [38] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 1–20, 2008.
 - [39] J. Yan, M. Zhu, H. Liu, and Y. Liu, "Visual saliency detection via sparsity pursuit," *IEEE Singal Processing Letters (SPL)*, vol. 17, pp. 739 – 742, 2010.
 - [40] C. Lang, G. Liu, J. Yu, and S. Yan, "Saliency detection by multi-task sparsity pursuit," *IEEE Trans. on Image Processing (TIP)*, vol. 21, no. 3, pp. 1327–1338, 2012.
 - [41] D. E. Jacobs, D. B. Goldman, and E. Shechtman, "Cosaliency: Where people look when comparing images," in *Proc. ACM symposium on User Interface Software and Technology (UIST)*, pp. 219–228, 2010.
 - [42] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2097–2104, 2011.
 - [43] L. Zhang, M. H. Tong, and G. W. Cottrell, "SUNDAY: Saliency using natural statistics for dynamic analysis of scenes," in *the Thirty-first Annual Cognitive Science Society Conference*, pp. 1–6, 2009.
 - [44] H. J. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, pp. 1–27, 2009.
 - [45] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, pp. 1–24, 2009.

-
- [46] S. Marat, T. HoPhuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Gurin-Dugu, “Modelling spatio-temporal saliency to predict gaze direction for short videos,” *International Journal of Computer Vision (IJCV)*, vol. 82, pp. 231–243, 2009.
 - [47] Y. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues,” in *Proc. ACM International Conference on Multimedia (MM)*, pp. 815–824, 2006.
 - [48] C. Chamaret, J. Chevet, and O. Le Meur, “Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies,” in *Proc. of International Conference on Image Processing (ICIP)*, pp. 1077–1080, 2010.
 - [49] T. Liu, Z. J. Yuan, J. Sun, J. D. Wang, N. N. Zheng, X. Tang, and H. Shum, “Learning to detect a salient object,” *Trans. Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 33, no. 2, pp. 353–367, 2010.
 - [50] J. Li, Y. Tian, T. Huang, and G. Wen, “Probabilistic multi-task learning for visual saliency estimation in video,” *International Journal of Computer Vision (IJCV)*, vol. 90, pp. 150–165, 2010.
 - [51] J. Li, Y. Tian, T. Huang, and W. Gao, “Multi-task rank learning for visual saliency estimation,” *IEEE Trans. on Circuits and Systems for Video Technology (TCSVT)*, vol. 21, no. 5, pp. 623–636, 2011.
 - [52] L. Itti and P. Baldi, “Bayesian surprise attracts human attention,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 19, pp. 547–554, 2006.
 - [53] K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, “Real-time estimation of human visual attention with dynamic bayesian network and MCMC-based particle filter,” in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pp. 250–257, 2009.
 - [54] V. Mahadevan and N. Vasconcelos, “Spatiotemporal saliency in dynamic scenes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, pp. 171–177, 2010.
 - [55] W. Kim, C. Jung, and C. Kim, “Spatiotemporal saliency detection and its applications in static and dynamic scenes,” *IEEE Trans. on Circuits and Systems for Video Technology (TCSVT)*, vol. 21, no. 4, pp. 446–456, 2011.
 - [56] F. Liu and M. Gleicher, “Automatic image retargeting with fisheye-view warping,” in *Proc. ACM symposium on User interface software and technology (UIST)*, pp. 153–162, 2005.

-
- [57] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Salient region detection by modeling distributions of color and orientation," *IEEE Trans. on Multimedia (TMM)*, vol. 11, no. 5, pp. 892–905, 2009.
 - [58] P. Khuwuthyakorn, A. Robles-Kelly, and J. Zhou, "Object of interest detection by saliency learning," in *European Conference on Computer Vision (ECCV)*, pp. 636–649, September 2010.
 - [59] J. W. Han, K. N. Ngan, M. Li, and H. J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. on Circuits and Systems for Video Technology (TCSVT)*, vol. 16, no. 1, pp. 141–145, 2006.
 - [60] B. C. Ko and J.-Y. Nam, "Object-of-interest image segmentation based on human attention and semantic region clustering," *Virtual Journal for Biomedical Optics*, vol. 1, no. 11, pp. 2462–2470, 2006.
 - [61] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Efficient subwindow search: A branch and bound framework for object localization," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 31, no. 12, pp. 2129–2142, 2009.
 - [62] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *International Conference on Computer Vision (ICCV)*, pp. 2232–2239, 2009.
 - [63] X. Fan, X. Xie, H. Zhou, and W. Ma, "Looking into video frames on small displays," in *Proc. of ACM Multimedia 2003 (MM)*, pp. 247–250, 2003.
 - [64] T. Liu, N. Zheng, W. Ding, and Z. Yuan, "Video attention: Learning to detect a salient object sequence," in *Proc. IEEE Conference on Pattern Recognition (ICPR)*, pp. 1–4, 2008.
 - [65] C. Tao, J. Jia, and H. Sun, "Active window oriented dynamic video retargeting," in *Workshop on Dynamic Vision at International Conference on Computer Vision (ICCV-WDV)*, pp. 1–12, 2007.
 - [66] F. Liu and M. Gleicher, "Video retargeting: Automating pan and scan," in *Proc. ACM Multimedia (MM)*, pp. 241–250, 2006.
 - [67] G. Hua, C. Zhang, Z. Liu, Z. Zhang, and Y. Shan, "Efficient scalespace spatiotemporal saliency tracking for distortion-free video retargeting," in *Asian Conference on Computer Vision (ACCV)*, pp. 1597–1604, 2009.

-
- [68] M.-M. Wang, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, “Global contrast based salient region detection,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 409–416, 2011.
 - [69] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2376–2383, 2010.
 - [70] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. Rowley, “Image saliency from intrinsic to extrinsic context,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 417–424, 2011.
 - [71] H. Li and K. N. Ngan, “A co-saliency model of image pairs,” *IEEE Trans. on Image Processing (TIP)*, vol. 110, no. 3, pp. 346 – 359, 2011.
 - [72] G. Zhao, J. Yuan, J. Xu, and Y. Wu, “Discovering the thematic object in commercial videos,” *IEEE MultiMedia Magazine*, vol. 18, no. 3, pp. 56–65, 2011.
 - [73] V. Mahadevan and N. Vasconcelos, “Spatiotemporal saliency in dynamic scenes,” *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, p-p. 171–177, 2010.
 - [74] X. Shen and Y. Wu, “A unified approach to salient object detection via low rank matrix recovery,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 853–860, 2012.
 - [75] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li, “Salient object detection for searched web images via global saliency,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3194 – 3201, 2012.
 - [76] H. El-Alfy, D. Jacobs, and L. Davis, “Multi-scale video cropping,” in *Proc. ACM Multimedia (MM)*, pp. 97–106, 2007.
 - [77] L. Wolf, M. Guttman, and D. Cohen-Or, “Non-homogeneous content-driven video-retargeting,” in *International Conference on Computer Vision (ICCV)*, pp. 1–6, 2007.
 - [78] W. Yin, J. Luo, and C. W. Chen, “User guided semantic image adaptation for mobile display devices,” in *Proc. Intl. Conf. on Multimedia and Expo (ICME)*, pp. 418–423, 2010.
 - [79] Y. Ma, X. Hua, L. Lu, and H. Zhang, “A generic framework of user attention model and its application in video summarization,” *IEEE Transactions on Multimedia (TMM)*, vol. 7, no. 5, pp. 907–919, 2005.

-
- [80] J. You, G. Liu, and H. Li, “A multiple visual models based perceptive analysis framework for multilevel video summarization,” *IEEE Trans. on Circuits and Systems for Video Technology (TCSVT)*, vol. 17, no. 3, pp. 273–285, 2007.
 - [81] H. Lee, A. Battle, R. Raina, and A. Y. Ng, “Efficient sparse coding algorithms,” in *Advances in Neural Information Processing Systems (NIPS)*, pp. 801–808, 2007.
 - [82] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, “Saliency-based video segmentation with graph cuts and sequentially updated priors,” in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pp. 638–641, 2009.
 - [83] E. L. Lawler and D. E. Wood, “Branch-and-bound methods: A survey,” *Operations Research*, vol. 14, no. 4, pp. 699–719, 1966.
 - [84] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, pp. 62–66, 1979.
 - [85] Y.-F. Ma, X.-S. Hua, L. Lie, and H.-J. Zhang, “A generic framework of user attention model and its application in video summarization,” *IEEE Trans. on Multimedia (TMM)*, vol. 7, no. 5, pp. 907–919, 2005.
 - [86] L. Shi, J. Wang, L. Duan, and H. Lu, “Consumer video retargeting: Context assisted spatial-temporal grid optimization,” in *Proc. ACM Multimedia (MM)*, pp. 301–310, 2009.
 - [87] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.
 - [88] G. Sperling and E. Weichselgartner, “Episodic theory of the dynamics of spatial attention,” *Psychological Review*, vol. 102, no. 3, pp. 503–532, 1995.
 - [89] D. Walther, U. Rutishauser, C. Koch, and P. Perona, “Selective visual attention enables learning and recognition of multiple objects in cluttered scenes,” *Computer Vision and Image Understanding (CVIU)*, vol. 100, pp. 41–63, 2005.
 - [90] N. A. Goussies, Z. Liu, and J. Yuan, “Efficient search of top-k video subvolumes for multi-instance action detection,” in *Proc. of International Conference on Multimedia and Expo.(ICME)*, pp. 328–333, 2010.
 - [91] C. H. Lampert, M. B. Blaschko, and T. Hofmann, “Beyond sliding windows: Object localization by efficient subwindow search,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

- [92] D. Tran and J. Yuan, “Optimal spatio-temporal path discovery for video event detection,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3321–3328, 2011.
- [93] C. Liu, *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, MIT, 2009.
- [94] M. D. Rodriguez, J. Ahmed, and M. Shah, “Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.
- [95] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, “Slic superpixels compared to state-of-the-art superpixel methods,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.
- [96] P. Viola and M. J. Jones, “Robust real-time face detection,” *International Journal of Computer Vision (IJCV)*, vol. 57, pp. 137–154, 2004.
- [97] G. Zhao, “Cohesive sub-graph mining using trust region method,” tech. rep., Nanyang Technological University, 2012.
- [98] J. Xu, J. Yuan, and Y. Wu, “Learning spatio-temporal dependency of local patches for complex motion segmentation,” *Computer Vision and Image Understanding (CVIU)*, vol. 115, pp. 334–351, March 2011.
- [99] M. Pavan and M. Pelillo, “Dominant sets and pairwise clustering,” *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 29, pp. 167–172, 2007.
- [100] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 22, pp. 888–905, 2000.
- [101] Y. J. Lee and K. Grauman, “Object-graphs for context-aware visual category discovery,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 99, no. PrePrints, pp. 346–58, 2011.
- [102] C.-J. Lin and J. J. More, “Newton’s method for large bound-constrained optimization problems,” *SIAM Journal on Optimization*, vol. 9, pp. 1100–1127, 1998.