Detection of visual attention regions in images and videos

Hu, Yiqun

2008

Hu, Y. Q. (2008). Detection of visual attention regions in images and videos. Doctoral thesis,
Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/18864>

<https://doi.org/10.32657/10356/18864>

Detection of Visual Attention Regions in Images and Videos

Yiqun Hu

School of Computer Engineering

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

2008

Abstract

The explosive growth of multimedia content and advances in the development of hardware with multimedia functionalities call for techniques to enable users to access such content anywhere and anytime and with similarly pleasing experience each time. This requires intelligent search, transmission, analysis and display of multimedia data. However, in addition to the data being very large in size, it is inherently complex due to the variety of features (color, texture, shapes, motion, etc.) that it contains. The challenge then is to detect information from the clutter for further processing. The relevant information is the visual attention region (VAR) whose detection in images and videos is the topic of this dissertation.

The bottom-up model for detecting VAR in an image involves generation of a saliency map that highlights contrasts in features like color, intensity and orientation. The saliency map itself is obtained through a combination of each feature map that highlights the contrast for that particular feature. We investigate the process of good selection and proper combination strategies for the features. We propose a novel Composite Saliency Indicator (CSI) to determine the contribution of each feature map to the salient region. CSI is designed to capture the spatial compactness as well as the density of candidate regions in the feature maps. We also propose a Context Suppression Model that provides a measure to determine similarity among candidate attention regions in a feature map. This measure is used to find a suppression factor for a particular patch in the scene, which is then used to highlight actual attention regions. We also demonstrate an application in multimedia adaptation that benefits from the improved VAR detection.

An object-based attention paradigm has been corroborated by research in the psychophysics and neuroscience. Therefore, we investigate the possibility of developing algorithms for VAR detection based on visual attention as a global process that considers

properties of regions instead of as a process of local contrast measurement. A polar transformation maps homogeneity in the features into a linear subspace that also encodes spatial information of a region. A new subspace estimation algorithm based on the Generalized Principal Component Analysis (GPCA) is proposed to estimate multiple linear subspaces. The robustness of subspace estimation is improved in two ways. Sensitivity to outliers is achieved by weighted least squares estimate of the subspaces in which weights calculated from the distribution of K nearest neighbors are assigned to data points. Iterative refinement of the weights is proposed to handle the issue of estimation bias when the number of data points in each subspace is very different (imbalanced subspaces). A new region attention measure is defined to calculate the visual attention of each region by considering both feature contrast and spatial geometric properties of the regions.

Moving from images to videos, we introduce the notion of attention-from-motion in which the objective is to identify, from an image sequence, only those objects in motion that capture visual attention. Following an important concept in film production, viz., the tracking shot, we define attention objects in motion (AOM) as those that are tracked by the camera. When the camera tracks an object in motion, the factorization relating to the motion matrix attains a special structure. We describe the method to identify this structure without carrying out the complete factorization. Finally, a recursive framework is proposed to robustly extract more than one AOM without resorting to full-fledged motion segmentation of all moving objects. The proposed solution to the attention-from-motion problem is not constrained by the type of camera motion or object motion including degenerate motion as well as dependent motion. Furthermore, the algorithm can detect AOMs even when only part of the trajectory information is available (i.e., missing data) or when only part of a period is observable in the case of periodic motion.

Acknowledgments

First and foremost, I would like to express my great appreciation to my supervisors, Dr. Deepu Rajan and Dr Liang-Tien Chia, for their invaluable instructions and support throughout my research efforts towards this dissertation. Their insights and suggestions to the research area enlightened me in various detailed aspects throughout the work. Their instructions with regards to the development of my academic thinking and writing were inspiring and helpful for my current and future work. I would also like to thank Dr Tan Ah Hwee for his job offer, which also support me to continue my research in this dissertation.

During the past four years, my colleagues in the Center for Multimedia and Networking Technology (CeMNet) have provided me much appreciated help in my daily work. I would like to thank all of them for their encouragement on my study and their invaluable support. I would like to thank the School of Computer Engineering of Nanyang Technological University for the financial support and excellent working environment during my past years of study. I would also like to thank Microsoft Research Asia (MSRA) and Dr. Xing Xie for providing me two internships in MSRA.

Furthermore, I would give my thankfulness to all staffs and students in CeMNet. They're Mrs Ling Chen, Mrs Song Liu, Mr Haoran Yi, Mrs Surong Wang, Ms Huan Wang, Mr Jiaming Li, Mr Ziming Zhang, Mr Xiangang Cheng, Ms Chua Poo Hua, et al.

Finally, I will give my special thanks to my dearest mother. Without the strong support from her, I cannot finish the research in this dissertation.

Contents

Abstract	i
Acknowledgments	iii
List of Figures	vii
List of Tables	xi
1 Introduction	1
1.1 Visual Attention	1
1.2 Motivation	3
1.3 Research Objectives and Contributions	5
1.3.1 Main Contributions	6
1.3.2 Novelty and Originality	7
1.4 Dissertation Organization	8
2 Literature Survey	10
2.1 Computational Models of Visual Attention	10
2.1.1 Biologically Plausible Models	11
2.2 Computational Bottom-up Models	14
2.2.1 Feature Contrast based Methods	17
2.2.2 Region and Object based Methods	18
2.2.3 Other Methods	20
2.3 Top-down Modulation & Learning	22
2.3.1 Top-down Modulation	22
2.3.2 Learning for Visual Attention Detection	24
2.4 Applications of Visual Attention Detection	25
2.4.1 Computer Vision Applications	25

2.4.2	Multimedia Information Management	26
2.4.3	Multimedia Adaptation	27
3	Adaptive Feature Selection and Combination	29
3.1	Introduction	29
3.2	Composite Saliency Indicator (CSI)	31
3.2.1	Salient Point Selection	31
3.2.2	Spatial Compactness	32
3.2.3	Saliency Density	33
3.2.4	Feature Combination Strategy	33
3.3	Spatial Context Suppression	39
3.3.1	Texture Attention Cue	39
3.3.2	Feature combination through Context Suppression	41
3.4	More experiments	45
3.4.1	Evaluation in MPEG-21 DIA Adaptation System	47
3.5	Conclusion	51
4	Subspace analysis for detection of VARs in images	53
4.1	Introduction	53
4.2	Image representation	56
4.2.1	Scale-sentient features	57
4.2.2	Polar transformation of features	60
4.3	NN-GPCA Subspace Estimation	63
4.3.1	Review of <i>GPCA</i>	63
4.3.2	Assigning weights to data points	66
4.3.3	Weighted Least Square Approximation of c	68
4.3.4	Iterative Refinement of c	70
4.3.5	Computational Complexity	74
4.4	Subspace based region attention measure	76
4.4.1	Cluster Compactness	76
4.4.2	Cumulative Projection	78
4.5	Experimental Results	81

4.5.1	Experiment 1	82
4.5.2	Experiment 2	85
4.5.3	Experiment 3	85
4.5.4	Experiment 4	90
4.6	Conclusions	92
5	Attention-from-Motion	94
5.1	Introduction	94
5.1.1	Attention-from-Motion	95
5.2	Factorization Method for AOM	97
5.2.1	Dynamic Scene Factorization	97
5.2.2	Factorization of AOM	100
5.2.3	Why not simply check for stationary centroid?	101
5.3	Attention-from-Motion for a single AOM	103
5.3.1	Constraints on $\hat{M}$ and Q	104
5.3.2	Attention Measure from $\bar{Q}_T$ and Q	106
5.3.3	Integrated Attention Measure	109
5.4	Attention-from-motion for multiple AOMs	110
5.4.1	Shape Space Initialization	110
5.4.2	Motion Space Expansion	111
5.4.3	Iterative AOM Extraction by Voting	113
5.5	Experimental Results	117
5.5.1	Synthetic Data	117
5.5.2	Real image sequences	120
5.6	Conclusion	126
6	Conclusions	127
6.1	Summary of Contributions	127
6.2	Future Research Issues	131
	References	133
	List of Publications	149

List of Figures

3.1	Example of a convex hull to determine the spatial compactness measure; (a) Original Image; (b) Intensity Feature Map; (c) Salient Point Set; (d) Convex Hull.	33
3.2	Feature combination using (a) CSI (b) global non-linear combination of [1] and (c) iterative local competition [2].	36
3.3	Examples of feature combination using CSI.	37
3.4	Use of CSI on the visual attention model of [3]; (a) Original image, Saliency map generated (b) without CSI and (c) its thresholded VAR; (d) using CSI and (e) its thresholded VAR.	38
3.5	(a) Original image (b) Texture map (c) Intensity map and (d) Color map using [1].	41
3.6	(a) Original image and (b) Eigen ellipsoids for each of the patches in (a).	42
3.7	The process of local context suppression.	44
3.8	Illustration of the usage of Context Suppression Model to combine multi-scale feature contrasts of [1] (a) saliency map generated without context suppression and (c) its thresholded visual attention region; (d) saliency map using Context Suppression and (e) its thresholded visual attention region.	45
3.9	Comparison of CSI and [1] (a) Original Image; (b) Saliency map of [1] and (c) its thresholded attention region (d) Saliency map using CSI and (e) its thresholded attention region.	46
3.10	Comparison of Context Suppression Model with [2] (a) Original Image; saliency maps using (b) proposed context suppression model and (c) the thresholded attention region; (d) Itti's iterative local competition model and (e) the thresholded attention region.	48

3.11	Mean DR v/s Gaussian Noise for various models.	49
3.12	(a) Metadata related to the extracted VAR in XML form; (b) Adaptation of an image from PC to PDA.	50
3.13	Example of Digital Item BSD Adaptation; (a) Adaptation Decision Description; (b) JPEG2000 BSD Adaptation (Green - Original BSD, Blue - Adapted BSD).	51
4.1	Illustration of local scale estimation on a synthetic image.	59
4.2	(a) Synthetic monochrome image and (b) polar transformation of its intensity. (c) Synthetic color image with noise and (d) polar transformation of its hue.	62
4.3	Effect of outliers on subspace estimation by <i>GPCA</i> (a) Synthetic data (b) estimation using true data only (without outliers); (c) initial estimate used and (d) final estimate after optimization.	66
4.4	Effect of subspace skew on subspace estimation by <i>GPCA</i> (a) Synthetic data (b) initial estimate and (c) final estimate after optimization.	67
4.5	Improved subspace estimation using <i>NN-GPCA</i> for data with outliers (a) initial estimate and (b) final estimate after optimization (compare with Figure 4.3).	73
4.6	Improved estimation using <i>NN-GPCA</i> for data with subspace skew (a) estimation after 3rd iteration; (b) initial estimate after convergence and (c) final estimate after optimization (compare with Figure 4.4).	73
4.7	<i>NN-GPCA</i> algorithm on real image (a) Original image and (b) estimation of the three subspaces using <i>hue</i> feature.	74
4.8	Images to illustrate effect of region size and location on Cumulative Projection	79
4.9	Illustration of <i>CP</i> function in hue and intensity,	81
4.10	Multiple feature analysis: (a) feature map, (b) subspace estimation and (c) extracted attention region using Top row: Hue feature and Bottom row: Intensity feature.	82

4.11	Comparison of <i>NN-GPCA</i> and <i>GPCA</i> algorithms in terms of estimation error of subspaces: (a) error v/s percentage of outliers in data, (b) error v/s ratio of subspace skew, (c) the sum and (d) the variance of the residuals during estimation.	84
4.12	Comparison of proposed subspace method with [1] (a) saliency map; (b) VAR from (a); (c) attention region using the proposed method; (d) VAR region from (c).	86
4.13	VAR extraction across different scales: (left) small, (center) medium and (right) large using (i) the proposed method and (ii) method of [3].	87
4.14	The invariance to the change of reference point (a) original image; attention regions when reference point is at (b) image center (160,120); (c) (35,150); (d) (100,60). (The origin is at the top-left corner.)	88
4.15	The invariance to rotation and to illumination change due to change in contrast (a) original images; (b) the detected attention regions; (c) the transformed images after rotating by 45° and reducing contrast by 50%; (d) the detected attention regions in (c).	88
4.16	(a) Original image; (b) attention region and (c) VAR when scale-sentient features are not used; (d) attention region and (e) VAR when scale-sentient features are used.	89
4.17	Two examples of automatic feature selection.	90
4.18	Two failed examples where the original images are shown in (a), the polar transformation and the detected subspaces are shown in (b) and (c, d) show the wrongly detected VAR and the corresponding bounding box. . .	91
5.1	Illustration of Attention Object in Motion (the dashed square indicates the field of view of the camera). The camera follows the car on the top-left (AOM) while ignoring the car on the bottom-right (not AOM).	96
5.2	The reference coordinate system used in the factorization of dynamic scene.	99
5.3	An example of unstable centroid of AOM where only partial feature points on an AOM are observed.	102
5.4	Another example of unstable centroid of AOM where the centroid of motion is followed and only partial frames are observed.	103

5.5	An example of the failure to detect feature points (indicated by small blue circles). There are two types of errors introduced: 1. out-of-view/occlusion error (area indicated by green ellipse); 2. tracking error (area indicated by red ellipse).	116
5.6	(a), (b), (c) Top row: Objects in synthetic image sequences; Bottom row: Detected AOMs in red.	118
5.7	Precision and recall of feature points of AOM(s) <i>w.r.t</i> noise level.	120
5.8	Top row: All feature points; Bottom row: Detected AOMs for (a) Seq 1, (b) Seq 2, (c) Seq 3. In the top row, the number of feature points <i>on</i> an AOM need not be larger than that outside the AOM (figure best viewed in color).	122
5.9	Top row: All feature points; Bottom row: Detected AOMs for (a) Seq 4, (b) Seq 5 and (c) Seq 6. In the top row, the number of feature points <i>on</i> an AOM need not be larger than that outside the AOM (figure best viewed in color).	123
5.10	Detection of AOM with partial occlusion over a very short interval (6 frames). First row shows 3 frames with all feature points and second row shows the detected feature points belonging to AOM.	125

List of Tables

2.1	Comparative summary of representative computational models of bottom-up visual attention.	16
3.1	Rules for Feature Map Combination	35
4.1	Cumulative Projection values	79
4.2	Estimate Error on synthetic data of 2 subspaces + 20% outliers	84
5.1	Average and maximum error of image coordinates in W under different noise level.	119
5.2	Comparison with simply checking for stationary centroid. We calculate the proposed attention measure L_q/S_q and S_c of (Eq. 5.32) on the objects for the 3 sequences in Figure 5.9. The ground truth AOM(s) are highlighted in bold	124

Chapter 1

Introduction

The past decade has witnessed an explosive growth of multimedia information, due to the advances in hardware technologies for processing multimedia content and the ease of creating multimedia content thanks to low cost video cameras and camera-enabled handphones. More and more applications utilize, process and analyze multimedia data instead of plain text. For information retrieval, the search engines for images and videos are becoming popular. For content publishing and consumption, image sharing and video podcast may provide more interesting content than text blog. More and more users consume the multimedia content on mobile devices, which require multimedia transmission over internet. For content analysis, multimedia content mining has the same importance as text mining. For many human computer interaction related applications, the captured image and video is another form of input.

1.1 Visual Attention

Visual Attention (VA) refers to the ability of any vision system to select a certain subset of visual sensory information for further processing within the first few milliseconds that the scene is presented [4]. This mechanism serves as the information processing bottleneck that allows only some part of information to reach short-term memory and visual awareness [5]. Only the visual information in the "fovea" region is processed and other

CHAPTER 1. INTRODUCTION

information outside this visual field is suppressed in such "focus-of-attention" mechanism. The most important function of such selective visual attention is to direct the eye gaze rapidly towards object-of-interest in a visual environment [4]. This ability is of evolutionary significance because it allows an organism to detect quickly possible prey, mates or predators in the visual world. In the human visual system (HVS) too, VA plays the role of directing the gaze rapidly towards *interesting* information related to current behaviors or tasks in the environment by ignoring irrelevant information [4]. Given the large amount of visual sensor data, only the selected attentional information is further analyzed and processed. Visual attention is a complex concept and has diverse interpretations in psychology, neuroscience and vision research, leading to different research methodologies as well as the evaluation criteria in these areas.

The taxonomies of visual attention are derived from specific tasks that use it. Considering eye movements, visual attention can be either *overt* to drive and guide eye movements, picking up useful information over time [6, 7], or *covert*, where the focus of attention shifts internally from one locus to another without explicit eye movements [8]. In the case of fixing attention to a target, visual attention has three forms: (i) feature-based attention where attention is deployed to different features; (ii) space-based attention where attention is deployed to different locations; (iii) object-based attention where attention is deployed to different objects or perceptual groups. According to the control of attention deployment, visual attention can be driven in either *bottom-up* manner or *top-down* manner [4]. In the bottom-up manner, visual attention is purely driven by visual data. In top-down manner, high-level information like the goal and preferences of the observers can modulate and guide visual attention. These two types of visual attention can also interact with each other. Considering the time of occurrence, visual attention can be either *early selection* [9] or *late selection* [10]. In early selection model, a selective filter selects among different channels where visual data are partitioned. A

CHAPTER 1. INTRODUCTION

limited capacity channel then processes the selected channel. In late selection model, all information is completely processed and recognized before it receives attention of a limited capacity processor.

Visual attention works like a "spotlight" or "zoom lens", scanning the scene by shifting attention from one part to the next to limit processing to a variable amount of information in the visual field [11]. When observing a scene, the *fovea area* [12, 13] around the center of gaze fixes on some part and captures visual information at high resolution while the resolution is reduced in the *peripheral area*. Such fixation is shifted from one part to another and these consequent shifts of fixations are represented as *saccades* [14]. Through this specific scanpath [15], the relevant visual information are integrated and processed to obtain the understanding of the observed scene. The development of algorithms for automatic detection of visual attention regions (VARs) allows irrelevant data to be removed for further processing. In this way a balance can be achieved between the allocation of limited computing resources and time for computation. In addition, complex tasks could be handled in cluttered and dynamic scenes.

1.2 Motivation

Although there are many interesting applications of multimedia, there are two issues that limit even more widespread use of multimedia content:

Huge capacity Compared to static text, both image and video content can grow explosively in number as well as size. This property make it impossible to process all the visual data at the same time. This property affects the storage, transmission as well as display of multimedia content on heterogeneous environments, especially for mobile computing platform;

High complexity The visual information is much more complex than other modalities of information. The interesting visual information are often presented in the

CHAPTER 1. INTRODUCTION

cluttered environment. Processing and analyzing all visual information is computationally expensive and not robust.

Current research on vision and multimedia has proposed different solutions for different problems such as object recognition, image retrieval, etc. However, most of the proposed algorithms are not robust enough to be applied in real applications.

There are two directions to overcome this limitation. One way is to develop more advanced techniques to improve robustness in the presence of noise and outliers. For example, statistical methods can be applied to improve robustness, but because of the high complexity of visual data, robust statistical methods could result in increased computational complexity. Another way to overcome the limitation is to select relevant visual information from multimedia content and process only that part of the data. This can simultaneously improve the efficiency and robustness. The concept of active vision has been proposed in the beginning of last decade in computer vision community [16, 17], which confirms that vision system should be able to actively obtain relevant visual information to understand the visual world. However, many recent vision algorithms do not have an attentional content detector because an effective solution to detect relevant visual information has not been proposed, especially for multimedia applications.

The motivation of this research comes from its usefulness to many potential applications in multimedia and computer vision. In the literature of multimedia information management, images can be adapted [18] and browsed [19], video sequences can be progressively transmitted [20] for display on small screen devices according to the visual attention region. For the applications of image retrieval, visual attention regions can improve the system performance by reducing the influence of cluttered background [21, 22]. In the literature of computer vision, algorithms for different vision tasks can benefit from the information of visual attention region. In active vision [16, 17], the vision system should be able to select the relevant information for processing while removing irrelevant

CHAPTER 1. INTRODUCTION

information. Visual attention region detection can naturally facilitate solutions to the visual tracking problem because of the common issues that they address: 1) selection of a salient (tracking) target from background; 2) transfer of the attention (tracking) information to a higher processing stage; 3) shift of attention (tracking) to a new location [23]. When a tracking target is lost, we can rely on visual attention region detection to recover the target [24]. Changes of shifting pattern of visual attention have also been used for video segmentation [25]. In summary, many multimedia and computer vision applications can benefit from the detection of visual attention region(s).

1.3 Research Objectives and Contributions

Visual attention has long been studied in the areas of psychology and neuroscience [26, 4]. It is only recently that the image processing/computer vision community has started to research actively on automating the detection of VA in images and videos. Some of this research inherits the biologically plausible architecture [27, 28].

A multitude of features like color, texture, intensity, orientation etc. can guide the visual attention in an image. The underlying assumption in detection of VARs is that visual attention at different locations is manifested by the difference in features or combinations of features. The features that have high contrast in one region might not have a similarly high contrast in another region. One of the objectives in this thesis is to develop algorithms to combine adaptively the contrasts presented by multiple features.

In most cases, the human visual attention is deployed on objects in a scene, provided there are objects that stand out from the background, unlike in landscape scenes where there are no such objects. Detection of objects that belong to the VAR assumes that candidate objects can be first isolated from an image; object detection is itself a challenging problem, especially in cluttered environments. The second objective here is to

CHAPTER 1. INTRODUCTION

effectively detect image regions that correspond to semantically meaningful objects and assign a measure of relevance to the region.

The third objective is to ascertain as to what constitutes a visual attention region in videos and to develop algorithms to extract VARs in image sequences. Given the temporal information in videos, it is imperative that VAR detection in videos is not perceived as merely extracting VAR from individual frames. The temporal information is likely to contain cues as to which area of the frames in a sequence attracts visual attention.

1.3.1 Main Contributions

The main contributions of the thesis are as follows:

- We propose a feature selection strategy that selects the appropriate feature that contributes to visual attention in a particular region of the image. The compactness and the density of the points contributing to the feature contrast in a particular feature map are calculated from a convex hull analysis. These parameters determine the corresponding weights of individual feature maps that are combined to form saliency maps.
- A novel feature combination technique is proposed to combine the feature maps into the saliency map by developing a context suppression model in which spurious attention patches are suppressed and the true attention patches are retained. The eigen values of an Attention Cue Covariance Matrix derived from feature contrast vectors over an attention patch is shown to encapsulate the extent of similarity or dissimilarity among feature vectors.
- A polar representation of an image is proposed in which homogeneous features cause image regions to be mapped into linear subspaces that also encode the spatial

CHAPTER 1. INTRODUCTION

information in a region. A new subspace estimation algorithm based on Generalized Principal Component Analysis is developed to estimate multiple linear subspaces.

- A new region based attention measure is defined to calculate the visual attention of each region estimated from the subspaces. It is calculated by accumulating the projection of all data onto individual subspaces. This measure considers both feature-based appearance contrast as well as spatial geometric properties of the region.
- We define attention objects in motion (AOM) in order to capture visual attention regions in videos using 2D trajectories of a set of points. We develop an algorithm to analyze the factorization of a measurement matrix that contains the co-ordinates of the moving points. The structure of the measurement matrix leads to detection of an AOM even if only a subset of points in the AOM is available.
- We extend the notion of AOM to develop a recursive framework to detect multiple moving objects based on a very simple multi-body segmentation algorithm.

1.3.2 Novelty and Originality

The novelty and originality of the research in this thesis lies in the following aspects:

- For saliency map based attention region detection methods, we have proposed a novel feature selection and combination strategy for generating better saliency map. The novelty is that we calculate the spatial compactness of salient points to adaptive combine different feature maps.
- A context suppression model is originally proposed to remove spurious saliency according to the central-surround suppression mechanism. Such suppression mechanism does not exist in any existing saliency map based method.

CHAPTER 1. INTRODUCTION

- Different from existing object-based attention region detection methods, we for the first time convert the problem of visual attention region detection into the problem of subspace analysis by applying polar transformation onto image data and discovering both the subspace and cluster constraints. A novel extension of GPCA was also proposed to robustly detect the cluster segment in linear subspaces.
- A new subspace attention was proposed to measure saliency of image region based on subspace representation. This measure is beyond other contrast-based attention measures in two-fold: (i) attention is measured on the region level; (ii) Both feature contrast and geometric properties of region are analyzed.
- Despite the heuristic definitions of salient motion, we proposed a clear problem definition for detecting the subject attention of cameraman through a special motion pattern. Through the extension of factorization for dynamic scene, we proposed a novel algorithm to detect such special motion pattern without directly estimating either camera or object motion. In the situation of multiple moving objects, the proposed new motion grouping method using iterative voting scheme can extract the motion group more robustly, compared to existing methods.

1.4 Dissertation Organization

The remainder of this dissertation is organized as follows.

In chapter 2, we briefly introduce the related work about visual attention region detection in the literature. Both biological plausible and computational models of bottom-up visual attention are reviewed. We also discuss top-down modulation and learning methods of visual attention as well as the applications of VAR detection.

In chapter 3, we discuss the strategy of combining the feature contrasts of multiple features for VAR detection. Two novel methods, Composite Saliency Indicator (CSI)

CHAPTER 1. INTRODUCTION

for adaptive selection of feature contrasts and Spatial Context Suppression Model for adaptive combination of feature contrasts are introduced. Benefiting from the proposed methods, we also implement a prototype image adaptation system by combining color, intensity as well as new texture contrast for VAR detection.

In chapter 4, we investigate the region-based method for detecting visual attention region. We describe a subspace analysis framework for VAR detection. It includes a polar transformation to map the image regions into subspaces, a new method *NN-GPCA* to robustly estimate multiple subspaces without segmenting data, and a novel subspace-based attention measure to calculate region attention from subspace representation by considering both feature contrast and region geometry.

In chapter 5, we target at detecting visual attention object from image sequence, from a more important cue, i.e., motion. We first define attention objects in motion (AOM) as those that are tracked by the camera by following an important concept in film production, viz., the tracking shot. By analyzing the factorization of AOM, a special structure attained in the motion matrix of AOM is discovered and we introduce a method to identify it. In the case of multiple moving objects, a recursive framework which can extract more than one AOM is presented. We report the experiment result on both synthetic data and real image sequence.

Finally, we conclude this dissertation and point out the future research direction and issues in chapter 6.

Chapter 2

Literature Survey

In this chapter we provide an in-depth introduction of visual attention and review the computational method for visual attention detection as well as its applications in both multimedia and vision communities. Through the analysis of both advantages and disadvantages of the existing methods for visual attention region detection, we find that detecting visual attention region in images is an important but immature problem for many applications of multimedia retrieval and vision analysis. New computational methods for detecting visual attention region are necessary to detect different types of visual attention and facilitate different applications.

2.1 Computational Models of Visual Attention

Due to the interdisciplinary of visual attention, biologically plausible models have been proposed from neuroscience and psychology communities while computationally plausible models have been proposed from computer vision and multimedia communities. All of these models are designed to mimic the functionality of visual attention mechanism for different applications and purposes. In this section, we briefly review the computational model of visual attention. In the psychological and neuroscience literature, we briefly review the fundamental theory of different types of visual attention and some classic

CHAPTER 2. LITERATURE SURVEY

models which are biologically plausible. In the computer vision and multimedia literature, we categorize the existing methods for visual attention region detecting into three categories: (1) feature contrast based methods; (2) perceptual grouping based methods; (3) other miscellaneous methods. After reviewing bottom-up models for visual attention, the existing methods for top-down modulation of visual attention is also reviewed. Finally, the applications of visual attention method in different vision and multimedia tasks are introduced.

2.1.1 Biologically Plausible Models

From the perspective of psychology and neuroscience research, visual attention models are designed according to the processing of visual system of primate. As introduced before, there are two different approaches to bottom-up visual attention that are related to the *early selection theory* [9]: space-based and object-based. Some classic computational models have been proposed for space-based visual attention and have been used in many applications. Although experimental results already prove the existence of object-based attention mechanism, there are only some conceptual models for this type of visual attention in literature.

Space-based Models

The *Feature Integration Model* [27] proposed by Treisman is the most successful conceptual framework of space-based attention. It is a general framework which is useful for understanding space-based visual attention. The theory states that each feature contributes to perceptual conspicuity according to their own modality characteristics and there exists a central structure that encodes conspicuity over the visual scene [4] for a stimulus. Each feature saliency is integrated into a unique saliency map. Based on this theory, a number of computational models that have a biological plausible structure have been developed.

CHAPTER 2. LITERATURE SURVEY

One popular model is the *Saliency Map Model* proposed by Koch and Ullman [28]. This model suggests a unique saliency map that fuses the saliency of each feature into a central representation of visual attention. The visual attention of every spatial location is encoded in the saliency map and the shifting of focus of attention is controlled by this saliency map. Itti et al. extended the model by placing the saliency map in a biologically plausible framework [1]. In this framework, the contrast maps of individual features are calculated as the "center-surrounding difference" across different levels of feature pyramid. The saliency map is generated from the linear combination of all contrast maps wherein a winner-take-all neural network has only those synaptic interactions active that connect units corresponding to salient points. Based on this saliency map structure, the search mechanism for both overt and covert shifts of visual attention was analyzed by the same authors [29]. In [2], they analyze the combination strategies to combine individual contrast maps to saliency map. A global non-linear amplification method was proposed which boosts those feature maps that contain small number of high activity regions at the expense of similar high activities spread over various locations in the scene. This new combination strategy is shown to be superior than the method that normalizes the features and sums them up or the supervised learning method that learns the combination weights.

There are other space-based visual attention models which are proposed according to different biological or psychological hypothesis. Wolfe proposed a guided search model of visual attention in [30]. In this model, the visual search is guided by the visual attention control and the deployment of visual attention is guided by the preattentive parallel processing of basic visual features (e.g. color, motion, orientation etc.). The selective tuning model proposed by Tsotsos in [31] is another biologically plausible neural network model of visual attention. The space-based visual attention is predicted at the location where the connections to the surrounding location are inhibited. The functionality of the

CHAPTER 2. LITERATURE SURVEY

"Surrounding Inhibition" was introduced to both the space-based and the feature-based visual attention. Other models like the spotlight or zoom lens model of Eriksen et al. [12, 13] and the dynamic routing model of Olshausen et al. [32] introduce different architectures of bottom-up visual attention model which use different biologically plausible methods to combine the saliency of low-level visual features and control the shifts of visual attention.

Object-based Models

The deployment unit of visual attentional selection can be not only the spatial location, but also the discrete objects or group of objects. The latter form of visual attention is object-based attention. Research in object-based attention still has many open questions like what is the relationship between object-based attention and perceptual groupings, how attended objects are individuated in time, and the possibility of attending to simple dynamic motions and events. A review about the state-of-the-art of object-based attention has been conducted in [33]. However, some fundamental conceptual models about object-based visual attention have been developed in recent years.

The mechanism of object-based attention has been addressed by Desimone and Duncan [34] using "Biased Competition Model" and by Duncan [35] using the "Integrated Competition" hypothesis. In both these models, visual attention is deployed onto the discrete objects through the competition between multiple neural systems. Using this biased competition, Linda and Susan [36] use an active model to replicate the object-based attention. Such object-based effect allows the competition of objects to be biased by the knowledge of target. Grossberg [37] analyzed the neural structure of primary visual cortex and designed a model which simulates the feedforward, horizontal and feedback interaction in V1 and V2. The horizontal interactions achieve perceptual grouping of features preattentively and the feedback interaction selectively attend to these groupings.

CHAPTER 2. LITERATURE SURVEY

In [38], Behrmann et al. proposed an object-based visual attention model to simulate the mechanisms that groups together all related features by perceptual grouping and gates these features for later, selective processing. Another object-based attention model is the *selective attention identification model (SAIM)* developed in [39]. In SAIM, a continuous Hopfield network is used to satisfy multiple constraints and map an object as a focus of visual attention.

Combining Space and Object

Space-based visual attention and object-based visual attention are not always treated as mutually exclusive [33]. There are also some biological plausible models as well as underlying theories considering the integration of the two types of visual attention. There exists some correlations between them. For example, some researchers suggested that space-based attention occurs within the complex object [40] and object-based attention only occurs under some spatially attended location [41]. The dynamic interaction of object-based spatial selection with object-independent spatial mechanisms in directing attention is also suggested in [42]. Although the two types of visual attention dynamically interact with each other, it is suggested in [43] that space-based selection of location has priority over object-based selection. Only when the spatial selection is invalid does the object-based selection occur.

2.2 Computational Bottom-up Models

The biologically plausible models utilize the fundamental theories in neuroscience and psychology and their architectures are designed according to neural structure of primates. However, most of such visual attention models are only of theoretical interest and difficult to be applied to real images for multimedia and vision applications because such models focus on the analysis of the mechanism of visual attention, but not the detection of visual

CHAPTER 2. LITERATURE SURVEY

attention, which is the key role of visual attention in other applications. The saliency map model proposed by Itti et al. [1] and its extensions [4, 44] is an exception. It is the most popular biologically plausible visual attention model, but has also been successfully applied in many real-world applications. The difficulties in applying existing models to real images motivate the multimedia and computer vision communities to propose other visual attention models. These models may not have biologically plausible architecture, but fulfill the requirement of detecting attention region. In this section, we review some representative models for detecting attention regions from image or video. We categorize these models into three categories:

- Feature Contrast Based Methods: Different ways of calculating feature contrast around every spatial location to indicate visual attention;
- Perceptual Grouping Based Methods: Visual attention is estimated from groups of features/spatial locations, which is grouped together by perceptual grouping/segmentation;
- Other Methods: includes other visual attention methods not belonging to the above two categories.

Table 2.1 summarizes some representative bottom-up visual attention models. We can find the common limitations of both contrast-based method and region/object based method from the table which motivate the research in this thesis. The limitation of contrast-based method is that models either use single feature or simply combine multiple features by summation, which makes the saliency map not adaptive to different locations in different images. The limitation of region/object based methods is that the attention is measured in the local neighborhood although integrated using region information which is not scale-invariant. The detail review of existing computational visual attention models are included in the rest of this section.

CHAPTER 2. LITERATURE SURVEY

Table 2.1: Comparative summary of representative computational models of bottom-up visual attention.

Type	Method	Biological Plausible	Used Features	Adaptive	Scale Invariant	Attention Scope
Contrasts-based Methods	Itti's Models [1, 45, 2]	Yes	Color Intensity Orientation	No	Partial	Local Neighborhood
	MSRA's Model [3]	No	Color	No	No	Local Neighborhood
	Lindeberg's Methods[46, 47]	No	Intensity	No	Partial	Local Neighborhood
	Kadir's Methods [48]	No	Spatial/Scale Entropy	No	No	Local Neighborhood
	Stentiford's Mmodels[49, 50]	No	Local Configuration	No	No	Local Neighborhood
	Sun's Method [11]	Yes	Color Intensity	No	Yes	Region Object
Object-based Methods	Kim's Method [51]	No	Color	No	No	Central Object
	Kwak's Method [52]	No	Luminance Color Orientation	No	No	Local Neighborhood
	Liu's Method [53]	No	Color Intensity Orientation	No	Partial	Local Neighborhood
	Park's Method [54]	No	Edge Color	No	No	Local Neighborhood

2.2.1 Feature Contrast based Methods

Visual attention or saliency is commonly indicated by contrast of visual features. Most of the visual attention models, especially space-based attention models calculate different forms of feature contrast to estimate the saliency at different spatial location. Besides the "center-surround" difference in [1, 45, 2] which calculated the difference of the features between several pre-defined pairs of fine and coarse scales of gaussian pyramid, other computational visual attention models proposed in the multimedia and computer vision communities proposed different methods to calculate visual contrast for attention.

One category of methods is to analyze the feature contrast by scale-space analysis. For example, the scale-space primal sketch proposed by Lindeberg in [46] explicitly modeled the features in gaussian scale-space as well as the relations between structures at different scales. The four-dimensional volume of the scale-space blob, which considers not only the spatial property, feature contrast, but also the lifetime of the blob, is used to measure the saliency of blob structure. Gilles [55] defined saliency in terms of local signal complexity or uncertainty which can be calculated from Shannon entropy of the distribution of local descriptors. However, because this method picks only single salient points in the entropy space, the local maximum entropy from frame to frame could be the result of noise. Another problem of this method is that the local scale is a pre-selected fixed parameter. To overcome these shortcomings, Kadir and Brady [48] extend the method in [55] to a multiscale algorithm. They analyzed the local extrema in scale-space similar to [46, 47] in which the authors use the statistics of the local descriptor over a range of scales around the peak to measure the degree of self-similarity. This novel method works over both feature space and scale space for estimating visual saliency. Based on the information gain, Martin [56] proposed a saliency measure with respect to the spatial location and scales of the image content. The feature contrast is calculated from

CHAPTER 2. LITERATURE SURVEY

the Kullback difference between successive resolution lengths and the saliency map is generated from decomposing information gain into spatial locations.

Other forms of visual attention measure were also proposed for space-based attention. Ma and Zhang [3] simply calculated the color difference between central block and surrounding blocks in HSV color space and used it as the saliency value of every spatial location. A fuzzy-growing method is designed to extract visual attention region from the saliency map. Generalized from [48], an affine invariant salient region detector is introduced in [57]. The salient regions, which are unpredictable in both feature and scale space, can be detected invariant to viewpoint change, image perturbation as well as intra-class variance. Stentiford et al. [49, 50] used competitive novelty to estimate the visual saliency at every location. Instead of using visual contrast, the saliency of a location is decided by the number of other locations with the similar neighboring configuration. Less number of matches indicate high visual attention.

The existing feature contrast based methods proposed different attention measures calculated from feature contrasts to generate central saliency map. However, two adaptive problems: how to adaptively select different feature contrast and how to adaptively combine multiple feature contrasts at different locations of different images have not been carefully analyzed. Such adaptive abilities can generate a final saliency map which better indicate visual attention region(s). The research in Chapter 3 analyzed these problems and proposed a Composite Saliency Indicator as well as a Context Suppression Model for adaptive selection and combination of feature contrasts for VAR detection.

2.2.2 Region and Object based Methods

In the previous sub-section we have seen that different mechanisms have been proposed to calculate visual saliency of every spatial location in an image and that a central saliency map can be generated to indicate visual attention. However, the attention

CHAPTER 2. LITERATURE SURVEY

of only a spatial location is well-defined in feature contrast based methods. There is no definition of visual saliency at the region/object level and extracting such visual attention region/object directly from the generated saliency map is not trivial without grouping information. For most of feature-contrast methods, the scale of attention unit is either pre-defined heuristically or decided by using scale-space analysis. Such local scale information cannot make these attention models invariant to object scale. Visual attention region/object of inappropriate scale will be ignored (e.g. object of small scale) or detected incompletely (e.g. only edge for the object of large scale). Motivated by these open issues, there are some attention detection methods proposed to measure visual attention of region/object.

To calculate the visual saliency in the region/object level, the most direct method is to first segment the regions/objects and estimate the saliency of every region/object according to the characteristics of each group. Sun and Fisher [11] proposed such a model for detecting visual attention objects. In their method, the objects are extracted by segmentation on some primary features. Both the cooperation among the components within the same groups and the competition between a group and other surrounding groups are integrated to estimate the group saliency. The hierarchical structure of grouping is also utilized for hierarchical attention selection in this method. However, this method of detecting salient objects is limited by the performance of the segmentation problem. There are some methods proposed to overcome this limitation. Park and Moon [54] addressed the problem of how to integrate feature maps to better indicate salient object. They integrated the edge and color information into a reference map which emphasizes the salient object boundary while suppressing background. The eliminated object boundaries are further recovered by re-combining the reference map with individual feature maps. The salient object is extracted from this combination map by segmenting the convex hull candidates. Kim et al. [51] defined the salient object according to some prior knowledge.

CHAPTER 2. LITERATURE SURVEY

The object near the central part of the image with significant color (which is the more frequently co-occurred color near the center) is taken as salient. An iterative growing and merging algorithm is used to extract the salient object. Similar to [54], the method in [52] detects the attention window of flexible size from the salient points by integrating three contrast maps: luminance map, color map and orientation map. Liu and Michael [53] enhanced the saliency map with region information and provide a scale-invariant saliency to segment the salient region. The scale-invariant saliency map is generated from a contrast pyramid of multiscale. The region information is used to average the saliency value in every region. This method tries to combine the advantages of saliency and region information, although the detection result of both of them may not be perfect.

The existing region/object based attention models have to rely on the segmentation to obtain the region/object information. The region/object attention is simply calculated by integrating all the saliency values belonging to the same region/object. Such simple attention measure only calculates local saliency and cannot indicate the difference between region/object well. The research in Chapter 4 proposed a novel subspace analysis framework for this problem. Using the proposed method, we can obtain the subspace cluster corresponding to region/object without actually segmenting them. The region/object attention is then measured by considering both feature contrast and geometric properties of region/object between different subspace clusters, which is a global attention measure.

2.2.3 Other Methods

There are some other methods which propose to detect salient point/region/object from other data characteristics instead of contrast. One idea which has been introduced before is to define saliency as complexity. Several saliency measures have been proposed in this context. Besides [48, 57], Lowe [58] calculated the difference of Gaussian for intensity

CHAPTER 2. LITERATURE SURVEY

variation to measure the interestingness. Another method is proposed by Sebe and Lew in [59] which use the absolute value of wavelet coefficients to detect salient points.

More recently, some researchers tried to redefine visual saliency from the view of information theory [60, 61]. Rooted from the problem of visual recognition, Gao and Vasconcelos [62] equated the visual saliency to discrimination. In their definition of *discriminant saliency*, the salient attributes are those that most distinguish it from all other visual concepts that may be of possible interest. This discrimination criterion integrates different type of salient features into a single framework and is translated to the maximization of mutual information between features and class labels as the principle for salient feature selection. Similarly, the information maximization criterion was used by Renninger et al. [63] for programming eye movement. In their model, the next fixation is selected at the location that minimizes uncertainty. Instead of using Shannon entropy on a local neighborhood like [63], Neil et al. designed a saliency measure based on self-information of local content [64]. Compared to entropy which measures the local activity, the self-information is more related to the local contrast such that it is more proper to estimate saliency. This model is implemented based on a neural circuit resembling circuitry involved in early visual processing.

There are some other works in computational modeling of visual attention which approach this problem from some new directions. For example, Harel et al. proposed the idea of *Graph-based Visual Saliency* in [65]. This model is simple but biologically plausible for parallelization. A markov chain is modeled from a full-connected directed graph with the edge weight representing the dissimilarity between visual feature of two locations. The equilibrium distribution of this chain is used to indicate saliency according to the idea of random walk. The idea of using random walk to model saliency was also proposed in [66, 67]. Itti and Baldi [44] propose a Bayesian measure for "surprise". This surprise is estimated by calculating the difference between the posterior after observing

the data and the prior beliefs. Such framework has been shown to be consistent with the gaze of human eyes and can be extended to different spatial-temporal scales, modalities, and levels of abstraction.

There has been research on detecting salient contours from images. However, they are more related to perceptual grouping and are concerned about the saliency of edge/contour instead of pixel/region. Some interesting work includes the neural network for detecting structural saliency proposed in [68] and the surround suppression of texture edges for salient contour detection proposed by [69]. The graph cut based algorithms are also utilized to detect salient boundary like [70, 71, 72].

2.3 Top-down Modulation & Learning

We now introduce some representative work about another aspect of visual attention, namely, top-down modulation and learning of visual attention in this section. It has been experimentally discovered in psychology and neuroscience that visual attention can be modulated by the relevant task or the specific context. Another interesting question is to learn how to detect visual attention location/region/object from training data. Introducing machine learning methods for visual attention detection is reasonable because visual attention detection is a complex process to code manually. Learning the knowledge from data can benefit the algorithm. But the design of computational methods to modulate visual attention detection according to top-down knowledge is still an open issue.

2.3.1 Top-down Modulation

The top-down modulation of visual attention is to use top-down knowledge to constrain the deployment of visual attention. Different types of high-level knowledge can be utilized for such modulation. One type of knowledge is about the object of interest. Schill et

CHAPTER 2. LITERATURE SURVEY

al. [73] learned a hierarchical knowledge tree of object through training to gain the knowledge of scene/objects. The spatial shift of visual attention is then directed towards the location that maximizes the gain of information about the object. Rybak et al. [74] proposed to learn the scanpaths of the object to be recognized. In their model, the candidate scanpath are used to modulate the shifting of attention when recognizing the specific object.

Oliva and Torralba [75, 76, 77, 78, 79] utilized another type of knowledge for modulating visual attention - the **context**. They tried to model how the scene context information modulates the saliency of image regions to facilitate object detection and recognition. As the first step of their model, the global scene/context information is obtained by the proposed spatial envelope model [75, 76]. Without segmentation and specific object recognition, the semantic category of a scene is embedded in the space of spatial envelope model. With the context information, they continued to model the relationship between the context and the object in it. The probability of the object appearing in the scene is estimated from the correlation between the statistics of low-level features across the entire scene and the object properties. Not only the existence but also the scale and location of the object can be inferred. By differentiating the image features into local features and contextual features, the bottom-up saliency can be defined as the inverse of probability of finding local features within the context, and this bottom-up saliency is modulated by the probability of the object appearing at the specific location within the context.

The limitation of current top-down attention models lies in the heuristic definitions about how to decide attention from the top-down knowledge. Those heuristic top-down attention are not generally applicable for different types of images/video. The research in Chapter 5 targets at detecting a specific form of top-down attention - the subjective attention of cameraman, which is generally applicable. Specifically, we discover the top-down attention by detecting the motion pattern between and camera. The proposed

solution for detecting Attention Object in Motion (AOM) based on factorization can detect the motion pattern of following only from partial data without estimating the motion of either camera or object.

2.3.2 Learning for Visual Attention Detection

Detection of visual attention is such a complex process that directly designing detection algorithm may not be reasonable. Making the system learn from training examples automatically is another possible solution for visual attention detection. For example, Moonmann et al. proposed to learn saliency map for object categorization [80]. Similar to [80], the saliency of each subwindow is defined as the likelihood of finding an object of interest in it and forms a prior knowledge. Such saliency map is represented as a probability density function of subwindow and the performance of object categorization is improved by biasing the sampling of subwindows according to such saliency map. Instead of generating saliency map offline, the authors designed a simple iterative update method to generate saliency map online. Recently, Liu et al. [81] proposed a set of new features to represent salient object in different levels: multi-scale contrast locally, center-surround histogram regionally and color spatial distribution globally. To combine these feature for salient detection, they learned a conditional random field effectively. The optimal combination of features is learned from the labeling information about salient object of multiple users in a large image database. Gao and Vasconcelos proposed a novel procedure to integrate the salient region detection and object detection in [82] based on their discriminant saliency [62]. Starting from the simple features, the saliency map is iteratively learned from the discriminant feature selection. In each iteration, more complex feature set are generated by the linear combination of simple features, which is optimal for discriminating objects and form a feature hierarchy. For different object class, different level of feature complexity are selected according to the variability of object class.

2.4 Applications of Visual Attention Detection

Given the salient location/region/object detected from visual attention detection, different challenging problems in computer vision and multimedia management can benefit from the attention information. In the last part of literature survey, we review different applications which utilizes the result of visual attention detection. We categorize the potential applications proposed in the literature into three parts: (i) using visual attention detection to improve other computer vision problems; (ii) using visual attention detection to improve multimedia organization, retrieval and management; (iii) using visual attention detection to adapt multimedia for human computer interaction.

2.4.1 Computer Vision Applications

In computer vision tasks, visual attention detection has been shown to be useful for several problems. For object recognition, it has been shown that pure bottom-up visual attention can enhance object recognition in a clutter environment [83, 84, 85, 86]. Rutishauser et al. [85, 86] exploited the usefulness of visual attention for object recognition. They designed a method to extract salient region from Itti's saliency map [1]. The recognition algorithm [58, 87] based on the matching of SIFT descriptor only considers the keypoints in the salient regions. The experiments has demonstrated that attentional region selection can dramatically improve the recognition performance. Navalpakkam and Itti [88, 89] learned the representation of object from attentional locations and combined them into the object model. By recursively applying this process, a hierarchy of object model can be built from simple object to complex object. The object detection and recognition can benefit from the learned object model by applying top-down biasing on the attention deployment to make the object more salient.

Visual tracking is another important vision task. Visual attention region detection can naturally facilitate visual tracking problem because of their similar issues: 1) select

a salient (tracking) target from background; 2) transfer the attention information to a higher processing stage; 3) shift attention (tracking) to a new location. Brajovic and Kanade [23] focused on the hardware aspect of attentional visual tracking. They designed a VLSI computational sensor that focuses on the target of interest and suppresses the irrelevant information, thereby achieving high computational efficiency can be achieved. Toyama and Hager [24] proposed the idea of *Incremental Focus of Attention (IFA)* to combine visual attention detection with tracking algorithm for robust, real-time tracking. In this hierarchical architecture, the visual attention detection is conducted to re-locate the lost target and continue tracking. The latest work of Yang et al. [90] also targeted the problem of attentional visual tracking. The proposed method employed a two-step spatial selection of image regions for tracking: a pool of salient regions are selected in a bottom-up way in the early selection process and a set of discriminative attentional regions are identified by online discriminative learning.

Some other areas that integrate visual attention detection methods are object detection [75, 91], scene understanding [92, 93], facial recognition [67] and surprising event detection in video [94].

2.4.2 Multimedia Information Management

Multimedia information management consists of indexing, summarization and retrieval of multimedia content. Region-based image retrieval methods can benefit from salient region/object detection. The information about visual attention region/object can improve the image representation by either removing the background of non-salient regions [95, 51, 96] or mapping the saliency value to the important factor to emphasize the main subject of the image [21, 22, 97]. Such attention-based representation improves the robustness of image retrieval in a cluttered environment.

For the organization problem of multimedia content, Boccignone et al. [25] proposed to detect both abrupt and gradual transitions between shots in video based on the

CHAPTER 2. LITERATURE SURVEY

foveated representation of video. The shot boundary detection is related to the consistency of the fixation sequences. Such measure is calculated as the average of the local consistency between pairs of Focus-of-Attention, iteratively selected from the foveated sequences by prediction strategy. Bayesian inference is used to detect the shot-change given the change of foveated consistency. Ma et al. [98] designed a generic attention framework for video summarization. In this framework, not only bottom-up attention of color, motion etc., but also top-down knowledge and other modalities are combined together. According to the attention measure, the video content can be ranked based on their importance and the ranking provides clues for summarization.

For multimedia compression and transmission, JPEG2000 can take the detected salient region(s) as Region-of-Interest (ROI) for progressive coding [99, 100]. By identifying the visual attention regions, the progressive transmission scheme of image is proposed in [20], where attentional content are transmitted in higher resolution than others.

2.4.3 Multimedia Adaptation

Universal Multimedia Access (UMA) and Universal Multimedia Experience (UME) aim to let users consume multimedia content anytime and anywhere. However, the heterogeneities and limitations of different devices prevent the success of such pervasive computing and require the multimedia content to be adapted for different devices, and also for different users. Different types of adaptation algorithms have been proposed based on visual attention detection to reduce the content redundancy and fit the limitation of resource.

To adapt image on the mobile devices of small screen-factor, Chen et al. [18] designed a visual attention model and utilized it to preserve only the visually attentive content to convey the maximum information on the limited space. Liu et al. [19] on the other hand proposed another solution for viewing images on the small screen. In their method, a

CHAPTER 2. LITERATURE SURVEY

browsing path is generated automatically to optimally cover multiple attention regions that are detected by visual attention model. It is an alternate way to maximize the information in spatial-temporal 2D space. Automatic cropping of an image based on visual attention can enable applications such as snapshot generation, adaptive documents and thumbnailing. In [101], the authors formulated cropping as an optimization problem and use eye tracking data to infer the optimal crop for any given aspect ratio or size.

Beyond adapting single image, some work tried to investigate the arrangement of a group of images by utilizing visual attention information. One example application called *PictureCollage* [102] focused on the arrangement of a collection of images according to their attention regions. To maximize the visible information, especially the important information indicated by the salient regions, a Maximum a Posterior (MAP) problem is formulated such that as many salient regions as possible can be visible on a fixed size of canvas. Another interesting application is *Tilling Slideshow* proposed by Chen et al. [103]. In this application, the attention regions are the common region in a group of similar images. Given this contextual attention information as well as displaying space constraint, the photo-slide consisting of large topic photo as well as small supportive photos are organized by constraint optimization.

Reorganization of video content for mobile devices, and video streaming under different network conditions are topics related to video adaptation. Fan et al. [104] generated automatic zooming into the visual attention region(s) for displaying video on a small screen. Recently, Liu and Gleicher [105] analyzed the problem of *Video Retargeting* to adapt video to better suit the target display, while minimizing the important information lost. Both scaling and cropping are optimized to minimize information loss by balancing the loss of two operations. Also the virtual pans and scans are generated by moving the cropping windows to ensure the cinematic plausibility.

Chapter 3

Adaptive Feature Selection and Combination

3.1 Introduction

Feature-based visual attention is based on the fact that contrast between features is a good clue for detecting visual attention regions. For example, the saliency based visual attention models [1, 50, 3] assume that there is a central saliency map which encodes the bottom-up visual attention on every spatial location. The saliency map is generated from the contrast of low-level features. The contrast of different features such as color, orientation, shape, motion etc. are the candidates that draw VA to a particular part of a scene. For feature-based VAR detection, it is important to determine how to combine the contrasts of multiple features in addition to defining and calculating individual feature contrast.

Low level features like color and intensity, by themselves, are not useful in VAR detection because a strong local saliency around a certain region indicated by one feature does not imply the same for another feature. It is only when the features are combined in an appropriate manner that the resultant saliency map captures the attention regions. However, the optimal solution to the problem of combining features is hindered by the fact that the features have different dynamic ranges and are heterogeneous by definition

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

(else they would not be good features). Some work about the combination strategy of feature-based visual attention have been reported earlier. Itti et al. propose and compare several combination strategies in [2]. The simplest but most ineffective way to combine the features is to normalize them to the same range and to sum them up. A weighted feature combination method uses supervised learning to determine the weights for specific targets and hence, this method cannot be generalized. The global non-linear amplification method boosts those feature maps that contain small number of high activity regions at the expense of similar high activities spread over various locations in the scene. One cannot guarantee that the latter condition implies absence of saliency. The iterative localized interaction method simulates local competition between neighboring salient locations. However, the elimination of isolated but spurious VARs is not guaranteed. In [106], the authors combine features related to color, texture and wavelet coefficients into a saliency map by optimizing a cost function to calculate the weights of each feature. All of these methods only find a global solution to combine multiple feature contrast. However, it is observed that at any location, different features contribute differently to visual attention.

In this chapter, we present two methods to adaptively select and combine multiple feature contrasts for VAR detection. The first method uses a novel Composite Saliency Indicator (CSI) to determine the contribution of each feature map to the salient region. CSI is designed to capture the spatial compactness as well as the density of candidate regions in the feature maps. In the second method, a measure to determine similarity among candidate attention regions in a feature map is devised. This measure is used to find a suppression factor for a particular patch in the scene, which is then used to highlight actual attention regions. An image adaptation prototype system that benefits from the improved VAR detection using the proposed method is also demonstrated.

The rest of this chapter is organized as follows. Section 3.2 describes the spatial compactness and saliency density features that are combined to obtain the CSI. In section

3.3, a new texture attention cue is described and a feature combination strategy through context suppression is proposed. Section 3.4 presents the experimental results on a wide variety of images. Finally, conclusions are presented in section 3.5.

3.2 Composite Saliency Indicator (CSI)

We follow the model in [1] to generate the three feature maps corresponding to color, intensity and orientation. Clearly, the contribution of each feature map to a potential salient point is different. When human subjects capture visual attention region, they prefer salient regions with high compactness. If strong salient points occur in a small area compared to the total size of the image, the saliency in this area can be said to be compact and distinct. Inspired by this observation of biological behavior of human subject, the CSI is used to measure the contribution of each feature map to the salient region. The measure consists of two factors - Spatial Compactness and Saliency Density. In the following subsections we describe each of them in more detail.

3.2.1 Salient Point Selection

The first step is to detect the salient points. This is obtained by simply thresholding each individual feature map. Any thresholding methods can be used. Because we analyze the histogram of saliency value in this paper, the value of the threshold is decided by histogram entropy thresholding analysis [107] which optimally chooses thresholding value for separating a histogram. Accordingly, the threshold is obtained by maximizing

$$l' = \arg \max_l \left(- \sum_{\mu=1}^l \frac{p_{\mu}}{\sum_{v=1}^l p_v} \log \frac{p_{\mu}}{\sum_{v=1}^l p_v} - \sum_{\mu=l+1}^L \frac{p_{\mu}}{1 - \sum_{v=1}^l p_v} \log \frac{p_{\mu}}{1 - \sum_{v=1}^l p_v} \right) \quad (\text{Eq. 3.1})$$

where p_i is the number of pixels with intensity i , L is the total number of gray levels and l is the threshold. The salient point set is defined as the set of pixels whose value is above the threshold.

3.2.2 Spatial Compactness

The spatial compactness of salient point set indicates the conspicuousness of potential salient region. Because the salient point set obtained from feature points are not connected to form salient region, the Connected Component Analysis cannot be applied to approximate the covered region of salient points. In our work, we use convex hull to measure spatial compactness of salient point set. The procedure consists of two steps:

- (i) Compute convex hull polygon of salient point set using "Gift Wrapping" algorithm [108]. The algorithm begins by locating the lowest-rightmost point and then finds the point that has the smallest positive angle (with respect to the horizontal axis). A hull edge is found that joins these two points. The algorithm then proceeds to find the point with the smallest angle from this established hull edge in a counter-clockwise sense. The process continues until the lowest-rightmost point is reached again.
- (ii) Calculate the area of the polygon using the trapezoid method. We first identify the left-most vertex A of the polygon and then rearrange the vertices starting from A in a clockwise sense. The area is calculated as,

$$Size_{convexhull} = \sum_{k=1}^{K+1} \frac{(y_{k+1} + y_k) \times (x_{k+1} - x_k)}{2} \quad (\text{Eq. 3.2})$$

where (x_i, y_i) are co-ordinates of the vertex and K is the total number of vertices in the convex hull.

Figure 3.1 shows an example for determining the convex hull. For images with multiple salient regions, the above procedure can be applied to each connected salient point subset and summed up over all convex hulls to get a measure of spatial compactness.

Figure 3.1: Example of a convex hull to determine the spatial compactness measure; (a) Original Image; (b) Intensity Feature Map; (c) Salient Point Set; (d) Convex Hull.

3.2.3 Saliency Density

The second part of the CSI is the saliency density indicated by the gray-level value of the points. Spatial compactness only considers the spatial relationship of salient points. Two salient point sets with the same spatial compactness can have different effects for salient region indication. The feature map with strong saliency density in one or few specific areas is a stronger cue for extracting salient region. We calculate saliency density as

$$D_{saliency} = \frac{\sum_{p \in \theta} \frac{\sum_{q \in \theta_n(p)} |I(p) - I(q)|}{|\theta_n(p)|}}{|\theta|}, \quad (\text{Eq. 3.3})$$

where $I(x)$ is the intensity at location x , $\theta_n(p)$ is the set of all salient points in the neighborhood of p and θ is the set of salient points. If all salient points with similar saliency value are close to each other, $D_{saliency}$ will give a small value indicating that the saliency of this map is conspicuous. A large value of $D_{saliency}$ implies inconspicuity of saliency. Note that the saliency density is measured using intensity values from the feature map but only for those locations in the salient point set.

3.2.4 Feature Combination Strategy

Based on the saliency measures $Size_{convexhull}$ and $D_{saliency}$, a two level feature combination strategy is designed for feature map combination. The advantage of the proposed combination strategy lies in that it dynamically decides whether a feature will be selected, and if so, what will be its weight.

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

The feature maps are first classified into two categories according to its spatial compactness. If $Size_{convexhull} < 80\%$ of the feature map area, we call it a *Non-uniform Map*, else it is called a *Uniform Map*. Here, the thresholding value of 80% is empirically decided according to the prior knowledge about the usefulness of a feature map. In a *Uniform Map*, the saliency is not sparse implying that it does not contain much useful information with regard to human visual attention; hence, a uniform map is not considered while determining the saliency map. Among non-uniform feature maps, the feature map with smallest $Size_{convexhull}$ is selected as the *Reference Map* (RM). Then the similarity between each non-uniform map and the *Reference Map* is examined. The similarity measure used is

$$d(f_i, f_j) = \left(\sum_{n=1}^N \left| \frac{1}{M} (l_n^r(f_i) - l_n^r(f_j)) \right|^2 + \sum_{m=1}^M \left| \frac{1}{N} (l_m^c(f_i) - l_m^c(f_j)) \right|^2 \right)^{\frac{1}{2}} \quad (\text{Eq. 3.4})$$

where l_n^r and l_m^c are the projection vectors of the $M \times N$ feature map f for the n^{th} row and m^{th} column, respectively, i.e. $l_n^r(f) = \sum_{m=1}^M f(m, n)$ and $l_m^c(f) = \sum_{n=1}^N f(m, n)$. Based on this similarity measure, non-uniform maps are further divided into two categories - if $d(f_i, f_{RM}) < T$, where $T = (M + N)/2$ is a heuristic threshold chosen as the middle between minimum distance (0) and maximum distance (M+N) between two feature maps, then we call the non-uniform map as *Related Map*, else it is called an *Unrelated Map*. Unrelated maps are ignored in the combination. In the second level, the related maps are linearly combined with the reference map. The weights for the maps are calculated using spatial compactness and saliency density according to

$$W_i = \frac{\frac{W_{total}}{Size_{convexhull}(i) \times D_{saliency}(i)}}{\sqrt{\sum_{p \in \Phi} \left(\frac{W_{total}}{Size_{convexhull}(p) \times D_{saliency}(p)} \right)^2}} \quad (\text{Eq. 3.5})$$

where

$$W_{total} = \sum_{p \in \Phi} (Size_{convexhull}(p) \times D_{saliency}(p)) \quad (\text{Eq. 3.6})$$

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

Table 3.1: Rules for Feature Map Combination

Rule Number	Concept	Condition	Combination Strategy
1	Uniform Map	$Size_{convexhull} > 80\%$ of map area	Skip during combination
2	Reference Map	$\min \{Size_{convexhull}\}$	Weighted combination
3	Related Map	$d(f_i, f_{RM}) < T$	Weighted combination
4	Unrelated Map	$d(f_i, f_{RM}) \geq T$	Skip during combination

and ϕ is the set of all *Related Maps*. Table 3.1 summarizes the rules of the proposed feature combination strategy.

In the unlikely event that all maps are classified as *Uniform Map*, this method implies that none of the features is good enough to clearly indicate a visual attention region. In such instances, VAR can be determined by considering the interaction among multiple features, which is the second method for VAR detection described in this chapter. Figure 3.2 compares the result of the proposed combination strategy with those in [1] and in [2]. Since the salient point sets corresponding to the orientation map and the intensity map do not satisfy Rule 1 for feature map combination, they are ignored. As a result, the detected salient region is derived solely from the color feature map. We observe that the salient region is more compact and closer to the human visual system. The visual attention region (VAR) is extracted by thresholding the saliency map. The proposed weighting scheme is further illustrated in Figure 3.3. The *Reference Map* and *Related Maps* are different for different images and so are the corresponding weights for each map. For instance, in Figure 3.3 (d), the uniform intensity map is ignored while the color map is shown as the *Related Map* and the orientation map is chosen as the *Reference Map* and their weights are calculated from Equation (Eq. 3.5). After combining the feature maps and generating a global saliency map, any region extraction method can be used to extract the visual attention region. In the examples of Figure 3.2, we use the Otsu thresholding method [109] which optimally separates continue saliency value into binary

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

Figure 3.2: Feature combination using (a) CSI (b) global non-linear combination of [1] and (c) iterative local competition [2].

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

Figure 3.3: Examples of feature combination using CSI.

Figure 3.4: Use of CSI on the visual attention model of [3]; (a) Original image, Saliency map generated (b) without CSI and (c) its thresholded VAR; (d) using CSI and (e) its thresholded VAR.

e.g. salient and non-salient to obtain VAR for all three combination strategies. The similar result can be obtained by applying other thresholding methods e.g. [107] because our method is independent to the thresholding algorithm. Note that the proposed CSI does not assume that only one VAR exists in each image. For images containing multiple VARs in different features, CSI generates the final saliency map by combining all related feature maps which is useful for detecting *any* VAR. The advantage of the proposed CSI in this case is to refine the saliency map to better represent *all* attention regions.

Although the feature combination method described above uses the model described in [1] to generate the feature maps, we note that it is independent of the mechanism by which the feature maps are generated. The model for salient region detection in [3] simply considers the difference in feature values between a block of pixels and its neighbors. We consider three features, viz., HSV color, intensity, and a new texture feature (to be described in section 3.3.1) and generate feature maps using the spatial neighborhood contrast method of [3]. These feature maps are then combined using the proposed CSI algorithm and the VAR extracted as shown in Figure 3.4 (d) and (e). We see that the VAR generated by the simple procedure of normalization followed by

summation of the feature maps as shown in Figure 3.4 (b) and (c), is unsatisfactory.

3.3 Spatial Context Suppression

Next, we describe the second method for combining feature maps to detect VARs. The neurons in the visual receptive field are subjected to both self-excitatory and neighbor-induced inhibitory inputs. The resultant stimulus causes focus of attention areas in the visual field to be identified. Therefore, in dealing with feature maps to generate a saliency map, we must also consider contextual information in addition to the contrast information in the features. Contextual information can be employed to suppress spurious attention points in feature maps. Such a mechanism is used in [2] where the suppression is carried out on each feature map individually. But the interaction between multiple features has not been considered previously. In this section, we propose a context suppression scheme, which takes into account information contained across all feature maps simultaneously. First, we describe a new texture attention cue that is suitable for detection of local saliency in scenes with cluttered background.

3.3.1 Texture Attention Cue

Intensity, color and orientation have been used as cues for VA. Here, we describe a method to recover texture information in an image that serves to generate a texture map that facilitates detection of attention regions. Texture is especially useful to capture VA in images containing small objects present in a cluttered background.

An image is divided into blocks, called *texture patches*, each block containing $p \times q$ pixels. By taking the Gabor Wavelet Transform [110] of the image, each texture patch is represented by the mean μ_{sk} and the standard deviation σ_{sk} of the wavelet coefficients, where s and k stand for scale and orientation, respectively. For S scales and K orientations, we then have SK mean maps $MM_{s,k}$ and standard deviation maps

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

$SDM_{s,k}$, $s = 1 \dots S$, $k = 1 \dots K$. Since our objective is to capture the contrast/variation in texture over the image, we calculate the Average Mean Difference (AMD) and the Average Standard Deviation Difference (ASDD) over a neighborhood of patches, where AMD for a patch centered at (i, j) is given by

$$AMD_{s,k}(i, j) = \frac{1}{N_n} \sum_{u,v} |MM_{s,k}(i + u, j + v) - MM_{s,k}(i, j)| \quad (\text{Eq. 3.7})$$

and ASDD for a similar patch is given by

$$ASDD_{s,k}(i, j) = \frac{1}{N_n} \sum_{u,v} |SDM_{s,k}(i + u, j + v) - SDM_{s,k}(i, j)| \quad (\text{Eq. 3.8})$$

with N_n being the number of patches in the neighborhood. A measure for texture contrast at a patch (i, j) and at any scale s and orientation k is calculated as

$$TC_{s,k}(i, j) = AMD_{s,k}(i, j) \times ASDD_{s,k}(i, j) \quad (\text{Eq. 3.9})$$

while the final texture contrast at patch (i, j) is obtained as

$$TC(i, j) = \sum_s \sum_k TC_{s,k}(i, j) \quad (\text{Eq. 3.10})$$

The above texture cue captures a VAR even if other cues like intensity and color fail. There are several situations in which the VAR is captured using texture as an aid, e.g. texture foreground in non-textured background, non-textured foreground in texture background, regular texture in a random textured background, random texture in regular textured background etc. Figure 3.5 shows two examples where texture is useful as a cue to detect salient objects that capture VA. The texture contrast map obtained using the proposed algorithm is shown in Figure 3.5 (b). Also, we show that intensity and color contrast maps generated by the algorithm described in [1] fail to indicate local saliency in Figure 3.5 (c) and (d). We observe that the texture map is able to enhance the salient objects.

Figure 3.5: (a) Original image (b) Texture map (c) Intensity map and (d) Color map using [1].

3.3.2 Feature combination through Context Suppression

One of the problems with the existing combination algorithms and the CSI algorithm described in the previous section is that the contribution of an entire feature map is altered on the basis of salient regions within that map. For example, if the true VAR is highlighted in the intensity map along with spurious VARs, the weight of the entire map is reduced instead of suppressing only the outliers. Motivated by the mechanism of non-classical receptive field inhibition for contour detection [69], we argue that local context information is essential to discriminate between true and spurious VARs. The influence of context can be brought about by suppressing the saliency of a region whose neighborhood has similar contrast, through the process of surround suppression. In Figure 3.6 (a), for example, the VARs due to background bushes and grass need to be suppressed to capture the true VAR of the antelopes as perceived by the human visual system. Here, we propose a local context suppression strategy to adaptively combine multiple attention cues like intensity, color and texture. Consider an image divided into blocks, each containing $p \times q$ pixels. We call each block an *Attention Patch*. The variation or contrast of a particular feature at a particular patch centered at (i, j) with respect to

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

Figure 3.6: (a) Original image and (b) Eigen ellipsoids for each of the patches in (a).

its neighborhood is calculated as

$$FV(i, j) = \frac{1}{N} \sum_{u, v} |MF(i, j) - MF(i + u, j + v)| \quad (\text{Eq. 3.11})$$

where $MF(i, j)$ is the mean of the feature in patch (i, j) and N is the number of patches in its neighborhood. The contrasts at patch (i, j) for n features/attention cues $\{FV_1(i, j), FV_2(i, j), \dots, FV_n(i, j)\}$ are normalized to lie between $[0, 1]$ and collected into a *feature contrast vector*. For those features that themselves have components, e.g. color has hue and saturation, we sum up the contrasts for individual components to get the contrast for the feature. Each patch is now represented by the n dimensional feature contrast vector which is compared with other feature contrast vectors in its neighborhood and its contrast measure (Equation (Eq. 3.11)) is suppressed if the patch and its neighbors are ‘similar’. Similarity is estimated by the variance of data along eigen vectors of an $n \times n$ covariance matrix, called the Attention Cue Covariance Matrix (ACCM). The ACCM is the covariance matrix of the feature contrast vectors over a patch (i, j) and its neighborhood. The eigen values of ACCM represent the extent of similarity or dissimilarity among the attention cues. A large (small) eigen value indicates large (small) variance along the direction of its corresponding eigen vector, which in turn implies higher (lower)

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

discriminating power. It is shown that regions having such higher discriminating power correspond to the true VARs. Consider the image shown in Figure 3.6 (a) and a 3 dimensional feature contrast vector of intensity, color and texture. We examine 3 patches as potentially true VARs: patch A is from the background bushes, patch C is from the grass and patch B is from the true VAR of the antelopes. Figure 3.6 (b) shows eigen ellipsoids corresponding to each patch. For each patch, the ellipsoid is centered at the mean of the feature contrast vectors over the neighborhood of the patch. Each axis of the ellipsoids points towards the eigen vectors of the $ACCM$ while the lengths of semi-axes of the ellipsoid are proportional to the eigen values. As seen from the figure, patch C has small variance along all the axes of the ellipsoid while patch A has small variance along two of its axes. However, patch B has large variance along all the 3 axes indicating higher discriminating ability with respect to its neighborhood and should therefore belong to a true VAR, which indeed it does. Thus, it is required that the contribution to the saliency map of patches A and C should be suppressed while that of patch B should be enhanced. The suppression factor (SF) for patch (i, j) is obtained as

$$\tau(i, j) = \prod_{u=1}^p \bar{\lambda}_u \quad (\text{Eq. 3.12})$$

where the $\bar{\lambda}$'s are the eigen values of the corresponding $ACCM$ and are sorted in ascending order. The parameter p controls the degree of suppression. The saliency value $S(i, j)$ for patch (i, j) is obtained in two steps: first the multiple attention cues or the contrast maps are linearly combined and the result is modulated by the SF as

$$S(i, j) = \tau(i, j) \times \sum_{u=1}^k FV_u(i, j) \quad (\text{Eq. 3.13})$$

Figure 3.7 shows the steps leading to the final saliency map of the image shown in Figure 3.6 (a). The intensity and color contrast maps are obtained using Equation (Eq. 3.11) and the texture contrast map is obtained as described in Section 3.3.1. The

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

linear combination of the contrast maps is implemented simply as their sum and the SF for each patch is displayed as an image with darker regions representing high SF and brighter regions representing low SF. The product of the combined map and the SF yields the final saliency map which contains the true VAR. Note that both the color and texture contrast maps have been able to indicate the true VAR to some extent. However, the former has more spurious salient points than the latter. Moreover, using the proposed SF, these spurious regions have been successfully removed.

Figure 3.7: The process of local context suppression.

The selection of block size and neighborhood connectedness is not critical to the success of this algorithm. In all the different types of images used in the experiment, we use 8×8 block size and 8-connected neighbors. We also need to mention here that the proposed *context suppression model* is independent of how the feature contrasts are calculated. Using the center-surround difference and normalization scheme of [1] to generate the feature maps, the final saliency map and the VAR is obtained as shown in Figure 3.8 (b) and (c), respectively. However, when we impose context suppression on the saliency map, according to the procedure described above, we notice that the spurious saliency points are inhibited (Figure 3.8 (d)) and the VAR is accurately located (Figure 3.8 (e)).

Figure 3.8: Illustration of the usage of Context Suppression Model to combine multiscale feature contrasts of [1] (a) saliency map generated without context suppression and (c) its thresholded visual attention region; (d) saliency map using Context Suppression and (e) its thresholded visual attention region.

3.4 More experiments

In this section, we present additional experiments to compare the proposed methods for detecting VAR with the methods described in [1] and [2]. CSI combination is carried out on feature maps generated by [1], while context suppression is done on feature maps generated through the contrast measure defined in Equation (Eq. 3.11). However, as mentioned in the last paragraph of the previous section, success of the combination strategies is independent of the mechanism by which the feature maps are generated. Determining VARs in images is a highly subjective exercise. Nevertheless, in order to provide some form of objectivity, we propose a measure to quantify the effectiveness of VAR detection algorithms. Clearly, this presumes the availability of VARs, which in this case, we segment out manually based on our judgement of what comprises visual attention for a normal HVS. We call the new measure as the Discrimination Ratio defined as

$$DR = \frac{\bar{I}_S}{\bar{I}_S + \bar{I}_{NS}} \quad (\text{Eq. 3.14})$$

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

where $\bar{I}_S$ is the average saliency value of the salient points and $\bar{I}_{NS}$ is the average saliency value of the non-salient points in the saliency map. Different from the common measure such as calculating the overlapped region of the detected VAR and the manually segmented ground-truth for object detection, the proposed DR avoids the dependency on the VAR extraction and bounding box calculation, which may affect other algorithms e.g. [1] and may not be fair for comparison. DR explicitly utilizes the saliency value to calculate the ratio between the average saliencies of the labeled attention region and non-attention region, which estimates the goodness of the obtained saliency map.

Figure 3.9: Comparison of CSI and [1] (a) Original Image; (b) Saliency map of [1] and (c) its thresholded attention region (d) Saliency map using CSI and (e) its thresholded attention region.

Figure 3.9 shows the results of the proposed CSI combination method for some images taken from the Corel Photo Library. In comparison to the global non-linear optimization method of [1], whose saliency map and VAR region are shown in Figure 3.9 (b) and Figure 3.9 (c), respectively, our method can extract the VAR, shown in Figure 3.9 (e), more accurately from the saliency map shown in Figure 3.9 (d). The DR values for these

images are 0.85487, 0.76092, 0.77373 and 0.85025 using CSI and 0.72788, 0.72557, 0.7068 and 0.8157 using the iterative local non-linear combination method proposed in [2].

Figure 3.10 shows the results of the context suppression method for some other images from the Corel Photo Library. The features used to generate the saliency map are intensity ($I = (R + G + B)/3$), color (hue and saturation in the HSV space) and the proposed texture contrast. The size of an Attention Patch was 8×8 and the Gabor Wavelet Transform used 4 scales and 6 orientations. The parameter p to control SF was chosen as 2. Figure 3.10 (a), (b) and (c) show the original images, the saliency maps and the VARs, respectively, detected by the proposed method. We compare them with those obtained by [2] as shown in Figure 3.10 (d) and (e). Note that the last row shows one of the images from Figure 3.9 to illustrate that the two proposed methods achieve better performance.

We added White Gaussian Noise to the images and compared the mean of the Discrimination Ratios over 600 images. Figure 3.11 shows the plot of the mean DR with increasing variance of noise for each of the two methods proposed in the chapter and the methods proposed in [1] and [2]. The context suppression method has significantly higher mean DR of above 0.9 compared to the others, implying that the saliency value of the non salient regions is close to zero in this case. The CSI method performs better than [1] although compared to [2] there is a slight degradation in performance since it considers only the spatial characteristics of the salient points without considering the contextual information around them. It is interesting to note that all the methods perform consistently with increase in noise variance, the reason of which is that all methods process images in block level instead of pixel level and hence can tolerate gaussian noise.

3.4.1 Evaluation in MPEG-21 DIA Adaptation System

As one of the applications of the proposed VAR detection methods, images can be adapted for devices with different size of screen by cropping and downsizing *only* the VAR. In this

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

Figure 3.10: Comparison of Context Suppression Model with [2] (a) Original Image; saliency maps using (b) proposed context suppression model and (c) the thresholded attention region; (d) Itti's iterative local competition model and (e) the thresholded attention region.

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

Figure 3.11: Mean DR v/s Gaussian Noise for various models.

section, we describe a prototype image adaptation system that integrates VAR detection and the standard media adaptation framework of MPEG-21 DIA [111]. The system resizes an image so that the extracted VAR can be accommodated within the viewing area of a terminal device. We consider images in JPEG2000 format whose scalability is useful for adaptation. The digital item (DI) is comprised of the JPEG2000 image bitstream and its Bitstream Syntax Description (BSD) [112]. BSD describes the high-level structure of JPEG2000 bitstream and adaptation is performed on the bitstream according to BSD. The detected VAR information is represented as an XML file, part of which is shown in Figure 3.12 (a). The adaptation engine utilizes the Structured Scalable Meta-format (SSM) for Fully Content Agnostic Adaptation [113] proposed as a MPEG-21 reference software module by HP Research Labs. In our experiment, the SSM module adapts the size of the JPEG2000 image according to the VAR information and the terminal screen size constraints. Denoting the sizes (in pixels) of the image, VAR

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

and terminal screen as I_{size} , R_{size} and C_{size} , respectively, the adaption decision is carried out according to the following rules:

- $C_{size} > I_{size}$: No adaptation, the original image is displayed unchanged.
- $R_{size} < C_{size} < I_{size}$: Crop the extracted VAR and display the entire VAR.
- $C_{size} < R_{size}$: Crop the extracted VAR and decimate it to the same size as that of the terminal screen.

Figure 3.12 (b) shows an image that is displayed on a PC screen as well as on a PDA. To display the image on the PDA, it is resized so that the VAR extracted using the proposed algorithm is presented to the user. Insignificant background information which corresponds to the non-attention areas have been removed. An example of JPEG2000

```
- <ROIINFO img="../../OriginalJP2/demo1.jp2">
- <SIZE>
  <WIDTH>768</WIDTH>
  <HEIGHT>512</HEIGHT>
</SIZE>
- <ROI>
  <TOP>93</TOP>
  <BOTTOM>484</BOTTOM>
  <LEFT>150</LEFT>
  <RIGHT>591</RIGHT>
</ROI>
</ROIINFO>
```

(a)

(b)

Figure 3.12: (a) Metadata related to the extracted VAR in XML form; (b) Adaptation of an image from PC to PDA.

BSD adaptation is illustrated in Figure 3.13 (b) where the BSD of original image (green) is adapted to the BSD of adapted image (blue) guided by the decision file shown in Figure 3.13 (a). The decision file decides the adaptation parameters according to both terminal screen information and ROI information.

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

Figure 3.13: Example of Digital Item BSD Adaptation; (a) Adaptation Decision Description; (b) JPEG2000 BSD Adaptation (Green - Original BSD, Blue - Adapted BSD).

Through the subjective evaluation of the implemented prototype system, it is shown that the prototype system can improve the experience of viewing images on devices with different screen sizes by utilizing the VAR detection. Especially for mobile devices with extreme small screen, adaptive image resizing according to VAR is much useful than directly downscaling. Using standard MPEG-21 DIA framework makes the proposed prototype system both platform and format independent.

3.5 Conclusion

The focus of this chapter is on strategies for selecting and combining multiple feature contrasts to carry out a spatial bottom-up visual attention analysis for detection of VAR. The CSI method for feature combination considers the convex hull area of salient regions in each feature map together with the density of the saliency points. Based on the two parameters a weight is assigned to each feature map to obtain a weighted linear combination. CSI can be efficiently calculated from the individual feature maps with low computational complexity. The Context Suppression method analyzes the local contextual information to find potential similarity of contrast between an attention

CHAPTER 3. ADAPTIVE FEATURE SELECTION AND COMBINATION

patch and its neighbors. Patches with similar contrast are suppressed through a suppression factor evaluated at each patch. The suppression factor are estimated from the eigen-decomposition of a local 9×9 matrix which can be efficiently computed. We also propose a new texture feature that enables the capture of VAR in images containing small objects present in a cluttered background. Through experiments, we have shown that the proposed algorithms improve the accuracy of VAR extraction for adaptation of image resolution in the MPEG-21 DIA framework.

Chapter 4

Subspace analysis for detection of VARs in images

4.1 Introduction

In the previous chapter, we have proposed CSI and Context Suppression Model to adaptively select and combine multiple feature contrasts to improve the detection of VARs. A more rigorous analysis of the problem of detection of VARs in images opens up the following issues:

- **Global attention:** The mechanism of capturing visual attention region should be perceived as a global process rather than considering local contrast calculated as differences in features among individual pixels or blocks of pixels. An object-based attention paradigm has been corroborated by research in psychophysics and neuroscience [35, 33].
- **Scale Invariance:** The visual attention extraction algorithm should be scale invariant at both the local level and the global level. The former implies that the VAR extraction algorithm should accommodate for multiple features that have saliency at different scales. Globally, salient objects at different scales (sizes) should be detected as VAR consistently.

- **VAR extraction:** The distribution of attention measured at a pixel or block of pixels as represented in a *saliency map*, only indicates the locations of salient pixels and does not contain information about a region. Segmentation of saliency maps may result in erroneous VARs as shown later.

The techniques for extracting VAR described in [1] and [3] are based on local contrast at a pixel or block of pixels. In [11], segmentation is carried out as a pre-processing step to the VAR extraction algorithm. It is unrealistic to expect image segmentation to provide candidates for objects that capture visual attention. Such dependency on region segmentation was also utilized to enhance saliency detection in [53]. None of the models contains a measure to calculate attention between different regions competing to be the VAR. As for scale invariance, although [1] used multi-scale representation for each feature, the model contained a set of predefined scales that are treated equally without considering the selection of a particular scale at each pixel or block. The lack of a provision to select suitable scales at a particular location precludes an accurate representation of a VAR. The performance degradation of these algorithms on VARs at multiple scales will be shown. Thus, the use of spatial neighborhood configurations to extract visual attention regions could fail since (a) it is hard to establish correspondence between fovea locations and image regions and (b) segmentation of the saliency map only groups together pixels with similar contrasts without considering information about regions.

Here, we try to overcome the above problems by proposing a visual attention region detection algorithm based on subspace estimation and analysis. First, the image is mapped onto a linear subspace through a polar transformation so that possible visual attention regions are forced onto the subspace. This can be considered as capturing attention through an object-based attention paradigm at a global level. In doing so, we take into account the scale at which a feature is salient at a particular location in the image. The information within a region is encapsulated in the clusters belonging to a

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

subspace as opposed to a saliency map, which indicates only salient pixels. The subspace representation enables the development of an attention measure that integrates the effect of feature contrast, location and size of the region to extract the VAR. Thus, mapping of the image onto linear subspaces overcomes the three disadvantages of existing models. It also provides a new representation to measure attention of a region by considering the feature contrasts of regions as well as its size and location.

Secondly, multiple linear subspaces corresponding to homogeneous regions are estimated using a variant of the Generalized Principal Component Analysis (*GPCA*) [114]; no segmentation is involved here. *GPCA* is suitable for the problem since it can handle the general case of arbitrary number of subspaces of unknown and possibly different dimensions. Moreover, it does not require the data to be segmented, while other clustering algorithms need to iteratively segment the data and update the model. The robustness of *GPCA* to outliers is improved by embedding the distribution of K nearest neighbors within the *GPCA* framework. To reduce the bias of the estimation towards more densely populated subspaces, we describe an iterative refinement method to accurately estimate both dominant and minor subspaces. We call the new subspace estimation method as the *NN-GPCA* method. The attention region is determined by a new region-based attention measure that considers feature contrasts as well as geometric properties of regions. We show that the proposed visual attention detection method can solve the issue of scalability and determine attentive regions from a global rather than a local perspective. Not only are inter-region contrasts examined but intra-region similarities are also involved to explore visual attention at the region level. Evidently, detection of VARs is different from conventional image segmentation in that the former results in only those regions which are visually significant while the latter simply partitions an image into regions that are homogeneous in some respect.

The rest of the chapter is organized as follows. In section 4.2, we introduce the scale invariant image descriptors and their representation in a subspace. We propose the

novel *NN-GPCA* in section 4.3 and illustrate its robustness to outliers and its ability to be unbiased in the presence of subspaces that are populated very differently. The application of *NN-GPCA* algorithm to extract VARs in images using a novel visual attention measure is described in section 4.4. In section 4.5, experimental results on both synthetic data and real images are presented in order to evaluate the proposed method. Finally, conclusions are drawn in section 4.6.

4.2 Image representation

In this section, we introduce simple features and their transformation in order to represent an image as a collection of linear subspaces so that each subspace corresponds to an image region with some homogeneous property. The proposed transformation is such that it conserves the spatial information of a region, i.e., disconnected regions of similar homogeneity form clusters within the subspace. Homogeneity of regions in an image can be broadly viewed in terms of their smoothness or texturedness. Most feature extraction techniques attempt to incorporate this distinction in their choice of features. Texture features such as Gabor features reside in high dimensional feature spaces and their extraction requires *a priori* information, such as directions for Gabor transforms and the distances in the case of co-occurrence matrices. Here, we analyse only the hue and intensity channels in the *HSV* color space to extract very simple features whose transformation enables an effective representation of the image as a linear subspace. The transformation converts feature similarity of the homogeneous region into linearity property of image data and the spatial proximity of the homogeneous region to clustering property of data. Both these properties convert the problem of visual attention region detection into a linear subspace analysis problem.

4.2.1 Scale-sentient features

Feature extraction has been studied extensively for many years [115, 87]. It is well known that the observation of a specific phenomenon depends on the *scale* at which the world (where the phenomenon occurs) is modeled. Similarly, a notion of scale must be embedded within features to account for the fact that some features become prominent when an object is viewed at a particular scale [116]. Multi-scale representations of image data based on scale-space theory have been proposed [117, 118]. An interesting feature is recognized to present at several scales and/or orientations which discriminated it from noisy and non-relevant detail in [119]. However, mere multi-scale representation of data does not indicate the scale at which the features must be *observed*. Thus, the issue of scale selection is distinct from that of representation. Scale selection has been studied systematically by Lindeberg in [47]. We briefly review the method for automatic feature detection at an appropriate scale.

The scale-space representation of an image is generated by smoothing the image using a set of Gaussian kernels having different variances t as

$$L(\cdot, t) = g(\cdot, t) * f(\cdot) \quad (\text{Eq. 4.1})$$

where $g(\cdot, t)$ is the gaussian kernel and $f(\cdot)$ is the original image. Note that the parameter t is set from 0.5 to 12.5 with a step of 0.5 and form total 25 scales in our work. An m th-order γ -normalized derivative operator is defined as $\partial_{\xi, \gamma\text{-norm}} = t^{m\gamma/2} \partial_x$. Applying this operator on the scale-space representation of (Eq. 4.1), we get the normalized second-order derivatives as $t^\gamma L_{xx}$, $t^\gamma L_{yy}$, $t^\gamma L_{xy}$ and $t^\gamma L_{yx}$, with the last two derivatives being equal. The basic principle for scale selection states that the local maximum over scales for some combination of the γ -normalized derivatives is a likely candidate to correspond to some interesting structure in the image. It is shown in [47] that for detecting blobs, the

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

trace, $\text{trace}(H_{\text{norm}}L)$, and the determinant, $\|H_{\text{norm}}L\|$ of the normalized Hessian matrix can be used to select the correct scale, where

$$H_{\text{norm}}L = \begin{bmatrix} t^\gamma L_{xx} & t^\gamma L_{xy} \\ t^\gamma L_{xy} & t^\gamma L_{yy} \end{bmatrix}. \quad (\text{Eq. 4.2})$$

When $\gamma = 1$, the magnitudes of $\text{trace}(H_{\text{norm}}L) = t(L_{xx} + L_{yy}) = t\nabla^2 L$ at any two locations where the local maximum of $\text{trace}(H_{\text{norm}}L)$ occurs at different scales, are equal; this leads to the notion of scale invariance. Moreover, the scale t_{max} at which the normalized Laplacian-of-Gaussian $t(L_{xx} + L_{yy})$ attains a maximum is related to the characteristic size of the blob. t_{max} is set as the width of the block surrounding at the corresponding location to characterize the characteristic structure of the interesting location. The so-called “interesting points” are those points that are a local extremum of $\text{trace}(H_{\text{norm}}L)$ with respect to both the spatial co-ordinates and the scale parameter. The “interest point” as discussed in this chapter refers to salient blob structures associated with characteristic scale as opposed to extracting image features like lines and edges, e.g. [120].

Figure 4.1 illustrates the process of detecting scale-space extremum of $\|H_{\text{norm}}L\|$ in a synthetic image. The image consists of nine blobs of four different radii together with additive white Gaussian noise of variance 0.1. After generating the scale-space representation of the image, we calculate $\|H_{\text{norm}}L\|$ at each location and for every scale. The local maxima of these values in the scale-space volume are ranked and the top 9 values indicate the location of interesting points in the image. The scales associated with those 9 values reflect the characteristic size of the blob. Figure 4.1 shows the scale space representation of the original image for two of the four scales at which $\|H_{\text{norm}}L\|$ attains a local maximum at the centers of the blobs. The interesting points are represented in 3-D (x-y-scale) space and a ball is drawn at the location of the local maxima with a radius proportional to the scale. For further details on scale-space representation and automatic scale selection, please see [47].

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARS IN IMAGES

Figure 4.1: Illustration of local scale estimation on a synthetic image.

Returning to the extraction of scale sentient features, we first carry out the above scale-space analysis on the *hue* and *intensity* channels of an image. This results in the identification of interesting points and the associated scales. The feature used for further analysis in the VAR extraction algorithm is then simply the *mean* of hue/intensity over a square window that is centered at the interesting point and whose size is determined by the scale. However, at those locations where no interesting points were detected at any scale, the mean is computed over an 8×8 neighborhood which is actually a 9×9 window centered at the interesting point. This implies that the windows can overlap. The mean as a feature is attractive because of its simplicity and the adaptive scale at which it is computed. The combination of interesting points and 8×8 size blocks allows a systematic way to treat high frequency (textured) regions and smooth regions simultaneously. The algorithm automatically decides whether a particular pixel should be considered as an interesting point or as the center of an 8×8 size block. In the following, we show how such a simple feature can be used for accurate extraction of visual attention regions.

4.2.2 Polar transformation of features

Having extracted scale sentient features, we look for a transformation that maps image regions onto linear subspaces. We consider a simple polar transformation of a feature value at one location into a point denoted by (θ, r) . The angle θ is given by

$$\theta(f_i) = \frac{f_i - \min_j(f_j)}{\max_j(f_j) - \min_j(f_j)} \times \pi/2 \quad (\text{Eq. 4.3})$$

where f_i is the feature value at pixel location i and θ is restricted to $[0, \pi/2]$. The radius r is the Euclidean distance of a pixel from a reference point in the image that can be placed at any location. We aim to develop a model for global attention where an object-based attention paradigm is sought, instead of generating a saliency map that utilizes attention in a local region. At the same time, we wish to assign a measure to each region so that the VAR is extracted in an objective and systematic manner. Such an attention measure should include the effect of feature contrast between regions, locations and sizes of the regions so that a cumulative effect of these factors helps to identify the VAR. The linear subspace representation obtained by the polar transformation is appropriate to define such an attention measure where feature contrast is indicated by the angles between the subspaces, location of a region is given by the distance to a reference point, and region size is determined by the projections of the data points onto the normal to the subspace. Therefore, the polar representation embeds the feature information and geometrical properties of image data onto a 2D space and the extraction of VAR and the attendant measure considers both features as well as spatial properties of the region.

A homogeneous connected region contains pixels that have similar feature values and similar distances to a reference point. Therefore, when analyzing regions, both feature homogeneity and spatial proximity of the pixels should be considered. We consider feature values and distances of pixels belonging only to a small neighborhood of each pixel. The size of the neighborhood is either 8x8 or is decided by whether the pixel is an

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

interesting point. Since only pixels in a local context are considered, spatial proximity in local neighborhood can be measured in one dimension as the distance to a reference point.

It is interesting to note that this simple transformation satisfies two conditions that ensure the correspondence of an image region to a subspace [121]. They are

- **Subspace Constraint:** We require that each homogeneous region in an image corresponds to one and only one subspace (linear or nonlinear). The angle θ in the polar transformation ensures the mapping of feature values of one region onto a line in 2D space.
- **Cluster Constraint:** In addition to the features lying on a subspace, they should also reside in clusters within the subspace. Thus, data not belonging to a cluster are considered as outliers. The radius r in the polar transformation forces clusters to be formed within the subspace.

It is indeed possible that the transformed features may lie in subspaces that are very differently populated. More specifically, if an image consists of two regions of which one is very small, then the subspace corresponding to the smaller region will be significantly less populated than the one corresponding to the larger region. We call this phenomenon as *subspace skew*. Figure 4.2 (a) and (b) show a synthetic monochrome image and the polar transformation of its intensity feature. The reference point is chosen as the center of the image. The two linear subspaces corresponding to the two regions are evident. Figure 4.2 (c) and (d) show a noisy synthetic color image and the polar transformation of its hue. The four distinct regions are mapped into four linear subspaces although other subspaces are also formed due to noise. However, the sizes of the clusters formed by data and noise are significantly different, causing them to be distinguished easily. Hence, the true regions can be detected while noise is filtered out by imposing the subspace

Figure 4.2: (a) Synthetic monochrome image and (b) polar transformation of its intensity. (c) Synthetic color image with noise and (d) polar transformation of its hue.

and cluster constraint. Note also that the sizes of the clusters indicate the sizes of the corresponding regions.

In Figure 4.2, the reference point is taken as the image center. However, it can be set to any location in the image. The eventual identification of a region as VAR is the cumulative effect of feature contrast of regions as reflected by the angles between the corresponding subspaces, the region size as determined by the projections of the data points onto the normal to the subspace and location of a region as determined by its distance to a reference point. The polar transformation places no constraint on the shape of a region. For two regions with arbitrary shape but having similar features and at different distances to a reference point, their pixels will be transformed into two clusters within the same subspace in the polar representation.

4.3 NN-GPCA Subspace Estimation

The next step in the proposed VAR extraction mechanism is to estimate the subspaces onto which the image feature has been mapped. This involves determining the number of subspaces and their orientations in the presence of noise. In doing so, we do not rely on any segmentation algorithms. Vidal et al. [114, 122] have proposed an algebraic geometric approach called Generalized Principal Component Analysis (*GPCA*) for subspace modeling. As we will show later, the performance of *GPCA* degrades with outliers and skewed data. The *GPCA* algorithm is made more robust to outliers by combining it with the distribution of K nearest neighbors g_{knn} that yields a weighted estimation of the subspaces so that outliers are weighted less than the inliers. On the other hand, *subspace skew* might force the estimation to treat the less populated subspace as noise forcing the estimation to be biased towards densely populated subspaces. This problem is solved by refining the initial estimation of weights through an iterative refinement procedure. These two components form the proposed new *NN-GPCA* method for subspace estimation. It extends the robustness of *GPCA* to fit both the subspace constraint and the cluster constraint for noisy and skewed (in terms of relative sizes of regions) image data. In this section, we first give a brief review of the *GPCA* algorithm including its limitations followed by the description of the proposed *NN-GPCA* algorithm.

4.3.1 Review of *GPCA*

GPCA is an algebraic geometric approach to the problem of estimating a mixture of linear subspaces from sample data points. It utilizes the fact that each data point $x \in R^k$ satisfies $b_i^T x = \sum_{j=1}^k b_{ij} x_j = 0$ where b_i is the normal vector of the subspace it belongs to. Since every sample point lies on one of the subspaces, the following equation about homogeneous polynomial of degree n on x with real coefficients holds,

$$p_n(x) = \prod_{i=1}^n (b_i^T x) = 0 \quad (\text{Eq. 4.4})$$

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARS IN IMAGES

where n is the number of subspaces and $\{b_i\}_{i=1}^n$ are normal vectors of the subspaces. The problem of estimating subspaces is to solve this nonlinear system for all $\{b_i\}_{i=1}^n$. It can be converted to a linear expression by expanding the product of all $b_i^T x$ and viewing all monomials $x^n = x_1^{n_1} x_2^{n_2} \dots x_K^{n_K}$ of degree n as system unknowns. Using the definition of *Veronese map* [114] $v_n : [x_1, \dots, x_K]^T \mapsto [\dots, x^n, \dots]^T$, (Eq. 4.4) becomes the following linear expression: $p_n(x) = v_n(x)^T c = \sum c_{n_1 \dots n_K} x_1^{n_1} \dots x_K^{n_K} = 0$, where $c_n \in R$ represents the coefficient of the monomial x^n . Given a collection of N sample points $\{x^j\}_{j=1}^N$, a linear system can be generated as

$$L_n c \doteq [v_n(x^1), v_n(x^2), \dots, v_n(x^N)]^T c = 0 \quad (\text{Eq. 4.5})$$

to solve all normal vectors $\{b_i\}_{i=1}^n$ to the subspaces.

In the absence of noise, the number of subspaces can be estimated from the requirement that there is a unique solution for c . The normal vectors $\{b_i\}_{i=1}^n$ can be solved once c is determined. However, in the presence of noise, the subspace estimation problem is cast as a constrained nonlinear optimization problem which is initialized using the normal vectors obtained as above. Further details of the *GPCA* algorithm are available in [114] and [122].

While the *GPCA* algorithm provides an elegant solution to the problem of linear subspace estimation without segmenting data, there are some inherent limitations that we discuss next.

- (i) Estimation of the number of subspaces: *GPCA* relies on a predefined threshold to combine the subspaces that are very close to each other. This threshold does not have a meaningful interpretation and there is no easy way to decide its value.
- (ii) Effect of outliers: Each data point is either true data or noise that appears as outliers in the subspace representation. *GPCA* solve first c and then the normal

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARS IN IMAGES

vectors of the subspaces by using Least Square Approximation from Eq. 4.5. When even only a few outliers are introduced, least square approximation of c will heavily degrade and propagated to the solution of normal vectors. This will lead to an erroneous estimation of the subspaces.

- (iii) Bias due to *subspace skew*: Since the objective function in the optimization problem consists of the sum of approximation errors, for a fixed number of subspaces, the estimation will be biased towards those subspaces that are populated more.

We illustrate the failure of the *GPCA* algorithm using synthetic data containing outliers in Figure 4.3 where the red lines represent the estimated subspaces. The data lie on four linear subspaces shown in Figure 4.3 (a) of which two subspaces contain true data and the other two contain outliers. In the absence of outliers as shown in Figure 4.3 (b) the *GPCA* estimation performs very well. However, when noisy data is also taken into account, the initial estimate of the subspaces shown in Figure 4.3 (c) and the final estimate using nonlinear optimization shown in Figure 4.3 (d) are erroneous. (In the figures, the length of the lines representing subspaces have no implication; what is important is the orientation of these lines). Figure 4.4 illustrates how *GPCA* is biased towards subspaces that are more populated. Figure 4.4 (a) shows 600 data points lying on 2 subspaces with additive Gaussian noise with variance of 3. We have introduced subspace skew in the distribution of the data between the two subspaces so that 500 points lie on one subspace and the other 100 points on the other subspace. Figure 4.4 (b) shows the initial estimate of the subspaces by *GPCA* and figure Figure 4.4 (c) shows the final estimate after optimization. Both the estimates are biased towards the more populated subspace. These drawbacks of the *GPCA* algorithm are overcome by the proposed *NN-GPCA* algorithm.

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

Figure 4.3: Effect of outliers on subspace estimation by *GPCA* (a) Synthetic data (b) estimation using true data only (without outliers); (c) initial estimate used and (d) final estimate after optimization.

4.3.2 Assigning weights to data points

A subspace clustering method using the K^{th} nearest neighbor distance ($kNND$) metric is shown to detect and remove outliers and small data clusters [121]. The $kNND$ metric uses the fact that in a cluster larger than K points, the $kNND$ for a data point will be small; otherwise it will be large. According to the polar transformation of features, the true data lies not in just any cluster but in the cluster inside its subspace. Instead of using $kNND$, we utilize the distribution of all K nearest neighbors denoted as g_{knn} to differentiate inliers and outliers. Here, we assign a weight to each data point x^i calculated from $g_{knn}(x^i)$. This provides a simple method to reduce the sensitivity of outliers without

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

Figure 4.4: Effect of subspace skew on subspace estimation by *GPCA* (a) Synthetic data (b) initial estimate and (c) final estimate after optimization.

segmentation of data. The weight is related to the probability of a data point lying in any cluster of a subspace corresponding to an image region. Given a sample data x^i , its K nearest neighbors are detected and the variance $svar(g_{knn}(x^i))$ along the direction of the subspace of x^i (from origin to the current point) and the variance $nvar(g_{knn}(x^i))$ along the orthogonal direction are calculated using Principal Component Analysis (*PCA*). The sum $S(g_{knn}(x^i)) = svar(g_{knn}(x^i)) + nvar(g_{knn}(x^i))$ corresponds to the characteristic variance of K nearest neighbors of the current point. It will be small if these K neighbors form a cluster, otherwise it will be large. Since only the clusters inside the subspace are true data, we use the ratio $R(g_{knn}(x^i)) = (svar(g_{knn}(x^i)) + nvar(g_{knn}(x^i))) / svar(g_{knn}(x^i))$ as the factor to bias the weights to those clusters in the subspace that correspond to true data. Consider the product

$$\overline{T(x^i)} = S(g_{knn}(x^i)) \times R(g_{knn}(x^i)) \quad (\text{Eq. 4.6})$$

where $\overline{(\cdot)}$ means that the product has been normalized to lie in $[0, 1]$. The weight for x^i is calculated as

$$W(x^i) = 1 - \overline{T(x^i)}. \quad (\text{Eq. 4.7})$$

When the data point x^i lies in a cluster larger than K inside a subspace, $R(g_{knn}(x^i))$ is close to 1 ($R(g_{knn}(x^i)) = 1$ in the absence of noise) since $nvar(g_{knn}(x^i))$ is small

($nvar(g_{knn}(x^i)) = 0$ in the absence of noise). The sum $S(g_{knn}(x^i))$ is also small because these K data form a cluster. So $W(x^i)$ is equal to or close to 1. For data points that do not lie in any cluster or whose cluster does not align with a subspace, $R(g_{knn}(x^i))$ and/or $S(g_{knn}(x^i))$ are large and $W(x^i)$ is small or even close to zero. Additionally, the parameter K decides the minimum size of the cluster that is considered as true data. Since the outliers are always far away from the true data, any small value of K can differentiate outliers from true data and the selection of this value can be fixed for all cases. Hence the weight of each data point relates to the probability that they are inside the cluster of a specific subspace corresponding to one image region. Thus, the analysis of $g_{knn}(x^i)$ assigns small weights to outliers to reduce their effect on subspace estimation.

4.3.3 Weighted Least Square Approximation of c

The weights obtained from the analysis of the distribution g_{knn} of K nearest neighbors are used to improve robustness and accuracy of subspace estimation. By taking the weight of each data point x^i into account, the linear system of (Eq. 4.5) is modified as

$$WL_n c \doteq \begin{bmatrix} W(x^1) & & & \\ & W(x^2) & & \\ & & \ddots & \\ & & & W(x^N) \end{bmatrix} \begin{bmatrix} v_n(x^1)^T \\ v_n(x^2)^T \\ \vdots \\ v_n(x^N)^T \end{bmatrix} c = 0 \quad (\text{Eq. 4.8})$$

where $W(x^i)$ is the weight of x^i .

In order to estimate the number of subspaces, we first modulate L_n using the weights $W(x^i)$ for each x^i as

$$\tilde{v}_n(x^i) = \bar{v}_n + W(x^i)(v_n(x^i) - \bar{v}_n) \quad (\text{Eq. 4.9})$$

where $\bar{v}_n$ is the mean of the data. If x^i is an outlier, its small weight causes it to be pulled closer to the mean of the data. Next, we do a Singular Value Decomposition (SVD) on WL_n and eliminate the outliers using a very weak threshold. We emphasize that the

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

choice of threshold in this case is not crucial since the weights allow less dependency on the threshold. This is unlike the case for *GPCA* where the presence of outliers may cause the number of dominant singular values to be large.

The subspace estimation problem is formulated as an approximation problem using weighted least square technique for the purpose of calculating coefficient vector c . Since c can be obtained only up to a scale factor, we normalize it by the first component c_1 . Thus the left side of (Eq. 4.8) becomes

$$W \begin{bmatrix} v_n(x^1)(2..M)^T \\ v_n(x^2)(2..M)^T \\ \vdots \\ v_n(x^N)(2..M)^T \end{bmatrix} \begin{bmatrix} c_2 \\ c_3 \\ \vdots \\ c_N \end{bmatrix} + \begin{bmatrix} W(x^1)v_n(x^1)(1)c_1 \\ W(x^2)v_n(x^2)(1)c_1 \\ \vdots \\ W(x^N)v_n(x^N)(1)c_1 \end{bmatrix} \quad (\text{Eq. 4.10})$$

where $v_n(x^i)(2..M)$ represents a vector containing all components of $v_n(x^i)$ except for the first component $v_n(x^i)(1)$. With $c_1 = 1$, (Eq. 4.8) can now be rewritten as

$$W \begin{bmatrix} v_n(x^1)(2..M)^T \\ v_n(x^2)(2..M)^T \\ \vdots \\ v_n(x^N)(2..M)^T \end{bmatrix} \begin{bmatrix} c_2 \\ c_3 \\ \vdots \\ c_N \end{bmatrix} = \begin{bmatrix} -W(x^1)v_n(x^1)(1)^T \\ -W(x^2)v_n(x^2)(1)^T \\ \vdots \\ -W(x^N)v_n(x^N)(1)^T \end{bmatrix} \quad (\text{Eq. 4.11})$$

The above equation can be succinctly written as $WAc = d$, where A is the matrix whose rows are $v_n(x^i)(2..M)^T$, $i = 1, 2..N$ and d is the right side of (Eq. 4.11). By minimizing the objective function $\|d - Ac\|_W$, we can obtain the weighted least square approximation of c_i , $i = 1, 2, 3..N$ as

$$\begin{aligned} \|d - Ac\|_W &= [W(d - Ac)]^T [W(d - Ac)] \\ &= (d^T - c^T A^T) W^T W (d - Ac) \\ &= d^T W^T W d - d^T W^T W A c - c^T A^T W^T W d + c^T A^T W^T W A c \end{aligned} \quad (\text{Eq. 4.12})$$

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

By differentiating the above function with respect to c , we obtain

$$\frac{\partial \|d - Ac\|_w}{\partial c} = 2A^T W^T W A c - 2A^T W^T W d \quad (\text{Eq. 4.13})$$

By setting $\frac{\partial \|d - Ac\|_w}{\partial c} = 0$, we obtain the solution of c as follows

$$c_1 = 1 \quad \text{and} \quad [c_2, \dots, c_N]^T = (A^T W^T W A)^{-1} (A^T W^T W d) \quad (\text{Eq. 4.14})$$

The estimation error of coefficient vector c is reduced by the diagonal matrix of weight coefficient W .

4.3.4 Iterative Refinement of c

As mentioned in section 4.3.1, the phenomenon of subspace skew may cause two or more subspaces to be very differently populated. A simple least squares estimation of the subspaces based on minimizing the sum of square errors will draw the estimate closer to the more populated subspace. Moreover, in the *GPCA* algorithm, the linear subspaces are transformed into a single linear model in the high-dimensional embedding space via the Veronese map. A least squares estimate in the high-dimensional space only minimizes the sum of squared error in the embedding space and does not necessarily minimize the error in the low-dimensional space. Hence, we propose an iterative refinement algorithm that adjusts the weight of each data point so that the final estimates of the subspaces are close to the true subspaces.

We begin the iteration with the coefficient vector c obtained from the weighted least squares approximation described in the previous section. We calculate the error for each data point x^i according to

$$e_i = \frac{W(x^i) \times |v_n(x^i)^T c_0|}{\|v_n(x^i)\|} \quad (\text{Eq. 4.15})$$

where $W(x^i)$ is the weight of x^i , c_0 is the initial estimate of c , $|\cdot|$ is the absolute value of the projection of $v_n(x^i)$ onto c_0 and $\|\cdot\|$ is the length of the vector $v_n(x^i)$. The weight

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

of each data point is then updated as

$$W^{p+1}(x^i) = W^p(x^i) + e_i \quad (\text{Eq. 4.16})$$

where p is the iteration number. The new value of c is obtained from the weighted least squares solution of $W \cdot L_n \cdot c = 0$. During the initial steps of the iteration procedure, the weights of the data points in the smaller subspace change more than those in the dominant subspace. As iteration proceeds, the weights in the smaller subspace change less as the error decreases. The cumulative effect of the change in weights on all the subspaces is reflected in the sum and variance of the errors. Hence the sum (s_e) and variance (v_e) of the errors serve as the stopping criterion, (rather than the errors in individual data points) i.e. if $|s_e^{p+1} - s_e^p| < \epsilon \wedge |v_e^{p+1} - v_e^p| < \epsilon'$, the iteration stops. Contrary to usual gradient-descent type of iterative algorithms in which the weights are reduced, here the weights of those data points with large residual are increased so that in each iteration, the importance of data with large residuals is increased by their weights resulting in reduction of error in their new estimates. In other words, we iteratively equalize the importance of the inliers in the embedding space by adjusting the weights of data according to their residuals in the current estimation. Since the initial weights of the outliers calculated from the K nearest neighbors are small, they are not increased significantly in the refinement process. Thus, only the inliers are considered in this process and the optimal estimation of c fits all the inliers in the embedding space and results in a more accurate estimation of b in the original space. It was found through experiments that the iterative refinement of c converged in only about 5 to 10 iterations.

The normal vectors $\{b_i\}_{i=1}^n$ are calculated from c using the same method as in *GPCA* [114]. These vectors serve to initialize the following constrained nonlinear optimization which differs from [114] in the weight matrix:

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

$$\begin{aligned} & \min \sum_{j=1}^N W(x^j) \|\tilde{x}^j - x^j\|^2 \\ & \text{subject to } \prod_{i=1}^n (b_i^T \tilde{x}^j) = 0 \quad j = 1, \dots, N. \end{aligned} \quad (\text{Eq. 4.17})$$

By using Lagrange multipliers λ^j for each constraint, the above optimization problem is equivalent to minimizing the function

$$\sum_{j=1}^N (W(x^j) \|\tilde{x}^j - x^j\|^2 + \lambda^j \prod_{i=1}^n (b_i^T \tilde{x}^j)). \quad (\text{Eq. 4.18})$$

Taking partial derivatives w.r.t $\tilde{x}^j$ and equating it to 0, we can solve for $\lambda^j/2$ and $W(x^j) \|\tilde{x}^j - x^j\|^2$. Substituting them back into the objective function (Eq. 4.17), the simplified objective function on the normal vectors can be derived as

$$E_n(b_1, \dots, b_n) = \sum_{j=1}^N \frac{W(x^j) (n \prod_{i=1}^n (b_i^T x^j))^2}{\|\sum_{i=1}^n b_i \prod_{l \neq i} (b_l^T x^j)\|^2}. \quad (\text{Eq. 4.19})$$

It was observed that the convergence of (Eq. 4.19) was very slow. Hence, a weighted k-means iteration method was used to determine the optimized $\{b_i\}_{i=1}^n$. The weighted data points are assigned to the nearest subspace and the updated subspaces are estimated. This process is continued till there is no change in the subspaces. This method of optimizing the b_i 's achieves the same performance as (Eq. 4.19) but at a faster convergence rate.

We now illustrate the superiority of the proposed *NN-GPCA* algorithm in terms of its robustness in the presence of noise and subspace skew. We consider the same synthetic data of Figure 4.3 that was used to demonstrate the failure of the *GPCA* algorithm. Recall that the data consists of two subspaces and some outliers. The initial estimate of the subspaces calculated using the proposed *NN-GPCA* algorithm that reduces the weight of outliers is shown in Figure 4.5 (a). The effect of the two subspaces formed by outliers is reduced in the initial estimate. The final estimates of the subspaces obtained after optimization are shown as green lines in Figure 4.5 (b). The optimization process results

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

Figure 4.5: Improved subspace estimation using *NN-GPCA* for data with outliers (a) initial estimate and (b) final estimate after optimization (compare with Figure 4.3).

Figure 4.6: Improved estimation using *NN-GPCA* for data with subspace skew (a) estimation after 3rd iteration; (b) initial estimate after convergence and (c) final estimate after optimization (compare with Figure 4.4).

in the correct estimation of the two subspaces that satisfy the cluster constraint while ignoring the outlier subspaces. The robustness of the *NN-GPCA* algorithm to subspace skew is also illustrated with the same synthetic data as in Figure 4.4. As iteration proceeds, the subspace estimates converge to the true subspaces on which the data lie, as shown in Figure 4.6 (a) for the third iteration, Figure 4.6 (b) for the initial estimate obtained after convergence and Figure 4.6 (c) for the final estimate after optimization.

Figure 4.7 shows the result of the proposed *NN-GPCA* based subspace estimation algorithm on real data. Here, we consider *hue* as the feature and transform it into the polar 2-D space resulting in a noisy representation in the subspaces. However, the

Figure 4.7: *NN-GPCA* algorithm on real image (a) Original image and (b) estimation of the three subspaces using *hue* feature.

proposed *NN-GPCA* method can correctly estimate the three main subspaces (three red lines) corresponding to three image regions with different hue. All outliers not lying in a cluster are ignored because of their small weights. Notice that multiple objects with similar visual appearance at different locations will be mapped to a single subspace. Hence, the proposed approach can detect similar regions simultaneously and assign the same amount of attention to them using the attention measure introduced later.

4.3.5 Computational Complexity

We derive the computational complexity of the *NN-GPCA* algorithm for an image that contains n subspaces. The number of data points in the polar representation is denoted as N which consists of the number of interesting points and the number of blocks each of 8×8 pixels where no interesting points were located. It is much smaller than the total number of pixels.

Computation of weights

Estimation of the K nearest neighbors requires complexity of $O(N^2)$ without any optimization. Since K is much smaller than N , the subsequent process of calculating weights

using *SVD* on K nearest neighbors is of much lower order and hence can be ignored. Therefore, the complexity in this step is $O(N^2)$.

Subspace estimation

This process consists of the following steps: (i) estimating n , the number of subspaces, for which the complexity is $O(N)$ since n is a small number; (ii) solving for $c_i, i = 1, 2, 3 \dots N$ from a system of N linear equations in $n + 1$ unknowns. Without optimization, the complexity is $O(n^2 \cdot N)$ without noise and $O(n \cdot N^2)$ in the presence of noise; (iii) iterative refinement of c in which each iteration consists of updating the weights and resolving the linear system in (ii). The complexity of each iteration is $O(N + n \cdot N^2)$ in the presence of noise. Since the number of iterations, p , for convergence is very small (usually less than 10), the complexity of this step is $O(p \cdot N + p \cdot n \cdot N^2)$; (iv) solving for $\{b_i\}_{i=1}^n$, which, for the 2D case, requires determining n roots of a linear equation of degree n in one unknown. Its complexity can be ignored; (v) optimization using weighted k-means has complexity $O(N \cdot \log(N))$ which is lower than $O(N^2)$. Hence, the total complexity is $O((p + 1) \cdot N + (p + 1) \cdot n \cdot N^2 + N \cdot \log(N)) = O(p \cdot n \cdot N^2)$ which is in the same order as the complexity of *GPCA*.

From the above analysis, we can see that the computational complexity of *NN-GPCA* is $O(N^2 + p \cdot n \cdot N^2) = O(p \cdot n \cdot N^2)$ which, again is the same order as the complexity of *GPCA*. We conclude that our extension of *NN-GPCA* improves the robustness of multiple subspace estimation without heavily increasing the computational complexity. Since the number of data points N is much smaller than the number of pixels and the number of subspaces n and also the number of iterations p is small, the computational complexity of $O(p \cdot n \cdot N^2)$ is small. When applying the *NN-GPCA* algorithm on multiple features, its complexity increases only linearly with the number of features.

4.4 Subspace based region attention measure

Having described the *NN-GPCA* algorithm to estimate the subspaces of a transformed image, we now present a new attention measure to detect visual attention regions from these subspaces. In order for a region to be a likely candidate for visual attention, it should satisfy the following criteria: (i) in the image domain, it should differentiate itself from its neighboring region in terms of high feature contrast (e.g. region color) and some spatial property (e.g. size and shape) and (ii) in the subspace domain, it should satisfy the clustering constraint which causes identical regions to form clusters within a subspace. The proposed attention measure integrating these factors consists of two parts: (i) *Cluster Compactness (CC)* - it is a measure of the compactness of a single cluster or, in the case of multiple clusters, the 'net' compactness for all clusters in a subspace; however, before determining *CC*, we remove noise which do not belong to any cluster. This is different from removing noise during estimation of the subspaces; (ii) *Cumulative Projection (CP)* measures the feature contrast among candidate visual attention regions in terms of the projection of all data points onto the normal of a subspace. These two components will be described in the following subsections.

4.4.1 Cluster Compactness

As noted in section 4.2.2, the cluster constraint requires data points belonging to a region to form a cluster in a subspace. Points that do not lie in a cluster are noisy data and should be removed before calculating the attention measure. We present a cluster growing algorithm that removes noise during cluster formation. While in other clustering algorithms like *k - means* clustering, the selection of the number of clusters and corresponding seeds are critical issues, here we use the weights W associated with each data point (these weights were calculated earlier in section 4.3.2) to indicate the possibility of a point belonging to a cluster. Data points with large W are more likely to

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

belong to some region in the image. Hence, we sort the points according to their weights and pick the top m points to form an initial set of inliers. It is interesting to note that in the case of multiple clusters within a subspace, these m data points will naturally come from each of the clusters, which implies that the initial set contains points drawn from all regions falling on that subspace. Next, we determine the Euclidean distance between each inlier and its nearest neighbor in the set of inliers. A normalized average minimum distance $\bar{d}$ is defined as $\bar{d} = D/W$, where D is the sum of the minimum distances obtained for all inliers and W is the sum of weights of all inliers. Given a data point x^i , we calculate its distance $d(x^i)$ to the nearest inlier. If

$$\frac{d(x^i)}{W(x^i)} \leq u \times \bar{d}, \quad (\text{Eq. 4.20})$$

the data point x^i is included in the set of inliers. Here, u is a constant that is required to extend the threshold beyond the average minimum distance; from experiments, it was observed that the choice of u was not critical. If a point does not satisfy the condition in (Eq. 4.20), it is discarded as noise. Normalization by weight of each data point increases the discriminatory power across different images, since a small (large) weight associated with noise (inlier) causes the left hand side of (Eq. 4.20) to be even larger (smaller). Thus, the initial set of inliers is grown by checking if each data point satisfies (Eq. 4.20). When all data points have been considered, the algorithm returns the set of inliers that form clusters in each estimated subspace. We now define the cluster compactness for a subspace with normal b_j as

$$CC(b_j) = \left(\frac{\overline{d^{cl}}}{\overline{d_{b_j}^{cl}}} \right), \quad (\text{Eq. 4.21})$$

where $\overline{d^{cl}}$ is the mean of the same distance over all subspaces and $\overline{d_{b_j}^{cl}}$ is the mean of the $kNND$ of all inliers to the subspace with normal b_j . $CC(b_j)$ measures the compactness of the cluster(s) in one subspace compared to the other subspaces. This measure is used in conjunction with the Cumulative Projection to assign attention measures to regions in the image.

4.4.2 Cumulative Projection

Due to the polar transformation of image features, each region lies on a linear subspace and hence a measure of feature contrast between the regions is simply $|\theta_1 - \theta_2|$, where θ_1 and θ_2 are the angles made by the subspaces. The projection of data points onto the normal to a subspace is an indication of how visually important the corresponding region is to the human visual system and serves as a good measure for region attention because it considers both feature contrast as well as the geometric properties of the region. We call this measure as the *Cumulative Projection (CP)* and define it as

$$CP(b_j) = CC(b_j) \times \left(\frac{\sum_{x^i \in \text{Inliers}} |(x^i)^T b_j|}{\sum_{x^l \in \text{Inliers}} \|x^l\|} \right), \quad (\text{Eq. 4.22})$$

where *Inliers* denote all inliers in all subspaces. Thus, the *CP* for a subspace is the sum of the projections of all normalized inliers onto its normal with the sum weighted by the cluster compactness of the subspace. This measure ensures that in addition to the dissimilarity of data between two subspaces, the similarity as reflected by the compactness of data in a particular subspace is also taken into account. Therefore, a subspace onto whose normal the projections of the inliers is large and whose cluster compactness is also large will have a high value of *CP* and hence, the region(s) corresponding to that subspace will have a high value of visual attention.

Besides feature contrast, the *CP* attention measure inherits two important properties about the size and location of the region that make it consistent with the human visual system (*HVS*) to correctly detect VARs. Firstly, an image consisting of a small object within a large background draws attention to the object first. Even if the differences in the feature values between object and background are not large, the *CP* attention value biases the attention to the smaller object in the foreground. This is because there are more data belonging to large region projecting onto the subspace of small region although the feature contrast between two subspaces is fixed. Hence the projection of all

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARS IN IMAGES

Figure 4.8: Images to illustrate effect of region size and location on Cumulative Projection

Table 4.1: Cumulative Projection values

Image	Light Region	Dark Region
Image (a)	0.5772	0.0005
Image (b)	0.1549	0.4223
Image (c)	0.7071	0.5528

data onto the normal of the subspace representing the small object will be larger than that on the normal of the subspace representing the background. It is consistent to the sparsity property of human perception for visual attention. Secondly, CP also captures the variability of attention on *location* of regions within an image. Most often, attention is drawn to the center of an image first. Closer a region is to the center of the image, higher will be its CP value. We illustrate these properties through a synthetic example in Figure 4.8, which shows three images with (a) small object on a large background, (b) a large object in the foreground and (c) two objects in the foreground but at different distances from the center of the image. As shown in the Table 4.1, the smaller brighter region in Figure 4.8 (a) has larger CP attention value than the background. Similarly, the attention is drawn to the edges in Figure 4.8 (b) first, which is reflected by the CP in Table 4.1. In Figure 4.8 (c), because the sizes of the subspaces representing the white region and the black region are the same and the contrast between these regions and the background are identical, both would possibly capture visual attention to the same extent. However, a higher CP attention value is obtained for the white region by virtue

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

of it being closer to the center of the image. So the white region will be focused before black region. Coming back to Figure 4.8 (b), one could argue that since the location of the darker edge region is farther from the center of the image, it should garner lower CP value than the lighter region. But in this case, the effect of the region size overrides the location property to generate a higher CP value, which is once again consistent with the HVS. Furthermore, if we were to imagine the lighter square region to shrink so that its size (number of pixels) is the same as the size of the background region, attention would now be drawn to this region since it is closer to the center.

The *Cumulative Projection* measures visual attention in a global perspective by considering inter-region contrasts as well as other properties like size and location of the regions. In the case of multiple features, we do a similar analysis on each feature to obtain the values of CP for each subspace. Based on the CP value, we select the feature that is finally used to extract VARs. In selecting a particular feature, we must consider two factors: (i) the largest value of CP among all the features and (ii) the difference between the largest and second largest values of CP for a particular feature. This fact is illustrated in Figure 4.9 which shows the CP values of each subspace in the hue and intensity features. The attention region is distinguishable in the intensity feature space although the largest value of CP appears in the hue feature space as $CP_h^{(1)}$. The difference between the largest and second largest values, $CP_i^{(1)}$ and $CP_i^{(2)}$ respectively, in the intensity feature is more significant than the corresponding differences in the hue feature. Here, we implement this difference as a ratio so that the automatic selection of the feature which contains the subspace corresponding to the attention regions is obtained by

$$R_f = CP_f^{(1)} \times \frac{CP_f^{(1)}}{CP_f^{(2)}} \quad (\text{Eq. 4.23})$$

where $CP_f^{(1)}$ and $CP_f^{(2)}$ are the largest and second largest values of CP in the polar feature space of $f \in \{\text{hue}, \text{intensity}\}$. The feature that yields the highest value of this measure is automatically selected as the one which is finally used to extract VARs.

Figure 4.9: Illustration of CP function in hue and intensity,

In Figure 4.10 (a), the hue and intensity features of an image are shown. Their corresponding subspaces are shown in Figure 4.10 (b) in which the first and second largest CP for the hue feature is $CP_h^{(1)} = 0.4011$ and $CP_h^{(2)} = 0.2559$ and those for the intensity feature are $CP_i^{(1)} = 0.8522$ and $CP_i^{(2)} = 0.2651$ (both subspaces having largest CP are shown in green). Since $R_{hue} = 0.6287 < R_{intensity} = 2.7395$, the system automatically selects intensity to detect visual attention of the textured object. From the extracted attentive region, we can see that the detected region corresponds to the attentive object detected by the *HVS*.

4.5 Experimental Results

We conduct several experiments to demonstrate various properties of the proposed VAR extraction mechanism. We illustrate specifically the robustness and scale invariance of the algorithm, and also compare the performance of the proposed algorithm with Itti's model [1] and Ma's model [3]. We also show the ability of the algorithm to automatically select the appropriate feature for VAR detection. As noted earlier, the two features that we consider are *hue* and *intensity*. These are chosen to illustrate how such simple features can be used to detect VARs in challenging situations like a textured object on a

Figure 4.10: Multiple feature analysis: (a) feature map, (b) subspace estimation and (c) extracted attention region using Top row: Hue feature and Bottom row: Intensity feature.

differently textured background, using the proposed algorithm. The extraction of scale-sentient features through scale-space analysis involves combining the first 1500 interesting points having the largest responses with fixed 8×8 blocks at locations with no interest points.

4.5.1 Experiment 1

In the first experiment, we compare the proposed *NN-GPCA* subspace estimation method with *GPCA* on synthetic data. To evaluate the robustness to the outliers, we randomly pick $n = 2$ subspaces each containing $N = 400$ points on 1-dimensional subspaces of R^2 . Zero-mean Gaussian noise with standard deviation of 20 is added to the sample points. We consider six cases of outliers that are randomly added to the data such that the number of outliers are 0%, 2.5%, 5.0%, 7.5%, 10%, and 12.5% of the original number of data points. These outliers themselves lie on other multiple subspaces. For each of the six cases, the *NN-GPCA* and *GPCA* algorithms are run 500 times. The error between the true unit normals $\{b_i\}_{i=1}^n$ and their estimates $\{\tilde{b}_i\}_{i=1}^n$ is computed for each run as

[114]

$$error = \frac{1}{n} \sum_{i=1}^n \cos^{-1}(b_i^T \tilde{b}_i). \quad (\text{Eq. 4.24})$$

Figure 4.11 (a) plots the mean error calculated over all the 500 trials as a function of the percentage of outliers. We see that when there are no outliers, the estimation errors corresponding to *NN-GPCA* and *GPCA* are the same. As the number of outliers increases, *NN-GPCA* outperforms *GPCA*. RANSAC [123], as a well-known model fitting technique, can remove outlier for *GPCA*. We also compared the proposed *NN-GPCA* with this method (denoted as RANSAC+*GPCA*). Because RANSAC cannot be directly applied to a mixture of model (Either the data belonging to some models are removed as outliers or some outliers will be included as inliers in this case), we iteratively apply RANSAC for fitting multiple subspaces. Both methods are tested on the noisy synthetic data of 2 subspaces including 20% outliers. We can see from Table 4.2 that *NN-GPCA* outperforms either *GPCA* or RANSAC+*GPCA*. It is because RANSAC recognizes 20% outliers lying in some subspace as inliers but these outliers do not form any cluster. The cluster constraints cannot be integrated into the RANSAC framework, which is a local method based on sampling. In this experiment, we also compared the method of Hough Transform [124] which can detect parametric structure. From the error reported in Table 4.2, we also found the proposed *NN-GPCA* outperforms hough transform in such noisy situation. Note that the large error of hough transform caused by its sensitivities to noise as well as the discretization of parameter space.

To evaluate the robustness of the *NN-GPCA* algorithm to subspace skew, we generate $n = 2$ subspaces with very different populations and add Gaussian noise with variance of 10 to the sample points. We test the algorithm for different ratios of the number of data points in one subspace to that in the other starting from 1:1 till 1:10. We run the estimation 100 times on each ratio and calculate the error as in (Eq. 4.24). Figure 4.11 (b) shows that *GPCA* and *NN-GPCA* achieve the same performance for small ratios but

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

Figure 4.11: Comparison of *NN-GPCA* and *GPCA* algorithms in terms of estimation error of subspaces: (a) error v/s percentage of outliers in data, (b) error v/s ratio of subspace skew, (c) the sum and (d) the variance of the residuals during estimation.

Table 4.2: Estimate Error on synthetic data of 2 subspaces + 20% outliers

	GPCA	RANSAC+GPCA	HoughTransform	NN-GPCA
Estimate Error	0.186	0.651	0.556	0.103

as the ratio increases the *NN-GPCA* algorithm far outperforms the *GPCA* algorithm. In continuation with the analysis of subspace skew, we show the sum and variance of the residuals of each data point to its corresponding subspace. While the *NN-GPCA* is in perfect alignment with the ground truth error, the performance of *GPCA* degrades heavily as the ratio of the populations in the subspaces increases.

4.5.2 Experiment 2

The second experiment compares the attention detection results of the proposed method and the visual attention model of [1]. In [1], Gaussian pyramids of color, intensity and orientation are created by progressively low-pass filtering and subsampling the input image. Motivated by the “center-surround” operation in visual neurons, the authors generate feature maps by taking the difference in feature values across scales. Normalization of the feature maps is followed by a feature combination strategy that combines the feature maps into a saliency map in which the maximum value corresponds to the most salient image location. Several feature combination strategies are also described by the authors in [2]. Although the saliency map encodes the attention at every location in an image, it is left to a segmentation algorithm to identify the visual attention *region*. Figure 4.12 (a) shows the saliency maps generated from [1] and (b) shows the bounding boxes indicating VARs extracted by segmenting the saliency map using the method described in [86]. Here, the saliency map only indicates the foveats of attention. Figure 4.12 (c) and (d) show, respectively, the candidate attention regions and the smallest bounding boxes that include these regions using the proposed method. The proposed algorithm does not require the VAR to be positioned at the center of the image as is clear from the third and fourth rows of Figure 4.12. The proposed method can detect attention regions of different sizes and also multiple separated attention regions.

4.5.3 Experiment 3

Models for VARs proposed in the literature calculate contrast among regions using *a priori* knowledge about a block size or pyramid level. These limit the ability of the model to detect regions only within the prescribed scale. The proposed method achieves scale invariance through the use of scale-sentient features and the attention measure defined

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

Figure 4.12: Comparison of proposed subspace method with [1] (a) saliency map; (b) VAR from (a); (c) attention region using the proposed method; (d) VAR region from (c).

by the cumulative projection. These can identify regions of any size by considering both inter-region contrast and intra-region homogeneity.

Figure 4.13 shows the VARs extracted from the images which show flowers at different scales. We compare our method to that proposed in [3]. In [3], Ma and Zhang divide an image into 8×8 blocks and estimate the saliency of each block simply by taking the difference in color in HSV space between it and its neighboring blocks. To extract the VARs, they propose a fuzzy growing algorithm to grow the VAR from the block with the highest saliency. Although such a simple model achieves promising performance on

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

Figure 4.13: VAR extraction across different scales: (left) small, (center) medium and (right) large using (i) the proposed method and (ii) method of [3].

some images, the measurement of saliency from the feature differences in fixed block sizes makes the procedure dependent on the scale of the VAR. From the results shown in Figure 4.13 (ii), we can observe that the results are the same for the small scale but as the scale increases, the performance of [3] degrades rapidly.

The polar transformation uses the distance of a pixel to a reference point as the radius in the polar space. The proposed VAR detection method is invariant to the selection of the reference point. In Figure 4.14, we show that the same attention regions are obtained for three different positions of the reference point. It is because that the reference point

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

Figure 4.14: The invariance to the change of reference point (a) original image; attention regions when reference point is at (b) image center (160,120); (c) (35,150); (d) (100,60). (The origin is at the top-left corner.)

Figure 4.15: The invariance to rotation and to illumination change due to change in contrast (a) original images; (b) the detected attention regions; (c) the transformed images after rotating by 45° and reducing contrast by 50%; (d) the detected attention regions in (c).

is used only for measuring the proximity of data points in spatial domain. Wherever the reference point is, the data points corresponding to an image region which are close in spatial domain will have similar distances to the reference point and will form a cluster within some subspace after polar transformation. Different configurations of reference point will preserve the subspace cluster while the shape of the subspace clusters may change. The proposed method is also invariant to rotation and to illumination change

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

Figure 4.16: (a) Original image; (b) attention region and (c) VAR when scale-sentient features are not used; (d) attention region and (e) VAR when scale-sentient features are used.

due to change in contrast of the image. This is illustrated in Figure 4.15 in which the original images (Figure 4.15 (a)) are rotated by 45° and their contrasts are reduced by 50%. The transformed images are shown in Figure 4.15 (c). The attention regions of the original images and the transformed images are shown in Figure 4.15 (b) and Figure 4.15 (d), respectively. We see that the algorithm has accurately determined the attention regions even in transformed images.

The importance of scale-sentient features and its usefulness in VAR extraction is illustrated in Figure 4.16. When a fixed block of size 8×8 is used to compute the mean of the intensity feature, the mask corresponding to the attention region is shown in Figure 4.16 (b) and the inaccurate VAR thus obtained is shown in Figure 4.16 (c). In contrast, using the scale-adaptive selection of block sizes enables the textured region of the leopard and zebra to be observed at the appropriate scale leading to an accurate estimation of the mask corresponding to the attention region as well as the final VAR as shown in Figure 4.16 (d) and 4.16 (e), respectively. The proposed method uses very simple features, viz., mean of hue and intensity to tackle feature extraction at high frequency regions and low frequency regions simultaneously. The entire procedure is data-driven and is not heuristic in nature.

Figure 4.17: Two examples of automatic feature selection.

4.5.4 Experiment 4

The fourth experiment is designed to evaluate the automatic feature selection capability of the proposed method. The CP based measure of (Eq. 4.23) that is calculated for each feature acts as the cue to select the optimal feature for VAR extraction. In the examples shown in Figure 4.17, hue and intensity features are used. In Figure 4.17 (a) the $R_{intensity}$ of the subspaces in the intensity feature is larger than R_{hue} of the subspaces in hue. This is because the interesting object (butterfly) in this case is a textured object and is represented better using intensity information. The proposed attention detection method automatically selects the more suitable intensity feature for textured object. In Figure 4.17 (b), the color of the interesting object manifests its homogeneity in hue channel and the R_{hue} automatically selects hue to extract attention region. We note that R_f facilitates an ordering by priority of attention regions. We select the most attentive region as the final result. The proposed method can alternatively provide a set of regions

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

Figure 4.18: Two failed examples where the original images are shown in (a), the polar transformation and the detected subspaces are shown in (b) and (c), (d) show the wrongly detected VAR and the corresponding bounding box.

with decreasing attention. For instance, the proposed method can indicate two attention regions: a textured region with the highest attention and a non-textured region with the second highest attention for an image containing both kinds of objects. In Figure 4.17 (a), the VAR is not at the image center.

Although the proposed method provides an alternative computational model to detect visual attention regions in images, the underlying assumption is that every image region can manifest its homogeneity in terms of some feature. Such assumption allows an image region to form linear subspace through the polar transformation. We only use the simplest features, viz. color and intensity to represent an image region. We have shown that the proposed model using such simple features can be successful even in challenging images as in Figure 4.12 and Figure 4.16. However in Figure 4.18 we show two failed cases in which homogeneous regions based on the simple features cannot be found; consequently, linear subspaces are not formed in the polar space. We believe that using some advanced features in the proposed model might still result in accurate VAR detection.

The MPEG-21 multimedia framework is envisioned to enable a transparent and aug-

mented use of multimedia resources across a wide range of networks and devices [125]. It aims to provide solutions for Universal Multimedia Access (UMA) whose primary motivation is to enable terminals with limited communication, processing, storage and display capabilities to access rich media content. Part 7 of the MPEG-21 standard deals with Digital Item Adaptation and it addresses requirements with respect to the usage environment and to media adaptability. More specifically, it supports description of resource scaling according to user terminal capabilities. The extraction of VARs can be used to adapt images for small form-factor terminal devices by displaying only the VAR. The proposed algorithm can be used to build a prototype for image adaptation, similar to the one [126] in chapter 3.

4.6 Conclusions

In this chapter, we focus on the development of VAR detection algorithms that can overcome the issues of feature-based methods, according to the object-based attention paradigm. We propose a new algorithm for detection of VAR in images using linear subspace estimation and analysis techniques. We show that very simple features like hue and intensity can achieve this objective by incorporating scale-sentence in them. The polar representation of the features ensures that all pixels within a homogeneous region are mapped into a single linear subspace that also encodes the location of the region within the image. A novel linear subspace estimation method improves the robustness of the *GPCA* algorithm to outliers and subspace skew. The outlier effect is taken care of by weighting the data points according to the distribution of K nearest neighbors and the subspaces are estimated using a weighted least squares technique. A further iterative refinement is carried out to deal with subspace skew. A new attention measure for image regions is proposed that considers the contrast among regions measured by the cumulative projection of all data onto individual subspaces and also considers the

CHAPTER 4. SUBSPACE ANALYSIS FOR DETECTION OF VARs IN IMAGES

clustering of data within a subspace. Automatic feature selection for VAR extraction is achieved from the analysis of attention values of all subspaces in each feature. The experiments on both synthetic and real data are performed to evaluate the performance of the proposed method. The experimental results show that the proposed method offers several advantages such as the robustness to outliers and scale invariance.

Chapter 5

Attention-from-Motion

5.1 Introduction

In this chapter, we investigate visual attention about the motion in videos. The additional information about motion in a video sequence provides a strong cue for VAR, instead of other features e.g. color, intensity as well as texture. Research in cognition and neuroscience has shown that the primate visual system can selectively focus attention on some specific motion pattern in a scene, while ignoring other motion patterns in it [127, 128]. Detection of motion that captures visual attention is useful in several applications like video surveillance [129], video summarization [130], and video editing [131].

Previous research on detecting motion patterns that capture visual attention use heuristic notions of salient motion. Williams and Draper [132] studied the effect of motion on visual attention by adding the motion channel to a saliency map that encodes the attention value of every pixel. They concluded that motion contrast cannot improve the performance of saliency-based selective attention. This shows that using information about contrast that appears as a result of motion is not a good cue to find the "interesting" object in the video sequence. Ma et al. [130] used motion vector fields to compute the motion energy to identify salient motion regions. Tian and Hampapur [133] discriminated salient motion from unimportant motion using a consistency measure.

CHAPTER 5. ATTENTION-FROM-MOTION

Wixson [134] also used a consistency measure to detect interesting motion by accumulating directionally-consistent flow. In all the above, regions that attract visual attention due to motion are defined heuristically, e.g. based on features like motion energy and motion consistency. Such definitions are not applicable in many instances, for example, even if motions of all objects are consistent, only one of them might invite visual attention. Similarly, an object with low motion energy can also be the one on which visual attention is centered. Moreover, these features are also not robust to noise. In [135], Horaud et al. developed a framework in which a static camera is cooperated with an active camera to detect all moving objects. Here, there is no discrimination among detected objects in terms of visual attention. In [136], López et al. extracted motion features like motion presence, module and angle of velocity to segment the moving objects and the user is invited to specify the objects that capture visual attention. The extraction of interesting object in this case is not automatic. Other motion analysis algorithms like [137, 138] ignore the concept of visual attention. In this chapter, we introduce the notion of **Attention-from-Motion** in which the objective is to identify, from a video, only those object in motions that capture visual attention (VA). Following the important concept in film production, viz, the tracking shot, we define the attention object in motion (AOM) as those that are tracked by the camera.

5.1.1 Attention-from-Motion

Video production manuals provide insight into the procedures used during video creation and editing. One of the most important shots is the “tracking shot” in which the camera follows one or more moving objects [139]. Naturally, the intention of the cameraman is to focus the viewer’s attention on the object(s) which he is following. Thus, we define an **Attention Object in Motion (AOM)** as an object that is followed by the camera. As illustrated in Figure 5.1, although there are two cars in the field of view of the camera,

CHAPTER 5. ATTENTION-FROM-MOTION

Figure 5.1: Illustration of Attention Object in Motion (the dashed square indicates the field of view of the camera). The camera follows the car on the top-left (AOM) while ignoring the car on the bottom-right (not AOM).

it follows only one of them, so that the car on the top-left is an AOM while the one on the bottom-right is not. The notion of “camera following an object” embeds the subjective attention of the cameraman and will eventually guide the attention of the viewers. This definition of AOM is applicable for videos produced professionally as well as for videos taken by general consumers of hand-held video cameras. We define the problem of **attention-from-motion** as detecting AOMs from the motion information in a video sequence. The motion information is provided as a measurement matrix of P feature points across F frames. The definition of AOM embeds the “following” relationship between camera motion and object motion.

In this chapter, Three components are proposed to form an attention-from-motion framework: (i) a new factorization form of the measurement matrix to describe dynamic geometry of moving object observed by moving camera is developed in Section 5.2. This factorization is derived under the assumption of orthographic projection and rigid body motion; (ii) determination of single AOM based on the analysis of certain structure on the motion matrix is introduced in Section 5.3. We propose a method to identify this structure *without carrying out the complete factorization*; (iii) an iterative framework for

detecting multiple AOMs is proposed to robustly extract more than one AOM in Section 5.4. The proposed solution to the attention-from-motion problem is not constrained by the type of camera/object motion including degenerate as well as dependent motion. Furthermore, the algorithm can detect AOMs even when only partial data is available, e.g. due to occlusion or over-segmentation.

5.2 Factorization Method for AOM

Factorization methods have been used in [140, 141] to solve the structure-from-motion problem. In [141], a moving object is observed by a static camera while in [140] a static object is observed by a moving camera. We generalize the above two methods to consider the case of a moving camera observing an object in motion under the assumptions of orthographic projection and rigid body motion. The factorization results in a motion matrix and a shape matrix such that the motion matrix corresponding to an AOM has a special structure.

5.2.1 Dynamic Scene Factorization

The *measurement matrix* W consists of a $F \times P$ submatrix X of horizontal feature coordinates x_{fp} , with F being the number of frames and P being the number of feature points and a similar $F \times P$ submatrix Y of vertical coordinates y_{fp} . Thus $W = \begin{bmatrix} X \\ Y \end{bmatrix}$. According to rigid-body motion assumption, the 3-D coordinates of the p^{th} feature point on an object at the f^{th} frame is

$$s_{fp} = R_{o,f}s_{1p} + T_{o,f}, \quad p = 1, \dots, P \text{ and } f = 1, \dots, F, \quad (\text{Eq. 5.1})$$

where $R_{o,f}$ is the rotation matrix and $T_{o,f}$ is the translation vector of the object at the f^{th} frame with respect to the world coordinates. By placing the origin of the world

CHAPTER 5. ATTENTION-FROM-MOTION

coordinates at the centroid of object feature points at the 1st frame, we have

$$\frac{1}{P} \sum_{p=1}^P s_{1p} = 0. \quad (\text{Eq. 5.2})$$

Combining (Eq. 5.2) with (Eq. 5.1), we get

$$\frac{1}{P} \sum_{p=1}^P s_{fp} = \frac{1}{P} \sum_{p=1}^P (R_{o,f} s_{1p} + T_{o,f}) = T_{o,f}. \quad (\text{Eq. 5.3})$$

Under orthography, the image feature position (x_{fp}, y_{fp}) of point s_{fp} at frame f is given by the equations [140]

$$x_{fp} = i_f^T (s_{fp} - T_{c,f}), \quad y_{fp} = j_f^T (s_{fp} - T_{c,f}), \quad (\text{Eq. 5.4})$$

where, i_f and j_f are the unit vectors in frame f pointing along the rows and columns, respectively, of the image and defined w.r.t the world reference system, and $T_{c,f}$ is the translation vector of the camera w.r.t the world coordinates. If $R_{c,f}$ is the rotation matrix of the camera w.r.t. the world coordinates, then

$$\begin{aligned} i_f &= R_{c,f} i_0 + T_{c,f} - T_{c,f} = R_{c,f} i_0 \\ j_f &= R_{c,f} j_0 + T_{c,f} - T_{c,f} = R_{c,f} j_0, \end{aligned} \quad (\text{Eq. 5.5})$$

where $i_0 = [1, 0, 0]^T$ and $j_0 = [0, 1, 0]^T$ are the two axes of the canonical world coordinate system as shown in Figure 5.2. Any other representation of i_0 and j_0 will not affect the subsequent analysis.

Using (Eq. 5.1) and (Eq. 5.5), the entries of X can be derived as

$$\begin{aligned} x_{fp} &= i_f^T (s_{fp} - T_{c,f}) \\ &= i_0^T R_{c,f}^T R_{o,f} s_{1p} + i_0^T R_{c,f}^T (T_{o,f} - T_{c,f}). \end{aligned} \quad (\text{Eq. 5.6})$$

Similarly, the entries of Y can also be derived as

$$y_{fp} = j_0^T R_{c,f}^T R_{o,f} s_{1p} + j_0^T R_{c,f}^T (T_{o,f} - T_{c,f}). \quad (\text{Eq. 5.7})$$

CHAPTER 5. ATTENTION-FROM-MOTION

Figure 5.2: The reference coordinate system used in the factorization of dynamic scene.

The above two sets of $F \times P$ equations ((Eq. 5.6) and (Eq. 5.7)) form the unregistered measurement matrix W for the case of a moving camera observing a moving object. The matrix W can be factorized into $W = M \times S$ as

$$W = \begin{bmatrix} X \\ Y \end{bmatrix} = \underbrace{\begin{bmatrix} i_0^T R_{c,1}^T R_{o,1} & i_0^T R_{c,1}^T (T_{o,1} - T_{c,1}) \\ i_0^T R_{c,2}^T R_{o,2} & i_0^T R_{c,2}^T (T_{o,2} - T_{c,2}) \\ \vdots & \vdots \\ i_0^T R_{c,F}^T R_{o,F} & i_0^T R_{c,F}^T (T_{o,F} - T_{c,F}) \\ j_0^T R_{c,1}^T R_{o,1} & j_0^T R_{c,1}^T (T_{o,1} - T_{c,1}) \\ j_0^T R_{c,2}^T R_{o,2} & j_0^T R_{c,2}^T (T_{o,2} - T_{c,2}) \\ \vdots & \vdots \\ j_0^T R_{c,F}^T R_{o,F} & j_0^T R_{c,F}^T (T_{o,F} - T_{c,F}) \end{bmatrix}}_{M(2F \times 4)} \underbrace{\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1P} \\ 1 & 1 & \cdots & 1 \end{bmatrix}}_{S(4 \times P)} \quad (\text{Eq. 5.8})$$

The matrix M in (Eq. 5.8) contains the motion of both camera and object while in [141] and [140], M contains only object motion and only camera motion, respectively. Since M is $2F \times 4$ and S is $4 \times P$, (Eq. 5.8) yields the **Rank Theorem** that states that *without noise, the measurement matrix W of a moving object observed by a moving camera is at most of rank four.*

5.2.2 Factorization of AOM

Recall that we define an AOM as an object in motion that is followed by a moving camera. The fact that the moving object is followed by the moving camera implies that, in the absence of noise,

$$\frac{1}{P} \sum_{p=1}^P x_{fp} = \frac{1}{P} \sum_{p=1}^P x_{1p}, \quad \forall f \in \{1, \dots, F\} \quad (\text{Eq. 5.9})$$

and

$$\frac{1}{P} \sum_{p=1}^P y_{fp} = \frac{1}{P} \sum_{p=1}^P y_{1p}, \quad \forall f \in \{1, \dots, F\}. \quad (\text{Eq. 5.10})$$

If this condition is true over a small number of frames, but not true in the long term, then according to our definition, the feature points belong to an AOM initially, but it is no longer an AOM when the conditions in (Eq. 5.9) and (Eq. 5.10) fail. Using equations (Eq. 5.1), (Eq. 5.4) and (Eq. 5.5), (Eq. 5.9) can be rewritten as

$$\begin{aligned} & \frac{1}{P} \sum_{p=1}^P i_f^T (s_{fp} - T_{c,f}) = \frac{1}{P} \sum_{p=1}^P i_1^T (s_{1p} - T_{c,1}) \\ \Rightarrow & \frac{1}{P} \sum_{p=1}^P i_f^T (R_{o,f} s_{1p} + T_{o,f} - T_{c,f}) = \frac{1}{P} \sum_{p=1}^P i_1^T (s_{1p} - T_{c,1}) \\ \Rightarrow & \frac{1}{P} \sum_{p=1}^P i_0^T R_{c,f}^T (R_{o,f} s_{1p} + T_{o,f} - T_{c,f}) = \frac{1}{P} \sum_{p=1}^P i_0^T R_{c,1}^T (s_{1p} - T_{c,1}) \\ \Rightarrow & i_0^T R_{c,f}^T R_{o,f} \underbrace{\frac{1}{P} \sum_{p=1}^P s_{1p}}_{=0} + i_0^T R_{c,f}^T (T_{o,f} - T_{c,f}) = i_0^T R_{c,1}^T \underbrace{\frac{1}{P} \sum_{p=1}^P s_{1p}}_{=0} - i_0^T R_{c,1}^T T_{c,1} \\ \Rightarrow & i_0^T R_{c,f}^T (T_{o,f} - T_{c,f}) = -i_0^T R_{c,1}^T T_{c,1}, \quad \forall f \in \{1, \dots, F\} \end{aligned} \quad (\text{Eq. 5.11})$$

Similarly, it can be shown that

$$j_0^T R_{c,f}^T (T_{o,f} - T_{c,f}) = -j_0^T R_{c,1}^T T_{c,1}, \quad \forall f \in \{1, \dots, F\} \quad (\text{Eq. 5.12})$$

Interestingly, the left sides of (Eq. 5.11) and (Eq. 5.12) are the entries in the last column of M in (Eq. 5.8). Hence, for an AOM, the motion matrix M can be represented

CHAPTER 5. ATTENTION-FROM-MOTION

as:

$$M = \begin{bmatrix} i_0^T R_{c,1}^T R_{o,1} & -i_0^T R_{c,1}^T T_{c,1} \\ i_0^T R_{c,2}^T R_{o,2} & -i_0^T R_{c,1}^T T_{c,1} \\ \vdots & \vdots \\ i_0^T R_{c,F}^T R_{o,F} & -i_0^T R_{c,1}^T T_{c,1} \\ j_0^T R_{c,1}^T R_{o,1} & -j_0^T R_{c,1}^T T_{c,1} \\ j_0^T R_{c,2}^T R_{o,2} & -j_0^T R_{c,1}^T T_{c,1} \\ \vdots & \vdots \\ j_0^T R_{c,F}^T R_{o,F} & -j_0^T R_{c,1}^T T_{c,1} \end{bmatrix}. \quad (\text{Eq. 5.13})$$

In the above $2F \times 4$ matrix, the first 3 columns relate to the rotation component and the 4th column relates to the translation component of the dynamic scene. This structure allows us to impose the following motion constraints:

- **Rotation Constraint:** Each pair of rows in the first 3 columns of M are the rotations of two reference orthonormal basis (e.g. $[1, 0, 0]^T$ and $[0, 1, 0]^T$) of M since $R_{c,f}^T R_{o,f}$ is a rotation matrix such that $i_0^T R_{c,f}^T R_{o,f}$ and $j_0^T R_{c,f}^T R_{o,f}$ are the rotations of i_0 and j_0 , respectively.
- **Translation Constraint:** The elements of the first half of the last column of M are constants and so also are the elements of the second half of the last column.

5.2.3 Why not simply check for stationary centroid?

Since the special structure of M is derived from the property of a stable centroid of *all* the feature points belonging to an AOM, it would seem that checking for the stationarity of the centroid would suffice to identify the AOM. However, this would imply that all the feature points are available to cover the whole object, otherwise the centroid would not be stable. This is a very strong constraint because feature points could be lost for a variety of reasons like occlusion and change in appearance of the object due to motion or change in illumination. Moreover, the limitations of the feature detection algorithm itself may cause only some of the feature points on an object to be detected resulting

CHAPTER 5. ATTENTION-FROM-MOTION

Figure 5.3: An example of unstable centroid of AOM where only partial feature points on an AOM are observed.

in the tracking of an erroneous centroid. Figure 5.3 illustrates this problem in which four frames of a sequence show a rotating rectangle whose four corners are the feature points. Since the centroid of the feature points indicated by the black cross is at the same location, the rectangle is an AOM. However, if the feature detection algorithm returns only two feature points (red and yellow) then its centroid as indicated by the pink cross is not stationary.

Another case when checking for the stationarity of the centroid of feature points fails is that of an object revolving about a point. In this case, the cameraman will most likely wish to capture the entire circular motion by focusing on the centroid of the motion rather than on the centroid of the feature points on the object. Figure 5.4 shows a few frames of this type of motion in which the ball is revolving about the point marked +; here the centroid of the feature points changes across all frames, but the camera will focus on the +. Furthermore, one would recognize the motion of the object as a circular motion about a point only if all the frames of at least one cycle are observable. If only some of the frames of the entire cycle are available, as shown in Figure 5.4, it is not possible to recognize the circular motion because the centroid of the motion will not be the point about which the object rotates.

The structure of M is valid in all of the above cases. For the case when only partial feature points are available, only a subset of feature points are required to calculate M

CHAPTER 5. ATTENTION-FROM-MOTION

Figure 5.4: Another example of unstable centroid of AOM where the centroid of motion is followed and only partial frames are observed.

and to analyze its structure to determine if they belong to an AOM. This is because it is enough to obtain the matrix M from the measurement matrix corresponding to a subset of feature points as long as its rank is equal to the rank of the measurement matrix corresponding to the full set of feature points. Also, in our analysis, the requirement that the last column of M is constant is valid independent of the number of frames observed. This makes the proposed method robust to occlusion and avoids the propagation of tracking error over long intervals. The structure of M is valid, even when the centroid of the moving object is not stationary (e.g. in Figure 5.4) because the AOM can be considered as a part of another object whose centroid is the centroid of motion. Thus, the proposed method works even if only partial data is available.

5.3 Attention-from-Motion for a single AOM

In [140], it is shown that the factorization of the measurement matrix W using *singular value decomposition (SVD)* gives

$$W = U\Sigma V^T = (U\Sigma^{\frac{1}{2}}) \cdot (\Sigma^{\frac{1}{2}}V^T) = \hat{M}\hat{S}. \quad (\text{Eq. 5.14})$$

This factorization is not unique since for any 4×4 **invertible** matrix Q , the matrices $\hat{M}Q$ and $Q^{-1}\hat{S}$ are also a valid factorization of W , i.e.,

$$(\hat{M}Q)(Q^{-1}\hat{S}) = \hat{M}(QQ^{-1})\hat{S} = \hat{M}\hat{S} = W. \quad (\text{Eq. 5.15})$$

CHAPTER 5. ATTENTION-FROM-MOTION

In the following, we describe a linear method to determine the existence and singularity of Q only from the affine version $\hat{M}$ of M . If there exists a Q that transforms $\hat{M}$ to M (of Equation (Eq. 5.13)), then the feature points in the measurement matrix W belong to an AOM. The result is then used to define an attention measure to identify an AOM.

5.3.1 Constraints on $\hat{M}$ and Q

By denoting the matrix Q as $Q = [Q_R|Q_T]$ where Q_R contains the first 3 columns of Q , and Q_T contains the last column of Q , M can be represented as

$$M = \hat{M}Q = [\hat{M}Q_R|\hat{M}Q_T]. \quad (\text{Eq. 5.16})$$

$\hat{M}Q_R$ corresponds to the first 3 columns of M which satisfies the rotation constraint and $\hat{M}Q_T$ corresponds to the last column of M which satisfies the translation constraint.

5.3.1.1 Rotation Constraint on $\hat{M}$ and Q

Each of the $2F$ rows of matrix $\hat{M}Q_R$ is a unit norm vector and the first and second set of F rows are pairwise orthogonal. This orthonormality can be represented as

$$\begin{aligned} \hat{M}_i Q_R Q_R^T \hat{M}_i^T &= 1, \\ \hat{M}_j Q_R Q_R^T \hat{M}_j^T &= 1, \\ \hat{M}_i Q_R Q_R^T \hat{M}_j^T &= 0, \end{aligned} \quad (\text{Eq. 5.17})$$

for $i = 1 \dots F$, $j = F + i$ and $\hat{M}_i$ denotes row i of $\hat{M}$. The set of equations in (Eq. 5.17) can be solved for Q_R by adding the additional constraints of the reference coordinate system (e.g., $i_0 = [1, 0, 0]^T$ and $j_0 = [0, 1, 0]^T$) to it. Instead, we solve for the entries of $Q_R Q_R^T$, represented as

$$Q_R Q_R^T = \begin{bmatrix} Q_1 & Q_5 & Q_6 & Q_7 \\ Q_5 & Q_2 & Q_8 & Q_9 \\ Q_6 & Q_8 & Q_3 & Q_{10} \\ Q_7 & Q_9 & Q_{10} & Q_4 \end{bmatrix}$$

CHAPTER 5. ATTENTION-FROM-MOTION

and whose elements are

$$\begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \\ Q_5 \\ Q_6 \\ Q_7 \\ Q_8 \\ Q_9 \\ Q_{10} \end{bmatrix} = \begin{bmatrix} Q_{11}^2 + Q_{12}^2 + Q_{13}^2 \\ Q_{21}^2 + Q_{22}^2 + Q_{23}^2 \\ Q_{31}^2 + Q_{32}^2 + Q_{33}^2 \\ Q_{41}^2 + Q_{42}^2 + Q_{43}^2 \\ Q_{11}Q_{21} + Q_{12}Q_{22} + Q_{13}Q_{23} \\ Q_{11}Q_{31} + Q_{12}Q_{32} + Q_{13}Q_{33} \\ Q_{11}Q_{41} + Q_{12}Q_{42} + Q_{13}Q_{43} \\ Q_{21}Q_{31} + Q_{22}Q_{32} + Q_{23}Q_{33} \\ Q_{21}Q_{41} + Q_{22}Q_{42} + Q_{23}Q_{43} \\ Q_{31}Q_{41} + Q_{32}Q_{42} + Q_{33}Q_{43} \end{bmatrix} \quad (\text{Eq. 5.18})$$

by regarding (Eq. 5.17) as a linear system about the 10 unique entries of $Q_R Q_R^T$. If the linear system of (Eq. 5.17) has a unique solution, we can directly obtain Q (the left hand side of (Eq. 5.18)) using least squares technique. Note that unlike in [141, 140] where Q is reconstructed after solving for $Q_R Q_R^T$ in (Eq. 5.17), *we do not need to reconstruct Q for carrying out attention analysis, as shown below.*

5.3.1.2 Translation Constraint on $\hat{M}$ and Q

The fact that two halves of the last column of M are constants implies that $\hat{M}_1 Q_T = \hat{M}_2 Q_T = \dots = \hat{M}_F Q_T$ and $\hat{M}_{F+1} Q_T = \hat{M}_{F+2} Q_T = \dots = \hat{M}_{2F} Q_T$. This constraint can be further transformed into a homogeneous linear system as

$$\underbrace{\begin{bmatrix} \hat{M}_1 - \hat{M}_2 \\ \vdots \\ \hat{M}_{F-1} - \hat{M}_F \\ \hat{M}_{F+1} - \hat{M}_{F+2} \\ \vdots \\ \hat{M}_{2F-1} - \hat{M}_{2F} \end{bmatrix}}_{P_T((2F-2) \times 4)} \underbrace{\begin{bmatrix} Q_{1,4} \\ Q_{2,4} \\ Q_{3,4} \\ Q_{4,4} \end{bmatrix}}_{Q_T} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (\text{Eq. 5.19})$$

For the existence of a nonzero Q_T , the rank of P_T should be smaller than 4 such that its null space is not empty. If $\text{rank}(P_T) = 3$, we can uniquely obtain a normalized version $\bar{Q}_T$ of Q_T by setting $\bar{Q}_{1,4} = 1$. If $\text{rank}(P_T) < 3$, then there is no unique solution. As we

CHAPTER 5. ATTENTION-FROM-MOTION

will see in the next section, the singularity of Q will have to be measured to determine a final attention measure for an AOM. This singularity is based on the linear dependence of rows/columns of Q . Since linear dependence is invariant to scale, the normalized version $\bar{Q}_T$ is enough to check for the singularity of Q . The trivial solution $[0, 0, 0, 0]^T$ is invalid because it corresponds to constant 0 of the last column of Q which is not allowable because that would make Q non-invertible and therefore, singular.

5.3.2 Attention Measure from $\bar{Q}_T$ and Q

Whether a group of feature points belongs to an AOM depends on whether there *exists* an *invertible* Q that satisfies the two constraints described in the previous section. We show that both the existence and singularity of Q can be analyzed from $\bar{Q}_T$ and Q , *without calculating* Q . In fact, for degenerate cases such as pure translation and pure rotation, it is not possible to calculate Q .

5.3.2.1 Existence of Q from $\bar{Q}_T$

In the absence of noise, if the feature points belong to an AOM, then there exists a solution for the system of equations in (Eq. 5.19) implying the existence of a $\bar{Q}_T$ that transforms $\hat{M}$ to a column consisting of two constant parts. However, it is not assured that the column will indeed be the last column of (Eq. 5.16). If the column is one of the first 3 columns of Q , $\bar{Q}_T$ will be linearly dependent with one column of Q_R and, consequently, Q will be singular, in which case we need to check for the singularity of Q . In the case of noisy data, a solution $\bar{Q}_T$ of (Eq. 5.19) always exists in the least square sense. Due to its sensitivity to noise, use of the smallest *SVD* value of P_T in (Eq. 5.19) to measure the existence of $\bar{Q}_T$ is not robust.

To overcome this problem, we propose a measure to check for the existence of $\bar{Q}_T$ based on the variance of $\tilde{Q}_T = \hat{M} \times \bar{Q}_T$. This measure consists of two factors. The first

CHAPTER 5. ATTENTION-FROM-MOTION

factor measures the variances of the first F rows and the last F rows of $\tilde{Q}_T$, i.e.,

$$M_c(i) = \text{var}(\tilde{Q}_T^i), \quad i = 1, 2, \quad (\text{Eq. 5.20})$$

where $\tilde{Q}_T^1$ and $\tilde{Q}_T^2$ are the first and second halves of $\tilde{Q}_T$, respectively. The second factor is related to the randomness in the change of $\tilde{Q}_T$. The change can be caused due to unstable following of the object by the camera or due to the feature points moving in a particular direction not being followed. The former causes more randomness in the values of $\tilde{Q}_T$ than the latter; this trait is captured in the following measure

$$R_c(i) = \frac{\max(\tilde{Q}_T^i) - \min(\tilde{Q}_T^i)}{\sum_{k=2}^F |\tilde{Q}_T^i(k) - \tilde{Q}_T^i(k-1)|}. \quad (\text{Eq. 5.21})$$

R_c is closer to 1 when the randomness of the variation in $\tilde{Q}_T$ is small. These two factors are combined to obtain the final measure S_q computed as

$$S_q = \max(R_c(1) \cdot M_c(1), R_c(2) \cdot M_c(2)). \quad (\text{Eq. 5.22})$$

If S_q is small, then $\tilde{Q}_T$ is more close to constant. Otherwise, it is more possible that $\tilde{Q}_T$ is not constant column.

5.3.2.2 Singularity of Q from $\bar{Q}_T$ and $\mathcal{Q}$

The singularity of Q can be indicated by the linear dependence in either its rows or its columns. In this section, we show that the linear dependency can be established by analyzing only $\bar{Q}_T$ and $\mathcal{Q}$.

In the case of linear dependence in rows or in columns of Q , we have

$$\text{rank}([Q_R | \bar{Q}_T]) = \text{rank}(Q) < 4. \quad (\text{Eq. 5.23})$$

Since $[Q_R | \bar{Q}_T]^T$ has the same rank as $[Q_R | \bar{Q}_T]$, it is also true that

$$\text{rank}([Q_R | \bar{Q}_T][Q_R | \bar{Q}_T]^T) = \text{rank}([Q_R | \bar{Q}_T]) < 4 \quad (\text{Eq. 5.24})$$

CHAPTER 5. ATTENTION-FROM-MOTION

The matrix $[Q_R|\bar{Q}_T][Q_R|\bar{Q}_T]^T$ can be calculated directly from Q and $\bar{Q}_T$ without knowing Q , since

$$[Q_R|\bar{Q}_T][Q_R|\bar{Q}_T]^T = Q_R Q_R^T + \bar{Q}_T \bar{Q}_T^T = \begin{bmatrix} Q_1 & Q_5 & Q_6 & Q_7 \\ Q_5 & Q_2 & Q_8 & Q_9 \\ Q_6 & Q_8 & Q_3 & Q_{10} \\ Q_7 & Q_9 & Q_{10} & Q_4 \end{bmatrix} + \bar{Q}_T \bar{Q}_T^T \quad (\text{Eq. 5.25})$$

A measure of linear dependence of Q , denoted by L_q , can then be determined as $L_q = \sigma_{\min}$ where $\sigma_{\min}$ is the smallest singular value of $[Q_R|\bar{Q}_T][Q_R|\bar{Q}_T]^T$. For noisy case, we improve the robustness of this measure by considering the ratio of the two smallest singular values.

Degenerate Cases

Here we analyze the existence and singularity of Q for degenerate cases, i.e., the case in which the rank of the measurement matrix is less than 4. Examples of degenerate motions are pure rotation and pure translation. Equation (Eq. 5.18) can be divided into 6 homogeneous linear systems in the following form:

$$\begin{bmatrix} Q_{i1}^2 + Q_{i2}^2 + Q_{i3}^2 \\ Q_{j1}^2 + Q_{j2}^2 + Q_{j3}^2 \\ Q_{i1}Q_{j1} + Q_{i2}Q_{j2} + Q_{i3}Q_{j3} \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix} \quad (\text{Eq. 5.26})$$

where $i \neq j \in \{1, 2, 3, 4\}$ and C_1, C_2 and C_3 are corresponding entities in Q , e.g., Q_1, Q_2 and Q_5 form one such linear system. Each linear system relates 2 of the 4 rows of Q_R . For degenerate cases, only some of the rows of Q_R can be determined. The singularity of Q can still be indicated by the linear dependence among the available rows. We detect such linear dependence in degenerate cases by analyzing only $\bar{Q}_T$ and Q . For example, if only 3 rows of Q_R can be determined for degenerate motion, then the rotation constraint on $\hat{M}$ gives

$$\begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \\ Q_5 \\ Q_6 \end{bmatrix} = \begin{bmatrix} Q_{11}^2 + Q_{12}^2 + Q_{13}^2 \\ Q_{21}^2 + Q_{22}^2 + Q_{23}^2 \\ Q_{31}^2 + Q_{32}^2 + Q_{33}^2 \\ Q_{11}Q_{21} + Q_{12}Q_{22} + Q_{13}Q_{23} \\ Q_{11}Q_{31} + Q_{12}Q_{32} + Q_{13}Q_{33} \\ Q_{21}Q_{31} + Q_{22}Q_{32} + Q_{23}Q_{33} \end{bmatrix} \quad (\text{Eq. 5.27})$$

If we denote the matrix containing 3 rows of Q_R as Q_A and the corresponding three entries of $\bar{Q}_T$ as $\bar{Q}_{TA}$, we have the following property for singular Q :

$$\text{rank}([Q_A|\bar{Q}_{TA}][Q_A|\bar{Q}_{TA}]^T) = \text{rank}([Q_A|\bar{Q}_{TA}]) < 3 \quad (\text{Eq. 5.28})$$

As before, the matrix $[Q_A|\bar{Q}_{TA}][Q_A|\bar{Q}_{TA}]^T$ can be calculated from available Q_i in (Eq. 5.27) and the corresponding entries in $\bar{Q}_T$. So the linear dependence of Q can still be measured from the singular values of $[Q_A|\bar{Q}_{TA}][Q_A|\bar{Q}_{TA}]^T$.

5.3.3 Integrated Attention Measure

Finally, we combine both measures S_q and L_q to obtain an integrated measure to determine if a collection of feature points belongs to an AOM. We estimate the attention of a moving object as

$$\text{Attention} = L_q/S_q. \quad (\text{Eq. 5.29})$$

Since S_q measures the variance as well as the randomness of the last column of M , a small value of S_q indicates how constant the last column of M is. L_q is the smallest singular value of $[Q_R|\bar{Q}_T][Q_R|\bar{Q}_T]^T$ that indirectly indicates the singularity of Q . If L_q is close to 0, the matrix Q is singular (non-invertible). When L_q is large and S_q is small, the feature points are more likely to belong to AOM(s) and their attention values should be large. Hence, we use the ratio L_q/S_q as the measure of attention. If *Attention* is greater than a threshold, then the feature points whose coordinates form the measurement matrix W belong to an AOM. In this paper, the threshold value is decided by Otsu thresholding method [109] for all experiments. Other advanced measures can also be designed based on S_q and L_q . Because the only related operation of the proposed attention measure is about solving a small-scale linear system, the computational complexity of it is extremely low. If there are multiple solutions for Q_T in (Eq. 5.19) or for $Q_R Q_R^T$ in (Eq. 5.17), then we apply the above analysis individually on X and Y . If there are still multiple solutions, then we can conclude that the feature points belong to an AOM.

5.4 Attention-from-motion for multiple AOMs

In the previous section, we have presented the theory for detecting a single AOM in the video sequence. Clearly, in some situations, there could be multiple moving objects and the measurement matrix would contain feature points from all of them. We develop an iterative framework to detect multiple AOMs, in which motion segmentation is performed. Since the proposed method to detect single AOM can handle partial data, we allow over-segmentation of the object, i.e., feature points can belong to two or more groups. The proposed framework can robustly detect multiple AOM(s) for partial dependent motion [142] as well as for degenerate motion.

5.4.1 Shape Space Initialization

As noted in (Eq. 5.14), the SVD of W is carried out as $W = U\Sigma V^T$. We denote the space formed by the basis of V as *shape space*. It encodes the information in the row space of the measurement matrix W . In [141], multibody segmentation is performed through the shape interaction matrix $A = VV^T$.

In our algorithm, initially, each feature point constitutes a group and the interaction of each point with all the other points are computed as follows. At each iteration, the feature point with the maximum interaction, computed as the sum of interaction with the points already in the group, is appended to the group. This procedure is continued till all the feature points are included in a group. The above process of expanding a group is performed on all the P groups. This means that every group will have all the feature points in decreasing order of interaction. Based on the ordering, we hope that points having strong interaction belong to the same object. But, such sorting could be erroneous since two points i and j might belong to the same object even if $A_{i,j} = 0$. However, it is very likely that the first 4 points on a majority of the sorted groups belong to the same object. The group is expanded further based on motion information (next section) to

determine if there are other feature points belonging to this group. Instead of relying on the block-diagonal structure of canonical A for segmentation [141], which could degrade when there is even small amount of noise or when there is partial dependent motion [142], we only extract the first 4 points from every sorted list according to the rank theorem of dynamic scene.

Although we cannot determine whether two points with small interaction belong to same object or not, if there is large interaction between them, it is very likely that they belong to the same object. That is the reason why *majority* of the 4-tuple are correct. However, some of the 4-tuple may be still erroneous due to accidental interaction between points on different objects. A voting scheme will select the most reliable grouping with maximum support in the final stage, making the framework robust to such error in grouping.

5.4.2 Motion Space Expansion

For grouping feature points having similar motion, it is imperative to use motion information in addition to using shape information. Indeed, if two points have similar motion, they must be grouped together irrespective of their 3D coordinates. In this section, we expand the list of feature points previously obtained using shape information by adding points that have similar motion.

The similarity of motion among feature points is computed as follows. The matrix U obtained from the SVD decomposition of the measurement matrix W forms a basis for the column space of W , which is called the *motion space*. From the factorization analysis in Section 5.2.1, we know that the motion space of a moving object observed by a moving camera is of at most rank 4. Hence, we carry out the SVD on the measurement matrix W of the 4-tuple (obtained from the previous section) to determine the motion space of the corresponding object. We compute the projection of every column of W (there

CHAPTER 5. ATTENTION-FROM-MOTION

are P columns) onto the basis of this motion space (projection denoted as L) as well as orthogonal to it (projection denoted as H). The ratio $\frac{H}{L}$ indicates the similarity between the feature point for that column and the 4-tuple. If the two motions are similar, $\frac{H}{L}$ will be small; in the absence of noise, it will be 0 if the motions are exactly the same. The initial 4-tuple is expanded by appending to it, in increasing $\frac{H}{L}$ ratios, those feature points that satisfy the rank theorem of Section 5.2.1. The expanded list is then split into two, based on a thresholding mechanism such that the feature points on the top of the list are retained as candidates for voting. The threshold is chosen as that feature point which has the maximum difference of the ratio $\frac{H}{L}$ from its neighbor.

All the 4-tuples from the shape space initialization are expanded according to the above procedure. In order to ensure that, for each group, the measurement matrix corresponding to the expanded list of feature points is not rank deficient, we develop an iterative technique by which the list is expanded. Specifically, we repeat the same process as above starting from the SVD of the measurement matrix up to finding a new list of feature points after thresholding. If the new group is the same as the old one, iteration is stopped, otherwise a new measurement matrix for the new group is formed and the process is repeated.

In order to accommodate for possible errors in the groups, we assign a confidence measure to each of them. The confidence measure for a group G is defined as

$$C_G = \frac{\min_{i \notin G} (H(i)/L(i))}{\sum_{j \in G} (H(j)/L(j))}, \quad (\text{Eq. 5.30})$$

where $L(i)$ and $H(i)$ are the projections of the column in the measurement matrix W corresponding to feature point i , onto the motion space and orthogonal to it, respectively. When there is no clear difference among all $\frac{H}{L}$ ratios, it is not possible to obtain a good threshold. This could possibly lead to an erroneous group, whose confidence measure C_G will be small. Large confidence measure C_G indicates that the threshold can be easily found and the group is more likely to be correct.

CHAPTER 5. ATTENTION-FROM-MOTION

By combining both shape space as well as motion space, the proposed method for motion segmentation is more robust to previous methods, which either consider shape space or motion space. Compared to the methods using shape space only [141, 143], and the cluster-and-test approaches using motion space only [137, 138], forming the 4-tuple from shape space and relying on motion space expansion, provides a systematic way to define the motion space, overcoming major disadvantages of previous methods. First, the 4-tuple with largest interaction values has more confidence to belong to the same object than random selection of a 4-tuple. The incorporation of confidence measure increases the accuracy of the motion space and allows errors in the shape interaction matrix. Second, the iterative motion space expansion and voting scheme (introduced in next section) can improve the robustness of motion segmentation when some groups of feature points are erroneous. Third, we iteratively extract the most reliable grouping one by one and avoid estimating the number of objects. Note that the degenerate motion space (e.g. low-rank subspace of motion space) does not affect the proposed method since we allow for over-segmentation of the object. Interestingly, the case of partially dependent motion as enunciated in [142] is also handled efficiently using the proposed method. For two objects o_1 and o_2 having partially dependent motion, the trajectory of o_1 has only *some* components lying on the motion space of o_2 so that the orthogonal component H for o_1 will be larger than that for o_2 . It follows that the above process will separate the two objects.

5.4.3 Iterative AOM Extraction by Voting

At this point, we have P candidate groups and their associated confidence measures from which only the most reliable group is selected based on a voting scheme on the confidence measures. We first refine these candidate groups by combining those groups that contain any common feature points. Beginning with the first candidate group as the only existing

CHAPTER 5. ATTENTION-FROM-MOTION

group, we iteratively add a new candidate group G_{new} . Specifically, if G_{new} contains any common feature point with some existing group G_{exist} , we combine it with G_{exist} , i.e.,

$$G_{exist} = G_{exist} \cup G_{new}$$

and update the corresponding confidence measure as

$$C_{G_{exist}} = C_{G_{exist}} + \frac{|G_{exist} \cap G_{new}|}{|G_{new}|} \cdot C_{G_{new}} \quad (\text{Eq. 5.31})$$

where $\cup$ denotes the union operator, $\cap$ denotes the intersection operator and $|\cdot|$ indicates the size of the set. If G_{new} does not overlap with any existing group, it is formed as a new group with the associated confidence measure. After combination, the group with the maximum confidence measure is chosen as the most reliable motion group. This group is analyzed using the method described in Section 5.3 for a single AOM and a value of *Attention* is assigned to it. The measurement matrix W is updated by removing the columns corresponding to the points in the most reliable group and the entire process is repeated. The complete algorithm is summarized in Algorithm 1.

The reliability of a group depends not only on the number of instances of that group but also on their confidence measures. The confidence measure of every instance of a group may be different because of the difference in the initial motion space. The proposed scoring mechanism takes both these factors into account. Moreover, once the most reliable motion group is obtained, it is removed from further analysis so that the remaining groups can be determined more confidently one by one.

The failure to detect feature points occurs due to two reasons: occlusion (or out-of-view) and errors in tracking. Figure 5 shows an example in which both these phenomena occur. They introduce two types of errors in the measurement matrix W either an entry is missing due to occlusion or the entry is wrong due to tracking error. In the former case, since the proposed algorithm can handle partial data, we simply remove all columns

CHAPTER 5. ATTENTION-FROM-MOTION

Algorithm 1 Recursive Multiple AOM(s) Detection**Require:** measurement matrix W

```

1: while  $W$  is not empty do
2:   Calculate  $A$  from SVD of  $W$  (Equation (Eq. 5.14));
3:   for each feature point  $i$  do
4:     Obtain its 4-tuple from shape space (Section 5.4.1);
5:     Expand the 4-tuple from motion space to obtain the candidate group  $G_i$  (Sec-
       tion 5.4.2);
6:     for each existing candidate group  $G_{exist}$  do
7:       if  $G_i$  overlaps with  $G_{exist}$  then
8:         update  $G_{exist}$  as well as  $C_{G_{exist}}$  according to Equation (Eq. 5.31);
9:       else
10:        add  $G_i$  as a new existing group and associate its confidence measure as  $C_{G_i}$ ;
11:       end if
12:     end for
13:   end for
14:   Calculate Attention for the group with maximum confidence, say  $G_{CMax}$ , which is
       one of the existing groups with the maximum confidence measure  $CMax$  (Equa-
       tion (Eq. 5.29));
15:   Update  $W$  by removing the corresponding columns of those feature points in
        $G_{CMax}$ ;
16: end while
17: Threshold Attention to detect AOM(s)

```

in W corresponding to the occluded points even if they were available in other frames. In the case of missing data, if feature points are missing in a majority of the frames, only the available data is analyzed without degrading performance. On the other hand, if feature points are missing in only few frames, SVD in the above procedure can be replaced by the PowerFactorization method [144] to utilize the additional information that is available. However, the PowerFactorization method requires the rank of the matrix as the prior knowledge, which is not available when there are multiple objects. Hence, for the entire sequence, we remove all the feature points that have been occluded. For example, the points within the green ellipse of Figure 5.5 are removed from W . Since the proposed algorithm can handle partial data, it can still detect an AOM if, despite occlusion, sufficient number of points on it is available. The proposed algorithm can

Figure 5.5: An example of the failure to detect feature points (indicated by small blue circles). There are two types of errors introduced: 1. out-of-view/occlusion error (area indicated by green ellipse); 2. tracking error (area indicated by red ellipse).

detect AOM in a very short interval because of its large discriminating power as well as the ability to handle partial period of motion. Hence, for occlusion changing over time, we can detect AOM without occlusion in some interval, then detect AOM when the object is partially occluded and only miss it when it is fully occluded. However, when the object re-appears and there are sufficient feature points on it, the analysis resumes. In the case of tracking error, the proposed algorithm will be inevitably affected by the failure of feature point detection/tracking (e.g. the tracking error in red ellipse of Figure 5.5). For moderate error in W , the proposed method can still correctly detect AOM due to its large discriminant power. But the performance will degrade when the output of tracking algorithm is too erroneous to correctly reflect motion information. The large discriminant power and the robustness to error in W of the proposed algorithm will be evaluated in the experiment section.

5.5 Experimental Results

The performance of the proposed method for attention-from-motion problem was tested on synthetic and real image sequences. The algorithm was tested to evaluate its ability to handle arbitrary number of objects, different types of motion and the case of partial data.

5.5.1 Synthetic Data

We consider image sequences containing synthetic objects like cube, sphere and pyramid and a static background. We choose 7 feature points from the pyramid, 12 from the cube, 30 from the sphere and 20 from the background. The object motions are synthesized and the camera motion is synthesized to follow one or more objects. The 2D image coordinates of the feature points in each frame are calculated with round-off errors to form the measurement matrix W . The synthetic sequences incorporate a combination of the following characteristics - (i) different number of AOMs, (ii) different object motion and camera motion including degenerate and dependent cases, (iii) AOM with all feature points and, (iv) AOM with partial feature points. The top row in Figure 5.6 (a) shows a sequence containing a cube and a pyramid as well as some background points in which the cube undergoes a 3D motion including both rotation and translation. The camera motion is also synthesized to undergo a different 3D motion but to follow the cube. Hence, from our definition of AOM, the cube has been detected as AOM and marked in red in the bottom row. Since the pyramid undergoes different motion so that the camera does not follow it, it is not an AOM. In Figure 5.6 (b), the two objects undergo different 3D motions - rotation and translation for the cube and translation only for the truncated pyramid. Both objects are followed by a moving camera. We can see the proposed algorithm detected both of them as AOMs and they are marked in red in the bottom row. The sequence in Figure 5.6 (c) consists of a sphere in addition to the cube

CHAPTER 5. ATTENTION-FROM-MOTION

Figure 5.6: (a), (b), (c) Top row: Objects in synthetic image sequences; Bottom row: Detected AOMs in red.

CHAPTER 5. ATTENTION-FROM-MOTION

and pyramid. However, only half of the sphere is observable and only feature points from this half are included in W . In the sequence, the camera undergoes *only* translation, i.e., this is a case of degenerate camera motion. The pyramid and the sphere undergo different rotations and the same translation as the camera, while the cube has a different translation. The bottom row shows that the two AOMs - sphere and pyramid - are correctly identified in red.

We also test the robustness of the proposed algorithm to errors in the measurement matrix W . Gaussian noise with standard deviation varying from 0.5 to 10 was added to W of the synthetic data to introduce the errors in the trajectories of feature points. Such noise perturbs the image coordinates of feature points. Table 5.1 shows the average offset as well as the maximum offset in the image coordinates of the feature points. To measure the performance of AOM detection, we calculate the precision and recall of the detection based on 10 trials at each noise level. Here, the precision is defined as the ratio between the number of correctly detected AOM points and the number of total detected points. Similarly, the recall is defined as the ratio between the number of correctly detected AOM points and the number of all AOM points. Each trial generates a new synthetic sequence where motions are different and the number of AOMs is different. As shown in Figure 5.7, high precision and recall rates are obtained for small noise level indicating that the proposed algorithm is robust to moderate errors in W . When the noise level is very high, the performance of the proposed algorithm degrades because of the incorrect motion information of W . Even so, we note that at a noise level of 10 when the average

Table 5.1: Average and maximum error of image coordinates in W under different noise level.

Std Dev. of Noise	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
Ave Offset (pixel)	0.80	1.12	1.36	1.59	1.78	1.95	2.13	2.24	2.36	2.53
Max Offset (pixel)	3.62	4.89	6.00	6.78	8.29	8.77	9.79	10.07	10.34	11.03

CHAPTER 5. ATTENTION-FROM-MOTION

Figure 5.7: Precision and recall of feature points of AOM(s) *w.r.t* noise level.

offset is about 2 pixels and the maximum error is about 11 pixels, the precision is still very high and the recall is reasonably good at about 77%. We pushed the noise level further to standard deviations of 25 and 30 to obtain a recall of 38% and 12% respectively, while maintaining a precision of close to 1. This shows that the proposed algorithm never identifies an incorrect AOM, but will miss a true AOM when noise in the measurement matrix is very high, resulting in erroneous motion information.

5.5.2 Real image sequences

Next, we apply the proposed AOM detection method on real video sequences shown in Figure 5.8 and Figure 5.9. The top row shows the feature points used to form W , while the second row shows feature points belonging only to the AOMs (different AOMs are indicated by different colors). The feature points are detected as follows: The corner points in the first frame are detected and tracked in the subsequent frames using [145]. Only the feature points which are correctly tracked are used to form W . For some sequence, the AOM(s) do not have enough corner points, we manually label minimum

CHAPTER 5. ATTENTION-FROM-MOTION

points to cover them. Figure 5.8 shows the results of AOM detection on three sequences. In Figure 5.8 (a), the top row contains 70 feature points. The bottom row shows that the moving car as well as the time stamp have been correctly identified as AOM, while the points in the background have been removed. The yellow car undergoes a full 3D motion and is detected as an AOM. The time stamp is also identified as another AOM since it is fixed on the screen. The sequence in Figure 5.8 (b) contains two moving cars of which only one is followed by the camera. This is also a challenging sequence since the cluttered background results in numerous tracking errors resulting in a very noisy measurement matrix W . As the bottom row shows, the tracked moving car has been identified as the AOM. In Figure 5.8 (c), the crane in the foreground and the duck in the background are being followed by the camera and the algorithm correctly identifies them as AOMs. Note that the duck gets partially occluded by the crane so that there is missing data in the measurement matrix. However, the algorithm succeeds in identifying the AOMs, without explicitly interpolating for the missing data. Figure 5.9 shows the results of AOM detection on three more sequences. The sequence in Figure 5.9 (a) is the “mobile-and-calendar” in which there are 3 objects - ball, train and calendar - and a stationary background. The ball and train are followed by the camera and hence they are the two AOMs. Although the ball and train are close to each other and their motions are degenerate and partially dependent, the proposed method succeeds in separating them and identifying both of them as AOMs from a small number of points, e.g. only 6 on ball.

The sequence shown in Figure 5.9 (b) is a cluttered one with many persons. 58 feature points partially cover five persons. The proposed method again succeeds to separate them and identify the person followed by the camera. Although the assumption in this paper was of rigid body motion, we also experimented the proposed method on articulated motion. For the walking sideview sequence in Figure 5.9 (c), the centroids of the arm and the leg and also the centroids of their motion are not stationary. Here the joints

CHAPTER 5. ATTENTION-FROM-MOTION

Figure 5.8: Top row: All feature points; Bottom row: Detected AOMs for (a) Seq 1, (b) Seq 2, (c) Seq 3. In the top row, the number of feature points *on* an AOM need not be larger than that outside the AOM (figure best viewed in color).

CHAPTER 5. ATTENTION-FROM-MOTION

Figure 5.9: Top row: All feature points; Bottom row: Detected AOMs for (a) Seq 4, (b) Seq 5 and (c) Seq 6. In the top row, the number of feature points *on* an AOM need not be larger than that outside the AOM (figure best viewed in color).

CHAPTER 5. ATTENTION-FROM-MOTION

Table 5.2: Comparison with simply checking for stationary centroid. We calculate the proposed attention measure L_q/S_q and S_c of (Eq. 5.32) on the objects for the 3 sequences in Figure 5.9. The ground truth AOM(s) are highlighted in **bold**.

Seq 4 (Figure 5.9 (a))			Seq 5 (Figure 5.9 (b))			Seq 6 (Figure 5.9 (c))		
object	S_c	L_q/S_q	object	S_c	L_q/S_q	object	S_c	L_q/S_q
ball	6.49	383.64	person 1	28.38	0.24	head	5.73	5.26
train	1.59	103.57	person 2	25.35	0.004	arm	13.65	5.40
calendar	10.80	0.26	person 3	22.20	0.04	leg	13.58	2.93
bg	11.71	4.28	person 4	8.06	3.14	bg	18.48	0.04
/	/	/	person 5	9.44	0.37	/	/	/

of the shoulder and the knee are followed by camera and the rotation motions around them satisfy the special structure of the motion matrix M . The algorithm returned three AOMs corresponding to the three “rigid body motions” of the head (red), an arm (green) and a leg (yellow). All three AOMs cover the whole walking person followed by the camera.

To compare the proposed AOM detection method with the simple check for stationary centroid, we compare the proposed measure L_q/S_q and the following measure for the stationary of object centroid.

$$S_c(X, Y) = std(X) + std(Y) \quad (\text{Eq. 5.32})$$

where std denotes the standard deviation. From Table 5.2, we can see the proposed method is much more discriminant than the simple check for stationary of centroid. Although the S_c of AOM(s) are smaller than that of other objects which are not AOM, the fact that there is no optimal threshold to separate AOM(s) from others will degrade the AOM detection. For example, only the train will be detected as an AOM while the ball will be missed for Seq 4. Similarly, both arm and leg part will be missed in Seq 6. For Seq 5, another person (person 5 - the lady in left bottom part) which is not an AOM will be wrongly detected as an AOM. However, it is clear that the proposed method can

Figure 5.10: Detection of AOM with partial occlusion over a very short interval (6 frames). First row shows 3 frames with all feature points and second row shows the detected feature points belonging to AOM.

well separate AOM from others according to the attention values of L_q/S_q for all these 3 sequences.

We also conduct an experiment to evaluate the ability of the proposed method to detect AOM in a very short time interval and to handle partial occlusion that changes over time. In Figure 5.10, we show 3 frames from a short sequence of *only* 6 frames where the body of the monkey (which is the AOM) is partially occluded. The occlusion changes over time. Due to the discriminating power of the algorithm, the AOM can be detected over such very small intervals. From the results, we can see that the proposed algorithm can detect the feature points on the monkey's head as part of the AOM, while ignoring those in the occluded parts and in the background.

Through the experiments on synthetic data and real video sequence, it is empirically shown that the proposed multi-body motion grouping algorithm is efficient for the moderate number of feature points (e.g. 300-500). The current bottleneck of the proposed

framework is about the tracking of feature points, which can be further improved by investigating some real-time tracking algorithms.

5.6 Conclusion

We target at the problem of detecting a specific form of VAR in motion from video sequence in this chapter. We have introduced the concept of attention object in motion, which is the object that the cameraman wishes to focus on. Then we define the problem of attention-from-motion that extracts AOMs from a video sequence. We propose the factorization of measurement matrix to describe a dynamic scene where a moving camera observes a moving object. Based on this factorization, we analyze the special structure in the motion matrix of single AOM and propose a method to estimate the existence of such structure without complete factorization. Finally, we describe an iterative framework to detect multiple AOMs by integrating shape space and motion space as well as voting scheme.

The proposed algorithm provides a robust framework for identifying multiple AOMs. The analysis of structure in the motion matrix of an AOM from factorization enables the detection of AOM(s) when only partial data is available due to occlusion and over-segmentation, a partial observation of complete object motion, or a special following pattern where the centroid of object is not stationary. Without recovering the motion of either object or camera, the proposed method can detect AOM robustly from any combination of camera motion and object motion, even for degenerate motion. For multiple AOMs, partial groups of individual objects can be robustly extracted. Although the assumptions of orthographic projection and rigid motion make the proposed algorithm cannot handle the video including perspective effects e.g. zoom in/out, it can be extended to other camera model e.g. weak-perspective camera to eliminate this limitation.

Chapter 6

Conclusions

This dissertation addresses the problem of detecting visual attention regions in images and videos. For VAR detection in single static image, we have proposed novel answers to these two questions using both feature-based method and region-based method. For VAR detection in video sequence, we have proposed a new definition for moving attention regions and have provided a solution to detect them. In this chapter, we summarize the research contributions achieved in this dissertation and also discuss some future research issues about visual attention region detection in both single and image sequence/collection.

6.1 Summary of Contributions

- (i) We have proposed the idea of using spatial information of a set of salient points indicated by the contrast of a single feature to select those feature(s) that indicate visual attention region. After calculating the contrast map of every feature, a *Composite Saliency Indicator* is proposed to estimate the contribution of the contrast of every feature towards detecting VARs. The convex hull and the spatial density of the salient point set are calculated. Such spatial information of salient points from one feature is used to determine the contribution of this feature to the VAR according to the assumption that salient region is compact. CSI generates a refined

CHAPTER 6. CONCLUSIONS

saliency map where VAR can be detected more accurately. We have also proposed the idea of using spatial relationship between the contrasts of multiple features to improve VAR detection.

The idea of *Context Suppression Model* is motivated from the surrounding suppression mechanism of visual system. At every spatial location, the suppression factor is adaptively estimated from the similarity of different feature contrasts in a local neighborhood. Eigen-decomposition of the Attention Cue Covariance Matrix (ACCM) is used to calculate such similarity among different contrasts. The proposed model has been evaluated in both subjective and objective evaluation with addition of a new texture contrast as another feature candidate. The experiment results show that the proposed model can eliminate the spurious contrasts which cannot indicate the correct VAR and outperform than the state-of-the-art models, especially when the VAR is small and the background is cluttered. An image adaptation prototype system is also implemented by integrating the proposed methods with the MPEG-21 media adaptation framework.

- (ii) The feature-based method for VAR detection has some limitations: the saliency is measured in a local neighborhood and the saliency map does not contain any region information. We go beyond from feature-based method to region-based method to overcome these limitations. By transforming image descriptors of simple features, e.g. color and intensity into a polar space, we represent image regions as the subspaces in the polar space. We extract scale-sentient features to represent both the smooth regions as well as textured regions. Specifically, we first apply the scale-space analysis to detect interesting points as well as their associated scale. The means of color/intensity of the regions around the interesting points, whose scales are equal to the associated scales, are calculated as features for subspace analysis.

CHAPTER 6. CONCLUSIONS

To robustly extract multiple subspaces in the presence of noise and outliers, we propose a new subspace estimation algorithm called *NN-GPCA*. Compared to the original GPCA algorithm, the new method is robust to the presence of outliers due to the weighted least square scheme, where the weights are estimated from the distribution of the K nearest neighbors associated with every data. The weighting scheme constraints the estimated subspace to satisfy not only linearity property but also cluster property of the data corresponding to the image region in polar space. The proposed method also reduces the estimation bias for skew data of GPCA by refining the estimation of subspaces through iteratively changing the weights in high-dimensional embedding space. A new subspace-based attention measure is proposed to measure the attention of the detected regions by considering the feature and geometry properties of the region. This measure calculates the *Cumulative Projection* of the inliers which satisfy both linear subspace constraint and cluster constraint. From the experiment on synthetic data, we evaluate the robustness of NN-GPCA to outliers as well as the robustness to skew subspaces. From the experiments on real image data, we evaluate the invariance of the proposed method to the change of scale/contrast and rotation. In contrast to feature-based methods, the proposed method measure the global attention of every region instead of the local attention of every block. It can estimate the number of regions as well as their attentions without relying on segmentation. It has the ability to correctly detect both smooth VAR as well as textured VAR.

- (iii) For the problem of VAR detection from video sequence, we propose a new definition of *Attention Object in Motion (AOM)* according to an important concept in film production, viz., the tracking shot. AOM is defined as any moving object which is followed by moving camera in the tracking shot. The “following” relationship

CHAPTER 6. CONCLUSIONS

between camera and object is embedded in the definition of AOM and this relationship captures the subjective attention of the cameraman. Compared with other heuristic definitions of motion attention, this definition is generally applicable for both professional video and consumer video. We then propose an *Attention-from-Motion* framework, which consists of a factorization approach to detect AOM(s) from the trajectory matrix of feature points across a video sequence. We derive the factorization of a dynamic scene where a moving camera observes a moving object. By analyzing the factorization, we discover the last column of the motion matrix of the factorization is constant for an AOM. The analysis of single AOM is conducted through estimating the existence of such special structure in the motion matrix without full factorization. The proposed analysis of single AOM has several advantages: (i) it provides a unified solution to any combination of camera motion and object motion; (ii) for degenerate motion, it still can estimate the existence of the special structure from the available information; (iii) it can handle the case when only partial data is observable, where the simple method of checking the stationary centroid cannot work.

In the situation of multiple moving objects, we design a iterative framework to extract multiple AOMs one by one. Within this framework, we combine the *shape space* (the row space of the measurement matrix) and the *motion space* (the column space of the measurement matrix) for motion segmentation. A simple voting scheme is designed to improve the robustness by only extracting the most reliable motion group in each iteration and checking whether it is an AOM. The proposed motion segmentation method overcomes the disadvantages of the methods which only use shape space or motion space. Although sometimes the points belonging to an object can be grouped into multiple groups, they can all be detected to belong to AOM because of the ability of single AOM analysis to handle partial data. We conduct

the experiments on synthetic data and real video sequences to evaluate all these advantages of the proposed factorization approach for *attention-from-motion*.

6.2 Future Research Issues

There exist several potential improvements and open challenges in the area of VAR detection which we describe in this section.

- (i) Feature-based methods assume that feature contrast in a region reflects a potential VAR. This assumption can be used in detecting VARs by using machine learning techniques. If we label every pixel/block in an image as either *attention* or *non-attention*, we can convert the problem of VAR detection to a two-class classification problem. Learning methods like boosting can be applied to adaptively select and combine individual cues for VAR detection. A weak classifier can be designed for each individual feature contrast and the weak classifiers can be combined to obtain a more accurate classifier for VAR detection. The constraints of VAR and the properties of feature contrast data can be used to extend original boosting algorithm.
- (ii) Instead of detecting VAR in a bottom-up manner based on the content of image regions, a top-down process will be useful for object driven VAR detection. Here VAR is guided by the knowledge of the object we are interested in. In this case, VAR is defined as the region that discriminates between sets of images that contain a particular object from those images in which that object is not present. Detecting this form of VAR is helpful for image classification as well as for generic object recognition. Given multiple images containing the object of interest, new learning algorithms can be developed to learn the characteristics of class-specific VAR. The invariance to the intra-class variance, rotation as well as scale and illumination

CHAPTER 6. CONCLUSIONS

change of the object are desirable. New attention measures based on the ability to discriminate between objects must also be developed.

- (iii) The proposed method for detecting AOM can be extended to the whole family of affine camera models including weak-perspective camera as well as para-perspective camera. The robustness of multi-body motion segmentation and AOM analysis in the presence of an imperfect measurement matrix would also be investigated. A better solution to simultaneously detect and segment *only* the AOM(s) is the final objective. Besides the constant property of the special structure in the motion matrix of the AOM, more information in this structure can be studied, e.g., the information relating the camera motion and object motion can be utilized to derive algorithms for retargeting/repurposing video sequence and stabilizing AOMs.
- (iv) A new VAR detection approach could use the probabilistic graphical model. We need to determine how to apply the popular probabilistic graphical models like pLSA and LDA to the problem of VAR detection. Class-specific VAR detector must be learnt from the bounding boxes of the labeled VAR in the training data. For every object class, the most discriminant contrast pattern will be learned and used for VAR detection. New graphical models can be designed to explore spatial coherency of VAR, which can better indicate the object-of-interest. For VAR detection in videos, the graphical model can be extended to discover VAR from motion features. Specifically, the objective is to model, learn and inference VAR in motion as the spatial and temporal distribution of local interest regions in motion.

References

- [1] Laurent Itti, Christof Koch, and Ernst Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November 1998.
- [2] Laurent Itti and Christof Koch, “Feature combination strategies for saliency-based visual attention systems,” *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, January 2001.
- [3] Yu-Fei Ma and Hong-Jiang Zhang, “Contrast-based image attention analysis by using fuzzy growing,” in *Proceedings of the eleventh ACM international conference on Multimedia*, New York, NY, USA, 2003, pp. 228–241, ACM Press.
- [4] Laurent Itti and Christof Koch, “Computational modeling of visual attention,” *Nature Neuroscience Review*, vol. 2, no. 3, pp. 194–203, 2001.
- [5] R Desimone and J Duncan, “Neural mechanisms of selective visual attention,” *Annual Review of Neuroscience*, vol. 18, pp. 193–222, 1995.
- [6] Vincent P. Ferrera and Stephen G. Lisberger, “Attention and target selection for smooth pursuit eye movement,” *The Journal of Neuroscience*, vol. 15, no. 11, pp. 7472–7484, November 1995.
- [7] E. Kowler, E. Anderson, B. Doshier, and E. Blaser, “The role of attention in the programming saccades,” *Vision Research*, vol. 35, no. 13, pp. 1897–1916, 1995.
- [8] J. E. Hoffman, *Attention*, pp. 119–154, Psychology Press. 1998.
- [9] Donald E. Broadbent, *Perception and Communication*, London: Pergamon Press, 1958.

REFERENCES

- [10] J. A. Deutsch and D. Deutsch, "Attention: Some theoretical considerations," *Psychological Review*, vol. 70, pp. 80–90, 1963.
- [11] Yaoru Sun and Robert Fisher, "Object-based visual attention for computer vision," *Artificial Intelligence*, vol. 146, no. 1, pp. 77–123, May 2003.
- [12] C. W. Eriksen and Y. Y. Yeh, "Allocation of attention in the visual field," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 11, no. 5, pp. 583–597, 1985.
- [13] C. W. Eriksen and J. D. St. James, "Visual attention within and around the field of focus attention: A zoom lens model," *Perception and PsychoPhysics*, vol. 40, no. 5, pp. 225–240, 1986.
- [14] A. L. Yarbus, *Eye Movements and Vision*, Plenum Press, New York, NY, 1967.
- [15] D. Noton and L. Stark, "Scanpaths in saccadic eye movements while viewing and recognising patterns," *Vision Research*, vol. 11, pp. 929–942, 1971.
- [16] John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay, "Active vision," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 333–356, January 1988.
- [17] Andrew Blake and Alan Yuille, Eds., *Active Vision*, MIT Press, Cambridge, MA, USA, November 1992.
- [18] Li-Qun Chen, Xing Xie, Xin Fan, Wei-Ying Ma, Hong-Jiang Zhang, and He-Qin Zhou, "A visual attention model for adapting images on small displays," *ACM Multimedia Systems Journal*, vol. 9, no. 4, pp. 353–364, 2003.
- [19] Hao Liu, Xing Xie, Wei-Ying Ma, and Hong-Jiang Zhang, "Automatic browsing of large pictures on mobile devices," in *Proceedings of the eleventh ACM international conference on Multimedia*, New York, NY, USA, 2003, pp. 148–155, ACM Press.
- [20] Yusuo Hu, Xing Xie, Zonghai Chen, and Wei-Ying Ma, "Attention model based progressive image transmission," in *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo*, June 2004, vol. 2, pp. 1079–1082.

REFERENCES

- [21] A. Bamidele, Fred W.M. Stentiford, and J. Morphett, "An attention-based approach to content based image retrieval," *British Telecommunications Advanced Research Technology Journal on Intelligent Spaces (Pervasive Computing)*, vol. 22, no. 3, July 2004.
- [22] Xin-Jing Wang, Wei-Ying Ma, and Xing Li, "Data-driven approach for bridging the cognitive gap in image retrieval," in *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo*, June 2004, vol. 3, pp. 2231–2234.
- [23] Vladimir Brajovic and Takeo Kanade, "Computational sensor for visual tracking with attention," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 8, pp. 1199–1207, August 1998.
- [24] Kentaro Toyama and Gregory D. Hager, "Incremental focus of attention for robust vision-based tracking," *International Journal of Computer Vision*, vol. 35, no. 1, pp. 45–63, November 1999.
- [25] Giuseppe Boccignone, Angelo Chianese, Vincenzo Moscato, and Antonio Picariello, "Foveated shot detection for video segmentation," *IEEE Transaction on Circuit System and Video Technology*, vol. 15, no. 3, pp. 365–377, March 2005.
- [26] William James, *The Principles of Psychology*, New York:Holt, 1890.
- [27] Anne Treisman, "A feature integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.
- [28] Christof Koch and Shimon Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Humman Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.
- [29] Laurent Itti and Christof Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489–1506, 2000.
- [30] Jeremy M. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bulletin and Review*, vol. 1, pp. 202–238, 1994.

REFERENCES

- [31] John K. Tsotsos, Sean M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nufflo, "Modeling visual attention via selective tuning," *Special Volume on Computer Vision, Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, October 1995.
- [32] Bruno A. Olshausen, Charles H. Anderson, and David C. Van Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," *The Journal of Neuroscience*, vol. 13, no. 11, pp. 4700–4719, November 1993.
- [33] Brain J. Scholl, "Objects and attention: The state of the art," *Cognition*, vol. 80, no. 1-2, pp. 1–46, June 2001.
- [34] Robert Desimone and John Duncan, "Neural mechanisms of selective visual attention," *Annual Review of Neuroscience*, vol. 18, pp. 193–222, 1995.
- [35] John Duncan, "Converging levels of analysis in the cognitive neuroscience of visual attention," *Philosophical Transactions of the Royal Society of London. Series B, Biological Science*, vol. 353, pp. 1307–1317, 1998.
- [36] Linda Lanyon and Susan Denham, "A model of object-based attention that guides active visual search to behaviourally relevant locations," in *Attention and Performance in Computational Vision*. 2005, vol. 3368, Springer Berlin / Heidelberg.
- [37] Stephen Grossberg and Rajeev D. S. Raizada, "Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex," *Vision Research*, vol. 40, pp. 1413–1432, 2000.
- [38] Marlene Behrmann, Richard S. Zemel, and Michael C. Mozer, "Object-based attention and occlusion: Evidence from normal participants and a computational model," *Journal of Experimental Psychology, Human Perception and Performance*, vol. 24, no. 4, pp. 1011–36, August 1998.
- [39] Dietmar Heinke and Glyn W. Humphreys, "Saim: A model of visual attention and neglect," in *International Conference on Artificial Neural Networks*, pp. 913–918. 1997.

REFERENCES

- [40] C. Neely, D. Dagenbach, R. Thompson, and T. Carr, "Object-based visual attention: The spread of attention within objects and the movement of attention between objects," in *The 39th Annual Meeting of the Psychonomic Society*, 1998.
- [41] N. Lavie and J. Driver, "On the spatial extent of attention in object-based selection," *Perception & Psychophysics*, vol. 58, pp. 1238–1251, 1996.
- [42] Notger G. Müller and Andreas Kleinschmidt, "Dynamic interaction of object- and space-based attention in retinotopic visual areas," *The Journal of Neuroscience*, vol. 23, no. 30, pp. 9812–9816, October 2003.
- [43] David Soto and Manuel J. Blanco, "Spatial attention and object-based attention: A comparison within a single task," *Vision Research*, vol. 44, pp. 69–81, 2004.
- [44] Laurent Itti and Pierre Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds., pp. 547–554. MIT Press, Cambridge, MA, 2006.
- [45] Laurent Itti and Christof Koch, "A comparison of feature combination strategies for saliency-based visual attention systems," in *Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'99)*, San Jose, CA, 1999, vol. 3644, pp. 473–482.
- [46] Tony Lindeberg, "Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention," *International Journal of Computer Vision*, vol. 11, no. 3, pp. 283–318, December 1993.
- [47] Tony Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 77–116, 1998.
- [48] Timor Kadir and Michael Brady, "Scale, saliency and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.
- [49] Fred Stentiford, "An estimator for visual attention through competitive novelty with application to image compression," in *Picture Coding Symposium*, 2001, pp. 24–27.

REFERENCES

- [50] Fred Stentiford, Nick Morley, and Alison Curnow, "Automatic identification of region of interest with application to the quantification of dna damage in cells," in *Human Vision and Electronic Imaging VII, Proceeding of SPIE*, January 2002, vol. 4662, pp. 244–253.
- [51] Sungyoung Kim, Soyoun Park, and Minhwan Kim, "Central object extraction for object-based image retrieval," in *Proceeding of the Conference on Image and Video Retrieval (CVIR)*, July 2003, pp. 523–528.
- [52] SooYeong Kwak, ByoungChul Ko, and Hyeran Byun, "Automatic salient-object extraction using the contrast map and salient points," in *Proceeding of Fifth IEEE Pacific Rim Conference on Multimedia*, November 2004, pp. 138–145.
- [53] Feng Liu and Michael Gleicher, "Region enhanced scale-invariant saliency detection," in *Proceedings of IEEE International Conference on Multimedia & Expo (ICME)*, July 2006, pp. 1477–1480.
- [54] Ki Tae Park and Young Shik Moon, "Automatic extraction of salient objects using feature maps," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2007, pp. 617–620.
- [55] S. Gilles, *Robust Description and Matching of Images*, Ph.D. thesis, University of Oxford, 1998.
- [56] Martin Jägersand, "Saliency maps and attention selection in scale and spatial coordinates: An information theoretic approach," in *International Conference on Computer Vision*, pp. 195–202. 1995.
- [57] Timor Kadir, Andrew Zisserman, and Michael Brady, "An affine invariant salient region detector," in *Proceedings of the 8th European Conference on Computer Vision*, 2004, pp. 228–241.
- [58] David G. Lowe, "Object recognition from local scale-invariant features," in *Proceeding of IEEE International Conference on Computer Vision*, September 1999, vol. 2, pp. 1150–1157.

REFERENCES

- [59] Nicu Sebe and Michael S. Lew, “Comparing salient point detectors,” *Pattern Recognition Letters*, vol. 24, no. 1-3, pp. 89–96, January 2003.
- [60] Tai Sing Lee and Stella X. Yu, “A information-theoretic framework for understanding saccadic eye movements,” in *Advances in Neural Information Processing Systems 12*, Lawrence K. Saul, Yair Weiss, and Léon Bottou, Eds. 2000, MIT Press.
- [61] Gerald Fritz, Christin Seifert, Lucas Paletta, and Horst Bischof, “Attentive object detection using an information theoretic saliency measure,” in *International Workshop on Attention and Performance in Computational Vision (WAPCV)*, May 2004, pp. 1079–1082.
- [62] Dashan Gao and Nuno Vasconcelos, “Discriminant saliency for visual recognition from cluttered scenes,” in *Advances in Neural Information Processing Systems 17*, Lawrence K. Saul, Yair Weiss, and Léon Bottou, Eds., Cambridge, MA, 2005, pp. 481–488, MIT Press.
- [63] Laura Walker Renninger, James Coughlan, Preeti Verghese, and Jitendra Malik, “A information maximization model of eye movements,” in *Advances in Neural Information Processing Systems 17*, Lawrence K. Saul, Yair Weiss, and Léon Bottou, Eds., Cambridge, MA, 2005, pp. 1121–1128, MIT Press.
- [64] Neil Bruce and John Tsotsos, “Saliency based on information maximization,” in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds., pp. 155–162. MIT Press, Cambridge, MA, 2006.
- [65] Jonathan Harel, Christof Koch, and Pietro Perona, “Graph-based visual saliency,” in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., pp. 545–552. MIT Press, Cambridge, MA, 2007.
- [66] G. Boccignone and M. Ferraro, “Modelling gaze shift as a constrained random walk,” *Physica A*, vol. 331, no. 1-1, pp. 207–218, January 2004.
- [67] Luciano da Fontoura Costa, “Visual saliency and attention as random walks on complex networks,” *eprint arXiv:physics*, March 2006.

REFERENCES

- [68] Shimon Ullman and Amnon Shaashua, "Structural saliency: The detection of globally salient structures using a locally connected network," in *Proceedings of International Conference on Computer Vision*, pp. 321–327.
- [69] C. Grigorescu, N. Petkov, and M.A. Westenberg, "Contour and boundary detection improved by surround suppression of texture edges," *Journal of Image and Vision Computing*, vol. 22, no. 8, pp. 583–679, 2004.
- [70] Song Wang, Jun Wang, and Toshiro Kubota, "From fragments to salient closed boundaries: An in-depth study," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, July 2004, pp. 291–298.
- [71] Song Wang, Toshiro Kubota, Jeffery Mark Siskind, and Jun Wang, "Salient closed boundary extraction with ratio contour," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 546–561, April 2005.
- [72] Feng Ge, Song Wang, and Tiecheng Liu, "Image-segmentation evaluation from the perspective of salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, June 2006, pp. 1146–1153.
- [73] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, and C. Zetzsche, "Scene analysis with saccadic eye movements: Top-down and bottom-up modelling," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 152–160, 2001.
- [74] I. A. Rybak, V. I. Gusakova, A. V. Golovan, L. N. Podladchikova, and N. A. Shevtsova, "A model of attention-guided visual perception and recognition," *Vision Research*, vol. 38, no. 15-16, pp. 2387–2400, August 1998.
- [75] Aude Oliva, Antonio Torralba, Monica S. Castelhana, and John M. Henderson, "Top-down control of visual attention in object detection," in *Proceedings of the IEEE International Conference on Image Processing*, September 2003, vol. I, pp. 253–256.

REFERENCES

- [76] Antonio Torralba, “Modeling global scene factors in attention,” *Journal of Optical Society of America A. Special Issue on Bayesian and Statistical Approaches to Vision*, vol. 20, no. 7, pp. 1407–1418, 2003.
- [77] Aude Oliva, Antonio Torralba, Monica S. Castelhana, and John M. Henderson, “Saliency, objects and scenes: Global scene factors in attention and object detection,” in *Vision Sciences Society Annual Meeting*, 2004.
- [78] Barbara Hidalgo-Sotelo, Aude Oliva, and Antonio Torralba, “Human learning of contextual priors for object search: Where does the time go?,” in *Proceedings of the 3rd Workshop on Attention and Performance in Computer Vision (WAPCV)*, 2005.
- [79] Antonio Torralba, “Contextual influences on saliency,” in *Neurobiology of Attention*, Laurent Itti, Geraint Rees, and John K. Tsotsos, Eds., pp. 586–593. Academic Press/Elsevier, 2005.
- [80] Frank Moosmann, Diane Larlus, and Frederic Jurie, “Learning saliency maps for object categorization,” in *ECCV International Workshop on The Representation and Use of Prior Knowledge in Vision*. 2006, Springer.
- [81] Tie Liu, Jian Sun, Xiaoou Tang Nan-Ning Zheng, and Heung-Yeung Shum, “Learning to detect a salient object,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, June 2007.
- [82] Dashan Gao and Nuno Vasconcelos, “Integrated learning of saliency, complex features, and object detectors from cluttered scenes,” in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2005, vol. 2, pp. 282–287.
- [83] Dirk Walther, Laurent Itti, Maximilian Riesenhuber, Tomaso Poggio, and Christof Koch, “Attentional selection for object recognition - a gentle way,” in *Biologically Motivated Computer Vision*, November 2003.

REFERENCES

- [84] Ueli Rutishauser, Dirk Walther, Christof Koch, and Pietro Perona, "On the usefulness of attention for object recognition," in *2nd International Workshop on Attention and Performance in Computational Vision 2004*, Prague, Czech Republic, May 2004, pp. 96–103.
- [85] Ueli Rutishauser, Dirk Walther, Christof Koch, and Pietro Perona, "Is bottom-up attention useful for object recognition?," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, July 2004, pp. 37–44.
- [86] Dirk Walther, Ueli Rutishauser, Christof Koch, and Pietro Perona, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Computer Vision and Image Understanding*, vol. 100, no. 1-2, pp. 41–63, October 2005.
- [87] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, January 2004.
- [88] Vidhya Navalpakkam and Laurent Itti, "Towards a unified model for attention and recognition," in *Proceeding of Society for Neuroscience Annual Meeting (SFN'03)*, November 2003.
- [89] Vidhya Navalpakkam and Laurent Itti, "Sharing resources: Buy attention, get object recognition," in *Proceeding of International Workshop on Attention and Performance in Computer Vision (WAPCV'03)*, July 2003.
- [90] Ming Yang, Junsong Yuan, and Ying Wu, "Spatial selection for attentional visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, June 2007.
- [91] Vidhya Navalpakkam and Laurent Itti, "An integrated model of top-down and bottom-up attention for optimal object detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, June 2006, pp. 2049–2056.

REFERENCES

- [92] J. J. Bonaiuto and Laurent Itti, "Combining attention and recognition for rapid scene analysis," in *Proceedings of the 3rd Workshop on Attention and Performance in Computer Vision (WAPCV)*, 2005.
- [93] Vidhya Navalpakkam, Michael Arbib, and Laurent Itti, "Attention and scene understanding," in *Neurobiology of Attention*, L. Itti, G. Rees, and J. K. Tsotsos, Eds., pp. 197–203. Elsevier, San Diego, CA, January 2005.
- [94] Laurent Itti and Pierre Baldi, "A principled approach to detecting surprising events in video," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2005, vol. 2, pp. 631–637.
- [95] Hee-Kyung Lee and Yo-Sung Ho, "A region-based image retrieval system using salient point extraction and image segmentation," in *Proceeding of Third IEEE Pacific Rim Conference on Multimedia*, December 2002, pp. 209–216.
- [96] Oge Marques, Liam M Mayron, Gustavo B Borba, and Humberto R Gamba, "An attention-driven model for grouping similar images with image retrieval application," *Eurasip Journal on Advances in Signal Processing, Special Issue on Image Perception*, vol. 2007, 2007.
- [97] Hong Fu, Zheru Chi, and Dagan Feng, "Attention-driven image interpretation with application to image retrieval," *Pattern Recognition*, vol. 39, no. 9, pp. 1604–1621, September 2006.
- [98] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transaction on Multimedia*, vol. 7, no. 5, pp. 907–919, October 2005.
- [99] Sanghoon Lee, Marios S. Pattichis, and Alan Conrad Bovik, "Foveated video compression with optimal rate control," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 977–992, July 2001.
- [100] Andrew P. Bradley and Fred W.M. Stentiford, "Visual attention for region of interest coding in jpeg200," *Journal of Visual Communication and Image Representation*, vol. 14, pp. 232–250, 2003.

REFERENCES

- [101] Doug DeCarlo David Salesin Michael Cohen Anthony Santella, Maneesh Agrawala, "Gaze-based interaction for semi-automatic photo cropping," in *Proceedings of ACM Human Factors in Computing Systems (CHI)*, June 2006, pp. 771–780.
- [102] Jingdong Wang, Jian Sun, Long Quan, Xiaoou Tang, and Heung-Yeung Shum, "Picture collage," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, June 2006.
- [103] Jun-Cheng Chen, Wei-Ta Chu, Jin-Hau Kuo, Chung-Yi Weng, and Ja-Ling Wu, "Tiling slideshow," in *Proceedings of the 14th ACM international conference on Multimedia*. 2006, pp. 25–34, ACM Press.
- [104] He-Qin Zhou Xin Fan, Xing Xie and Wei-Ying Ma, "Looking into video frames on small displays," in *Proceedings of the eleventh ACM international conference on Multimedia*, New York, NY, USA, 2003, ACM Press.
- [105] Feng Liu and Michael Gleicher, "Video retargeting: Automating pan and scan," in *Proceedings of the 14th ACM international conference on Multimedia*. 2006, pp. 241–250, ACM Press.
- [106] Mehmet Celenk, Qiang Zhou, Vermund Vetnes, and Rakesh Godavari, "Saliency field map construction for region-of-interest-based color image querying," *Journal of Electronic Imaging*, vol. 14, no. 3, pp. 033012, July-September 2005.
- [107] A.K.C. Wong and P.K. Sahoo, "A gray-level threshold selection method based on maximum entropy principle," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 4, pp. 866–871, July 1989.
- [108] Kokichi Sugihara, "Robust gift wrapping for the three-dimensional convex hull," *Journal of Computer and System Sciences*, vol. 49, no. 2, pp. 391–407, 1994.
- [109] Nobuyuki Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Mar 1979.
- [110] B. S. Manjunath and Wei-Ying Ma, "Texture feature for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996.

REFERENCES

- [111] Anthony Vetro and Christian Timmerer, "Iso/iec 21000-7 fcd - part 7: Digital item adaptation," July 2003, ISO/IEC JTC 1/SC 29/WG 11/N5845.
- [112] Gabriel Panis, Andreas Hutter, Jörg Heuer, Hermann Hellwagner, Harald Kosch, Christian Timmerer, Sylvain Devillers, and Myriam Amielh, "Bitstream syntax description: a tool for multimedia resource adaptation within mpeg-21," *Signal Processing: Image Communication, EURASIP*, vol. 18, no. 8, pp. 721–747, 2003.
- [113] Debargha Mukherjee, Geraldine Kuo, and Amir Said, "Structured scalable meta-formats (ssm) version 2.0 for content agnostic digital item adaptation - principles and complete syntax," Tech. Rep., Imaging Systems Laboratory, HP Laboratories, 2003.
- [114] Rene Vidal, Yi Ma, and Shankar Sastry, "Generalized principal component analysis (gpca)," in *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA, June 2003, IEEE, vol. 1, pp. 621–628.
- [115] Todd R. Reed and J. M. Hans Dubuf, "A review of recent texture segmentation and feature extraction techniques," *CGVIP: Image Understanding*, vol. 57, no. 3, pp. 359–372, May 1993.
- [116] J. Martinez-Baena, J. A. Garcia, J. Fdez-Valdivia, and Rosa Rodriguez-Sanchez, "Scale selection using three different representations for images," *Pattern Recognition Letters*, vol. 18, pp. 1453–1467, December 1997.
- [117] Eero P Simoncelli and William T Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proceedings of the second IEEE International Conference on Image Processing*, Washington, DC., October 1995, pp. 444–447.
- [118] Jeremy S. De Bonet, "Multiresolution sampling procedure for analysis and synthesis of texture images," in *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, New York, NY, USA, 1997, pp. 361–368, ACM Press/Addison-Wesley Publishing Co.

REFERENCES

- [119] R. Rodriguez-Sanchez, J.A. Garcia, J. Fdez-Valdivia, and Xose R. Fdez-Vidal, "The rgff representational model: A system for the automatically learned partitioning of visual patterns in digital images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 1044–1073, October 1999.
- [120] Peter Kovesi, "Image features from phase congruency," *Videre: A Journal of Computer Vision Research*, vol. 1, no. 3, Summer 1999.
- [121] Qifa Ke and Takeo Kanade, "Robust subspace clustering by combined use of knnd metric and svd algorithm," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, July 2004, vol. 2, pp. 592–599.
- [122] Rene Vidal, *Generalized Principal Component Analysis (GPCA): an Algebraic Geometric Approach to Subspace Clustering and Motion Segmentation*, Ph.D. thesis, School of Electrical Engineering and Computer Sciences, University of California at Berkeley, August 2003.
- [123] Martin A. Fischler and Robert C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, June 1981.
- [124] David A. Forsyth and Jean Ponce, *Computer Vision: A Modern Approach*, Prentice Hall, August 2002.
- [125] Jan Bormains, Jean Gelissen, and Andrew Perkis, "Mpeg-21: the 21st century multimedia framework," *IEEE Signal Processing Magazine*, vol. 20, no. 2, pp. 53–62, March 2003.
- [126] Yiqun Hu, Liang-Tien Chia, and Deepu Rajan, "Region-of-interest based image resolution adaptation for mpeg-21 digital item," in *Proceedings of the 12th annual ACM international conference on Multimedia*, New York, NY, USA, 2004, pp. 340–343, ACM Press.

REFERENCES

- [127] Richard S. Zemel and Terrence J. Sejnowski, "A model for encoding multiple object motions and self-motion in area MST of primate visual cortex," *Journal of Neuroscience*, vol. 18, no. 1, pp. 531–547, 1998.
- [128] Christopher Pack, Stephen Grossberg, and Ennio Mingolla, "A neural model of smooth pursuit control and motion perception by cortical area mst," *Journal of Cognitive Neuroscience*, vol. 13, no. 1, pp. 102–120, 2001.
- [129] John K. Tsotsos, Yueju Liu, Julio C. Martinez-Trujillo, Marc Pomplun, Evgueni Simine, and Kunhao Zhou, "Attention to visual motion," *Computer Vision and Image Understanding*, vol. 100, no. 1-2, pp. 3–40, October 2005.
- [130] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transaction on Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.
- [131] Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang, "AVE - automated home video editing," in *Proceedings of the eleventh ACM international conference on Multimedia*, New York, NY, USA, 2003, pp. 490–497, ACM Press.
- [132] Trent J. Williams and Bruce A. Draper, "An evaluation of motion in artificial selective attention," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPRW'05) - Workshops*, Washington, DC, USA, June 2005, vol. 3, pp. 85–85.
- [133] Ying-Li Tian and Arun Hampapur, "Robust salient motion detection with complex background for real-time video surveillance," in *Proceedings of IEEE Computer Society Workshop on Motion and Video Computing*, Breckenridge, Colorado, USA, January 2005, vol. 2, pp. 30–35.
- [134] L. Wixson, "Detecting salient motion by accumulating directionally-consistent flow," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 774–780, August 2000.

REFERENCES

- [135] Radu P. Horaud, David Knossow, and Markus Michaelis, “Camera cooperation for achieving visual attention,” *Machine Vision and Applications*, vol. 16, no. 6, pp. 331–342, February 2006.
- [136] María T. López, Miguel A. Fernández, Antonio Fernández-Caballero, José Mira, and Ana E. Delgado, “Dynamic visual attention model in image sequence,” *Image and Vision Computing*, vol. 25, no. 5, pp. 597–613, May 2007.
- [137] Terrance E. Boult and Lisa Gottesfeld Brown, “Factorization-based segmentation of motions,” in *Proceedings of the IEEE Workshop on Visual Motion*, Princeton, NJ, October 1991, pp. 179–186.
- [138] C. W. Gear, “Multibody grouping from motion images,” *International Journal of Computer Vision*, vol. 29, no. 2, pp. 133–150, August 1998.
- [139] David Bordwell and Kristin Thompson, *Film Art: An Introduction*, McGraw-Hill, 7th edition, June 2003.
- [140] Carlo Tomasi and Takeo Kanade, “Shape and motion from image streams under orthography: A factorization method,” *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, November 1992.
- [141] Jo’ a’o Paulo Costeira and Takeo Kanade, “A multibody factorization method for independently moving objects,” *International Journal of Computer Vision*, vol. 29, no. 3, pp. 159–179, September 1998.
- [142] Lihi Zelnik-Manor, Moshe Machline, and Michal Irani, “Multi-body factorization with uncertainty: Revisiting motion consistency,” *International Journal of Computer Vision*, vol. 68, no. 1, pp. 27–41, June 2006.
- [143] Naoyuki Ichimura, “A robust and efficient motion segmentation based on orthogonal projection matrix of shape space,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, SC, June 2000, vol. 2, pp. 446–452.

REFERENCES

- [144] Richard Hartley and Frederik Schaffalitzky, “Powerfactorization: 3d reconstruction with missing or uncertain data,” in *Proceedings of Australia-Japan Advanced Workshop on Computer Vision*, Adelaide, Australia, September 2003.
- [145] Bogdan Georgescu and Peter Meer, “Point matching under large image deformations and illumination changes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 674–688, June 2004.

List of Publications

Conference Papers

1. Yiqun Hu, Liang-Tien Chia and Deepu Rajan, JPEG2000 Image Adaptation for MPEG-21 Digital Items, In proceedings of the 5th Pacific-Rim Conference on Multimedia (PCM), November 30-December 3, 2004, Tokyo, Japan.
2. Yiqun Hu, Xing Xie, Wei-Ying Ma, Liang-Tien Chia and Deepu Rajan, Salient Region Detection using Weighted Feature Maps based on the Human Visual Attention Model, In proceedings of the 5th Pacific-Rim Conference on Multimedia (PCM), November 30-December 3, 2004, Tokyo, Japan.
3. Yiqun Hu , Liang-Tien Chia and Deepu Rajan, Region-of-interest based image resolution adaptation for MPEG-21 digital item, In proceedings of the 12th annual ACM international conference on Multimedia, October 10-16, 2004, New York, NY, USA.
4. Yiqun Hu, Deepu Rajan and Liang-Tien Chia, Adaptive Local Context Suppression of Multiple Cues for Salient Visual Attention Detection, In proceedings of the IEEE International Conference on Multimedia & Expo (ICME), July 5-8, 2005, Amsterdam, Netherlands.
5. Yiqun Hu, Deepu Rajan and Liang-Tien Chia, Robust subspace analysis for detecting visual attention regions in images, In proceedings of the 13th annual ACM International Conference on Multimedia, November 6-11, 2005, Singapore.
6. Min Xu, Jiaming Li, Yiqun Hu, Liang-Tien Chia, Bu-Sung Lee, Deepu Rajan and Jian-fei Cai, An Event-Driven Sports Video Adaptation for the MPEG-21 DIA Framework, In proceedings of the IEEE International Conference on Multimedia & Expo (ICME), July 9-12, 2006, Toronto, Ontario, Canada.
7. Min Xu, Jiaming Li, Liang-Tien Chia, Yiqun Hu, Bu-Sung Lee, Deepu Rajan and Jesse S. Jin, Event on Demand with MPEG-21 Video Adaptation System, In proceedings

REFERENCES

of the 14th annual ACM International Conference on Multimedia, October 23-27, 2006, Santa Barbara, CA.

8. Yiqun Hu, Deepu Rajan and Liang-Tien Chia, Scale Adaptive Visual Attention Detection by Subspace Analysis, In proceedings of the 15th annual ACM International Conference on Multimedia, September 24-29, 2007, Augsburg, Germany.

9. Ling Chen, Yiqun Hu and Wolfgang Nejdl, Using Subspace Analysis for Event Detection from Web Click-through Data, In proceedings of the 17th International Conference on World Wide Web, WWW 2008, April 21-25, 2008, Beijing, China.

10. Xiangang Cheng, Yiqun Hu and Liang-Tien Chia, Image Near-Duplicate Retrieval Using Local Dependencies in Spatial-Scale Space, In proceedings of the 16th annual ACM International Conference on Multimedia, October 27-31, Vancouver BC, Canada.

Journal Papers

1. Yiqun Hu, Deepu Rajan and Liang-Tien Chia, Detection of Visual Attention Regions in Images using Robust Subspace Analysis, Journal of Visual Communication and Image Representation, Elsevier (Published).

2. Yiqun Hu, Deepu Rajan and Liang-Tien Chia, Attention-from-Motion: A Factorization Approach for Detecting Attention Objects in Motion, submitted to Computer Vision and Image Understanding, Elsevier (Accepted)