The Hong Kong Polytechnic University

Department of Computing

Multimedia Content Analysis via Computational Human Visual Model

ZHONG Shenghua

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

February 2013

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

____________________________ (Signed)

Shenghua ZHONG (Name of Student)

Abstract

Multimedia content analysis refers to the computerized understanding of the semantic meanings of a multimedia document. Despite more than twenty years of extensive research, multimedia content analysis for real-world applications remains a well-known challenge in the field of multimedia and computer vision. Due to the human-machine gap in multimedia content analysis, more and more researchers focused on constructing the computational human visual models to imitate human perception and intelligence.

This thesis proposes a novel framework to solve various problems in multimedia content analysis via computational human visual models. We want to provide a human-like judgment by referencing the architecture of the human visual system and the procedure of intelligent perception. The techniques based on four important processes in human visual system are designed as follows: 1) retinal image formation for object detection and recognition; 2) attention allocation for image saliency detection and quality assessment; 3) perceptual modeling for image annotation; and 4) visual cortex simulation for multimedia content analysis.

Retinal image formation is the first step of human visual system. Integrated with the limited frame rate of retinal image formation, a novel Invariant moment & Curvelet coefficient (IMCC) feature space with two novel algorithms are proposed for water reflection detection and recognition. The proposed feature space and algorithms demonstrate impressive results in the water reflection image classification, the reflection axis detection, and the retrieval of the images with water reflection.

The fovea produces the most accurate information. Therefore, to encode

detailed visual information, eyes need to be moved so that this area is focused on the visual locations. As a cognitive process of selectively concentrating on one aspect of the environment while ignoring other things, attention has been referred to as the allocation of processing resources. Construction of attention model in multimedia data is useful for applications in multimedia like object segmentation, object recognition and quality assessment. Our novel attention model Bi-directional saliency map (BSMP) integrates bottom-up saliency features and top-down targets information together. Empirical validations on standard datasets demonstrate the effectiveness of the bottom-up saliency detection, and top-down saliency detection. Furthermore, in the image quality assessment task, our technique based on Bi-directional saliency map (BSMP) outperforms the representative blurriness methods.

The visual process is not only based on the immediate visual features, but also relies on the past experience of regularities. As a perceptual processing in which human brain gathers the information from visual elements and their surroundings, contextual cueing provides spatial knowledge about the objects in image. Different from existing techniques only rely on the visual information, the proposed Fuzzy-based contextual-cueing label propagation (FCLP) model addresses the challenging problem in region level annotation to improve the semantic understanding of images. The proposed technique shows obvious performance improvement of label to region assignment for images with multiple objects and complex background.

The visual cortex of the brain is the most important part in the human visual system which is responsible for processing visual information. Deep architecture composed of multiple layers of parameterized nonlinear modules is a representative

paradigm that has achieved notable success in modeling the human visual system. By referencing the architecture of the visual cortex and the procedure of perception, we construct two novel deep networks models for the two classical and intelligent tasks in multimedia content analysis. The first novel deep networks model called Bilinear deep belief networks (BDBN) is proposed for the task of image classification. The second novel deep learning technique called Field effect bilinear deep belief networks (FBDBN) is proposed to seek the recognition discriminant boundary and estimate the missing features jointly. Extensive experiments on various standard datasets not only show the distinguishing ability of our model in various tasks but also clearly demonstrate our intention of providing a human-like image analysis by referencing the human visual system and perception procedure.

Computational human visual models have demonstrated good performance in multimedia content analysis. Further work will be explored from two aspects: how to propose a general deep learning model by simulating human visual system and how to explore deep learning model for video data analysis.

Keywords: Multimedia content analysis, human visual system, computational human visual model, deep learning.

Publications Arising from the Thesis

1. **Zhong**, **S.H.**, Liu, Y., Liu, Y., Chung, F.L.. 2012. Region level annotation by fuzzy based contextual cueing label propagation. In *Multimedia Tools and Applications*. Accepted.

2. **Zhong, S.H**., Liu, Y., Zhang, Y., Chung, F.L.. 2012. Attention modeling for face recognition via deep learning. In Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci 2012), pages 2609-2614.

3. **Zhong**, **S.H.**, Liu, Y., Chung, F.L., Wu, G.S.. 2012. Semiconducting bilinear deep learning for incomplete image recognition. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (ICMR 2012).

4. **Zhong**, **S.H.**, Liu, Y., Liu, Y.. 2011. Bilinear deep learning for image classification. In Proceedings of the 19th ACM International Conference on Multimedia (ACMMM 2011), pages 343-352.

5. **Zhong**, **S.H.**, Liu, Y., Shao, L., Chung, F.L.. 2011. Water reflection recognition via minimizing reflection cost based on motion blur invariant moments. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR 2011).

6. **Zhong**, **S.H.**, Liu, Y., Shao, L., Wu, G.S.. 2011. Unsupervised saliency detection based on 2D Gabor and Curvelets transforms. In Proceedings of the 3rd International Conference on Internet Multimedia Computing and Service (ICIMCS 2011), pages 146-149.

7. **Zhong**, **S.H.**, Liu, Y., Liu, Y., Chung, F.L.. 2010. A semantic no-reference image sharpness metric based on top-down and bottom-up saliency map modeling. In Proceedings of the 17th International Conference on Image Processing (ICIP 2010),

pages 1553-1556.

8. **Zhong**, **S.H.**, Liu, Y., Chung, F.L.. 2010. Fuzzy based contextual cueing for region level annotation. In Proceedings of the 2nd International Conference on Internet Multimedia Computing and Service (ICIMCS 2010), pages 1-6.

9. **Zhong**, **S.H.**, Liu, Y., Liu, Y., Li, C.S.. Water reflection recognition based on motion blur invariant moments in curvelet space. In *IEEE Transactions on Image Processing*. Submitted.

10**. Zhong**, **S.H.**, Liu, Y., Liu, Y.. Robust image classification with human visual cortex-like mechanisms. In *IEEE Transactions on Multimedia*. Submitted.

Acknowledgements

First, I would like to express my greatest thanks to my dear supervisor, Dr. Yan Liu for her best guidance and support to my whole Ph.D. study and my life. Dr. Liu tried her best to train me as a qualified researcher. Her wisdom and open-mindedness gave me great ingenuity and freedom to explore research world of my own interest. I grew up through the difficulties with her endless encourages and help. Without her, I have no chance to do research and learn knowledge in different fields, and I would be not the person I am now. I would like to thank my co-supervisor, Dr. Korris Fu-Lai Chung, for his great support during my Ph.D study life. It is with his valuable advices on my papers and generous help on my research that I achieved a remarkable progress in these years. Furthermore, I would like to thank Dr. Jonathan Flombaum and Dr. Justin Halberda for their inspiring and constructive advices and suggestions on my work during my visit in Johns Hopkins University. My thanks are given to my colleagues who discussed researches and shared ideas with me, especially Yang Liu, Zheng Ma and JinJing Wang. I also wanted to thank my friends, especially Shasha Liu, Weiling Liu, Yingjie Huang, Zuhui Xiao, and Hiu Fai Sze. They taught me a lot during these years and made me a better person.

Finally, I would like to thank my parents, Hanxun Zhong, Ji-An Hu, and my husband Heng Zhang, for their forever support, trust, and love through my life. Without them, I will not begin my research and this thesis will never appear.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

With the huge amount of visual data and rapid development of digital technologies, the understanding of the contents in visual data has taken on increased interest and importance (Liu et al., 2011). But until now, humans outperform the best machine visual systems with respect to almost any measure in multimedia content analysis tasks. How to build a model that emulates human visual system has long been an attractive but elusive goal. This thesis proposes a novel framework to solve various problems in multimedia content analysis via computational human visual models. In this chapter, we provide the motivation of using computational human visual modeling for multimedia content analysis; describe the structure of proposed computational human visual modeling framework, and give the organization of the dissertation.

1.1 Motivation of Computational Human Visual Modeling for Multimedia Content Analysis

Multimedia content analysis refers to the computerized understanding of the semantic meanings of a multimedia document (Wang et al., 2000). As we enter the digital multimedia information era, tools that enable automated analysis are becoming indispensable to be able to efficiently access, classify, digest, and retrieve information.

Unfortunately automatic and efficient analysis multimedia content is not an easy problem. In the past decade, great efforts have been devoted to many fundamental problems such as similarity measurements, indexing schemes, and relevance feedbacks on the multimedia layer. Despite the great amount of research efforts, the success of multimedia content analysis systems is quite limited (Gong & Xu, 2007). A main reason for the problem of poor performance is that the semantic gap between computable low-level features and the high-level semantics (Jiang et al., 2009). And how to bridge the semantic gap could be also considered as unique goal of multimedia content analysis.

Due to the human-machine gap in multimedia content analysis, more and more researchers focused on constructing the computational human visual models to imitate human perception and intelligence. Human visual system (HVS), as the part of the central neural system enabling organisms to process visual information (Marr, 1982), becomes one of research focuses in cognitive science and computer science recently. HVS can be roughly divided into four processes, including image formation on retina, visual cortex processing, attention allocation and perceptual processing. The existing work based on human visual system to multimedia content analysis all utilized the characteristics in these four processes. They demonstrated good performance on different tasks including: image quality enhancement (Lindner et al., 2012), image quality assessment (Lambrecht & Verscheure, 1996.) (Narwaria et al., 2012), object recognition (Oliva & Torralba, 2007) (Serre et al., 2007) (Jarrett et al., 2009), and image classification (Oliva & Torralba, 2001) (Siagian & Itti, 2007).

Inheriting the advantages of existing computational human visual models, the primary focus of this thesis is seeking to understand the computational underpinnings of human visual processing through concerted efforts in both reverse- and forward-engineering based on the given tasks of multimedia content analysis. We want to provide a human-like judgment by referencing the architecture of the human visual system and the procedure of intelligent perception.

1.2 Proposed Framework

In the dissertation, we propose a novel framework to solve various problems in multimedia content analysis via computational human visual models. The techniques are designed from four aspects: 1) retinal image formation for object detection and recognition; 2) attention allocation for image saliency detection and quality assessment; 3) perceptual modeling for image annotation; and 4) visual cortex simulation for multimedia content analysis.

1.2.1 Retinal image formation for object detection and recognition

The first step of human visual system is to format the image on the retinal surface. The meaningful and robust feature space with the characters of retinal image formation can essentially improve the performance of further processing, such as detection and recognition. With the limited frame rate of sensors in image formation, the widespread existence of the motion blur distortion influences the image formation and representation, which is more obvious in water reflection image. Due to this distortion, the existing feature space utilized in symmetry detection and recognition is invalid for water reflection image detection and recognition. To address this problem, we construct a novel Invariant moment & Curvelet coefficient (IMCC) feature space (Zhong et al., 2011a). The proposed algorithms based on the feature space demonstrate impressive results in the tasks of water reflection image classification, the reflection axis detection, and the retrieval of the images with water reflection.

1.2.2 Attention allocation for image saliency detection and quality assessment

As a cognitive process of selectively concentrating on one aspect of the environment while ignoring other things, attention has been referred to as the allocation of processing resources. Construction of attention model in multimedia data is useful for applications in multimedia like object segmentation, object recognition and quality assessment. Based on this consideration, a novel attention model called Bi-directional saliency map (BSMP) (Zhong et al., 2010a) (Zhong et al., 2011b) is proposed for the task of image saliency detection and quality assessment. In bottom-up direction, the proposed Gabor & Curvelets based saliency map relying on 2D Gabor and Curvelet transforms. Compared with the traditional bottom-up model based on DOG and wavelets, our model takes advantage of Garbor transform's spatial localization and Curvelet transform's edge and directional information. In top-down direction, the proposed Target based saliency map relying on the center priority and target location. Compared with existing top-down model emphasizes the importance of center or some empirical high-level features, our model concerns about the semantic meaning by referring the tag information.

Empirical validations on standard dataset demonstrate the effectiveness of the bottom-up saliency detection, and top-down saliency detection. Furthermore, in the image quality assessment task, our technique based on Bi-directional saliency map (BSMP) outperforms the representative blurriness metrics.

1.2.3 Perceptual modeling for image annotation

Contextual cueing is the perceptual processing in which human brain gathers information from the visual elements and their surroundings, which acquires incidental learning from past experiences of regularities and rules. The spatial invariants in contextual cueing are thought to be important to perceptual processing and finally guide the object recognition and region annotation. Relying on the advantages of contextual cueing in object recognition and region annotation, a novel label to region assignment (LRA) method called Fuzzy-based contextual-cueing label propagation (FCLP) is proposed (Zhong et al., 2010b) (Zhong et al., 2012a). Fuzzy representation and fuzzy reasoning are utilized to describe the spatial invariants in contextual cueing and imitate the contextual cueing process. The experiments are tested on two public datasets, the results demonstrate that the proposed technique has obvious performance improvement of label to region assignment for the images with multiple objects and complex background.

1.2.4 Visual cortex simulation for multimedia content analysis

The visual cortex of the brain is the most important part in the human visual system which is responsible for processing visual information. Deep architecture composed of multiple layers of parameterized nonlinear modules is a representative paradigm that has achieved notable success in modeling the human visual system. By referencing the architecture of the visual cortex and the procedure of perception, the first novel deep networks model called Bilinear deep belief networks (BDBN) is proposed for the task of image classification (Zhong et al., 2011c) (Zhong et al., 2012b). Unlike most existing deep models, BDBN utilizes a bilinear discriminant strategy to simulate the “initial guess” in human object recognition, and at the same time to avoid falling into a bad local optimum. To preserve the natural tensor structure of the image data, a novel deep architecture with greedy layer-wise reconstruction and global fine-tuning is proposed. To adapt real-world image classification tasks, we develop BDBN under a semi-supervised learning framework, which makes the deep model work well when labeled images are insufficient.

We propose the second novel deep learning technique called Field effect bilinear deep belief networks (FBDBN) to seek the recognition discriminant boundary and estimate the missing features jointly (Zhong et al., 2012c). To address the difficulties of incomplete data, we design a novel second-order deep architecture with the Field effect restricted boltzmann machines, which models the reliability of the delivered information according to the availability of the features. Based on the new architecture, two peaks activation with the bi-directional inference of human’s perception is implemented by three learning stages of Field effect bilinear discriminant initialization, Field effect layer-wise abstraction and estimation, and global fine-tuning with missing features adjustment.

Extensive experiments on different standard datasets not only show the distinguishing ability of proposed deep learning models on various tasks but also clearly demonstrate our intention of providing a human-like image analysis by referencing the human visual system and perception procedure.

1.3 Organization

The organization of this proposed dissertation is illustrated in Figure 1.1. The Chapter 2 proposes a novel feature space Invariant moment & Curvelet coefficient (IMCC) and corresponding novel algorithms based on the characteristics of the retinal image formation for the task of water reflection recognition. A novel attention model called Bi-directional saliency map (BSMP) is introduced in Chapter 3 with the application of image saliency detection and quality assessment. In Chapter 4, we introduce FCLP using characteristics of the perceptual processing for the task of label to region assignment. We propose two visual cortex simulation models BDBN and FBDBN for three different tasks including document summarization, image classification and incomplete data recognition from Chapter 5 to Chapter 6, respectively. In every chapter, we will provide the details of background information and related works. Extensive experiments and the evaluations results are reported in each chapter. Conclusion and future work are provided in Chapter 7.

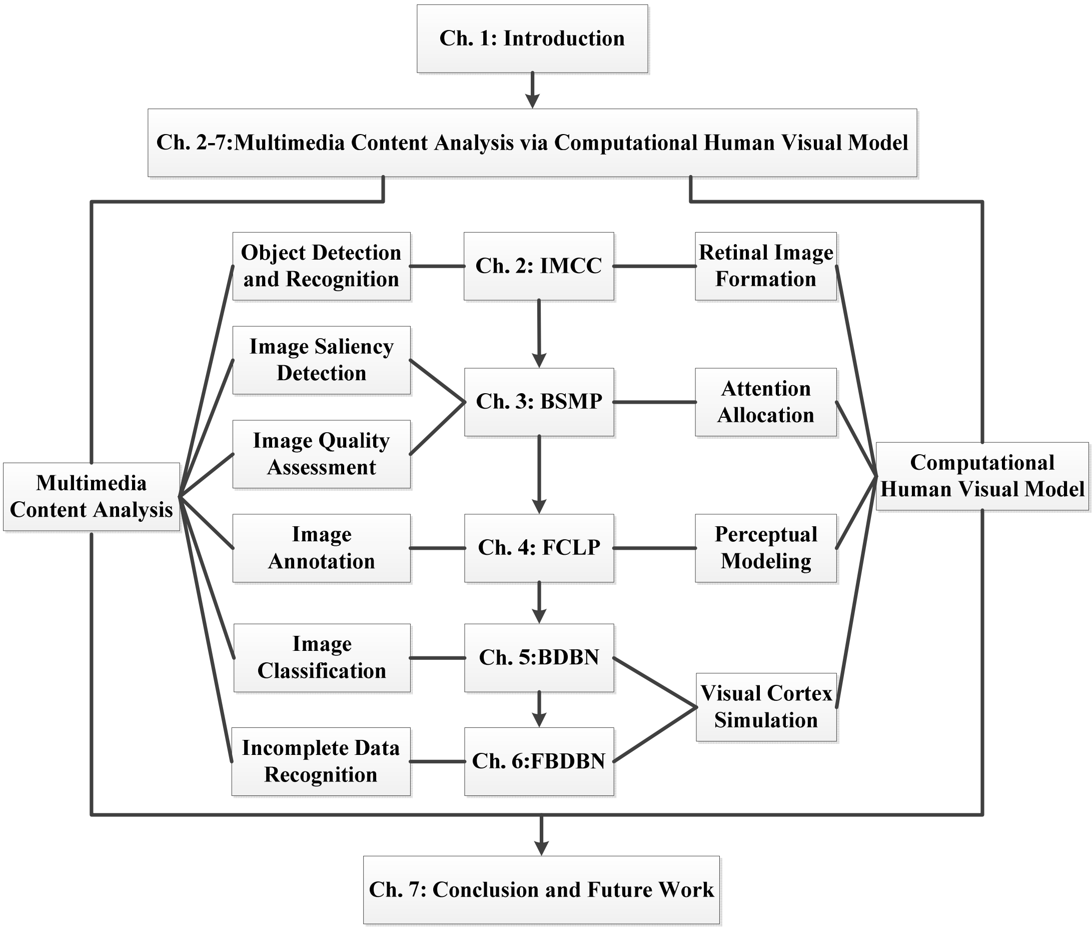


Figure 1.1. Organization of the dissertation.

Chapter 2 Retinal Image Formation for Object Detection and Recognition

2.1 Overview

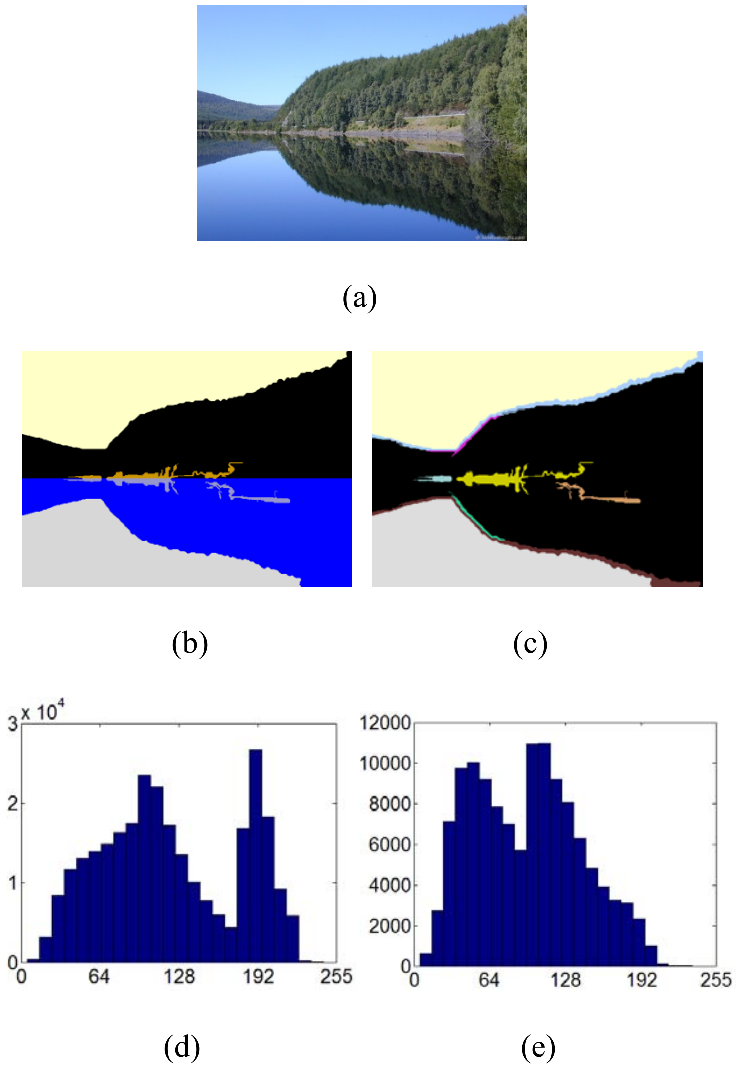
This chapter proposes a novel feature space Invariant moment & Curvelet coefficient (IMCC) and novel algorithms Reflection cost minimization (RCM) and High-frequency Curvelet coefficients discrimination (HCCD) using the characteristics of the retinal image formation (Zhong et al., 2011a) for the task of water reflection recognition.

Water reflection recognition, a typical imperfect reflection symmetry problem, plays an important role in multimedia content analysis. However, existing techniques of symmetry recognition cannot recognize water reflection images correctly because of the complex and various distortions caused by the water wave. Hence, this chapter proposes a novel water reflection recognition technique to solve the problem. First, a novel feature space Invariant moment & Curvelet coefficient (IMCC) is constructed with motion blur invariant moments in low-frequency Curvelet space and of Curvelet coefficients in high-frequency Curvelet space. Second, an efficient algorithm including two sub-algorithms is proposed: Reflection cost minimization (RCM) and High-frequency Curvelet coefficients discrimination (HCCD), to classify water reflection images and to determine reflection axis in these images.

By experimenting on datasets of authentic images in a series of tasks, the proposed techniques prove effective and reliable in classifying water reflection images and detecting the reflection axis, as well as in retrieving images with water reflection.

2.2 Introduction

Reflection happens between two different medias. The direction of a wavefront at the interface changes so that the wavefront returns into the medium from which it is originated. In natural image analysis, water reflection plays an important role. First, water reflection itself is an exciting natural landscape that attracts artists and photographers, so images with water reflection should be considered as one important category of natural images. Experiments from psychology reveal that subjects give favorable ratings to scenes with reflective water (Nasar & Li, 2004). Second, the awareness of the existence of water reflection will greatly influence further analysis of an image, such as image segmentation and object recognition. Figure 2.1(a) is an image with water reflection and the correct segmentation result is shown in Figure 2.1 (b). However, most existing segmentation algorithms, such as the state-of-the-art algorithms: graph-based technique, will consider the mountain and its reflection as one segment as shown in Figure 2.1 (c), if the existence of water reflection is not known prior to the analysis. As a result, the object mountain is not properly recognized due to the wrong segmentation. Obviously, the shape information in Figure 2.1 (c) will be helpless in detecting the mountain. Figure 2.1 (d) is the color histogram of the mountain part in Figure 2.1(a), which is quite different from that of the combination of the mountain and the reflection as shown in Figure 2.1(e). It is obviously that considering the object and the reflection as a whole will distort the color feature for recognition.



(a)

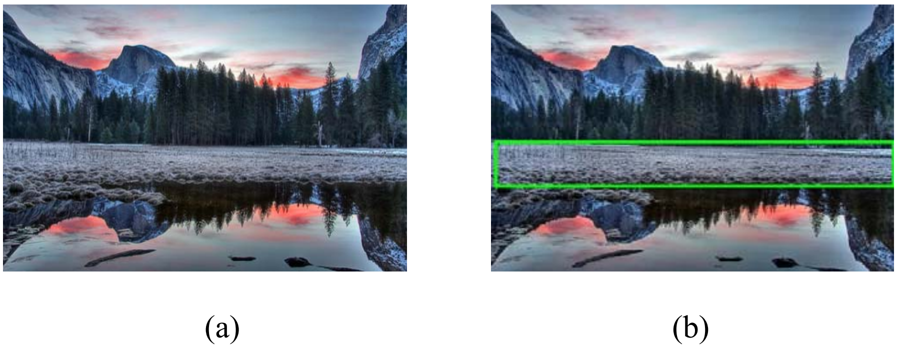
(b) (c)

(d) (e)

Figure 2.1. Example of the influence from water reflection to image segmentation and object recognition. (a) An example image with water reflection. (b) Desired segmentation result of the image. (c) Actual segmentation using existing algorithms. (d) Color histogram of the mountain. (e) Color histogram of the mountain and the reflection.

Although water reflection has great influence in many image processing tasks, currently, few research studies the water reflection images in view of vision (Huang, et al., 2011). To our knowledge, no effort has been made to address the classification, recognition and detection of water reflection images. Previously, only one study has been carried out on detecting the water reflection axis in water reflection image (Zhang, et al., 2010). The flip invariant shape detector utilized in (Zhang et al., 2010) relies on the complete and distinct shape of water reflection part, which cannot be easily satisfied as the water reflection is a complex phenomenon. For example, in Figure 2.2 (a) and (b), the snow mountain and trees are partially reflected because the ice above the water covers some area of the lake. Therefore, the method proposed in (Zhang et al., 2010) cannot be successfully applied to many images containing nature water reflection.

This chapter formulates the water reflection recognition as a special case of imperfect reflection symmetry problem. Reflection symmetry, namely mirror symmetry, is symmetry with respect to reflection. It refers to an object or figure that is indistinguishable from its transformed image. In real-world nature images, most reflection images are not perfect, such as tree leaf and human’s face. In water reflection images, the complexity of water part makes it impossible to keep the consistency between the object part and reflection part perfectly. To address the special characteristics of water wave, we construct a novel feature space that is composed of motion blur invariant moments in low-frequency Curvelet space and of Curvelet coefficients in high-frequency Curvelet space. With the help of moment invariants in low-frequency band, we could distinguish the imperfect symmetry images from other images. Utilizing Curvelet coefficients, water reflection images could be distinguished from other imperfect symmetry images. Based on the novel feature space, we propose an efficient algorithm including two sub-algorithms: Reflection cost minimization (RCM) and High-frequency Curvelet coefficients discrimination (HCCD). This algorithm is effective and reliable to classify water reflection images from other images, and to determine the reflection axis. Moreover, this algorithm has lower computational complexity than exhaust algorithm.



(a) (b)

Figure 2.2. Examples of images containing water reflections of incomplete and indistinct shapes.

The rest of this chapter is organized as follows. Section 2.3 reviews previous work about symmetry, focusing on imperfect reflection symmetry. Section 2.4 discusses the limitations of existing feature space used in symmetry detection and recognition tasks. Then, we propose a new feature space Invariant moment & Curvelet coefficient (IMCC) based on the characteristics of water waves. Section 2.5 provides an efficient solution to solve the problem of classification and recognition of water reflection images based on IMCC. Section 2.6 reports the experiments on authentic datasets. Finally, Section 2.7 draws on conclusion.

2.3 Related Work

Because most man-made and natural objects exhibit some extent of symmetry (Chertok & Keller, 2010), symmetry is an essential and ubiquitous concept in nature, science, and art. Issues relating to symmetry detection and recognition have attracted extensive attention in numerous fields including visual perception, computer vision, robotics, and computational geometry. In philosophy, symmetry, as one of the Gestalt grouping factors, is considered as a pre-attentive feature. For this reason, many artists and photographers prefer works of symmetry, e.g., “The Houses of Parliament, Sunset” shown in Figure 2.3 was painted by Claude Monet when he experimented with reflections in water. This pre-attentive feature is useful to enhance recognition and reconstruction of shapes and objects (Attneave, 1995). Based on the symmetry of an object, human beings can infer its structure and estimate its pose (orientation and positions) in the 3D space, even when certain parts of the object are lost or occluded. Thus, a symmetric object can be characterized efficiently by its symmetry groups, yielding a low dimensional set of features for object representation, recognition, matching, segmentation, and tracking (Lee et al., 2008). Therefore, symmetry is considered as a very important feature and should be processed in high priority.

The research on symmetry roughly includes recognition of the symmetry group (Lee et al., 2008), determination of the axis of symmetry or affinely (Kanade & Kender, 1983) (Shen et al., 2000) and perspectively distorted symmetry detections (Cornelius & Loy, 2006) (Lei & Wong, 1999). Reflection symmetry is one of the common basic symmetries (Weyl, 1952), in which one half of the object is indistinguishable from its mirror transformed image of the other. Reflection symmetry has been studied in many different fields for various applications from face analysis (Mitra & Liu, 2004), vehicle detection (Kuehnle, 1991) to medical image analysis (Mancas et al., 2005). Since the restriction to exact symmetries limits the use of these methods for real-world objects, more efforts have been focused on the imperfect symmetry (Guo et al., 2010). There are two types of imperfect symmetry: local symmetry, in which a portion of a model is perfectly symmetric while the rest is not; and approximate symmetry, in which the entire model is not symmetric but could be made symmetric with a slight deformation (Podolak et al., 2007). Images in Figure 2.4 represent different types of symmetries. Figure 2.4 (a) and (b) are perfect symmetry, and Figure 2.4 (c) to (f) the imperfect symmetry in which (c) and (d) are skewed symmetry by affine or perspective skewing; while (e) and (f) are curved glide-reflection symmetry.

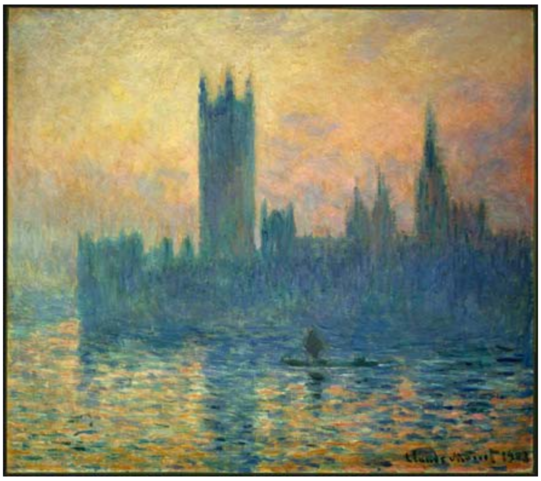
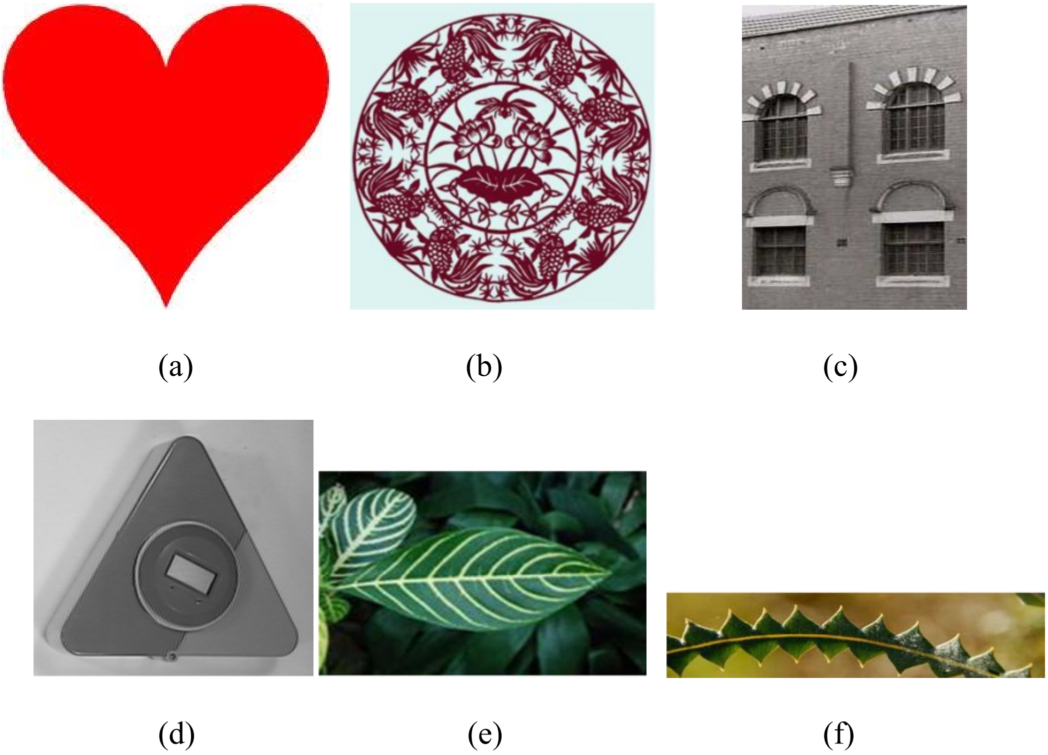


Figure 2.3. The Houses of Parliament, Sunset. Claude Monet. 1903. National Gallery of Art, Washington, D.C.



(a) (b) (c)

(d) (e) (f)

Figure 2.4. Examples of symmetry images. (a) and (b) are perfect symmetry, (c) and (d) are skewed symmetry by affine or perspective skewing. (e) and (f) are curved glide-reflection symmetry.

Based on the nature of the features extracted from images, the existing algorithms for reflection symmetry detection and recognition can be roughly classified into two general approaches (Gross & Boult, 1994), namely, the global approaches and the local approaches.

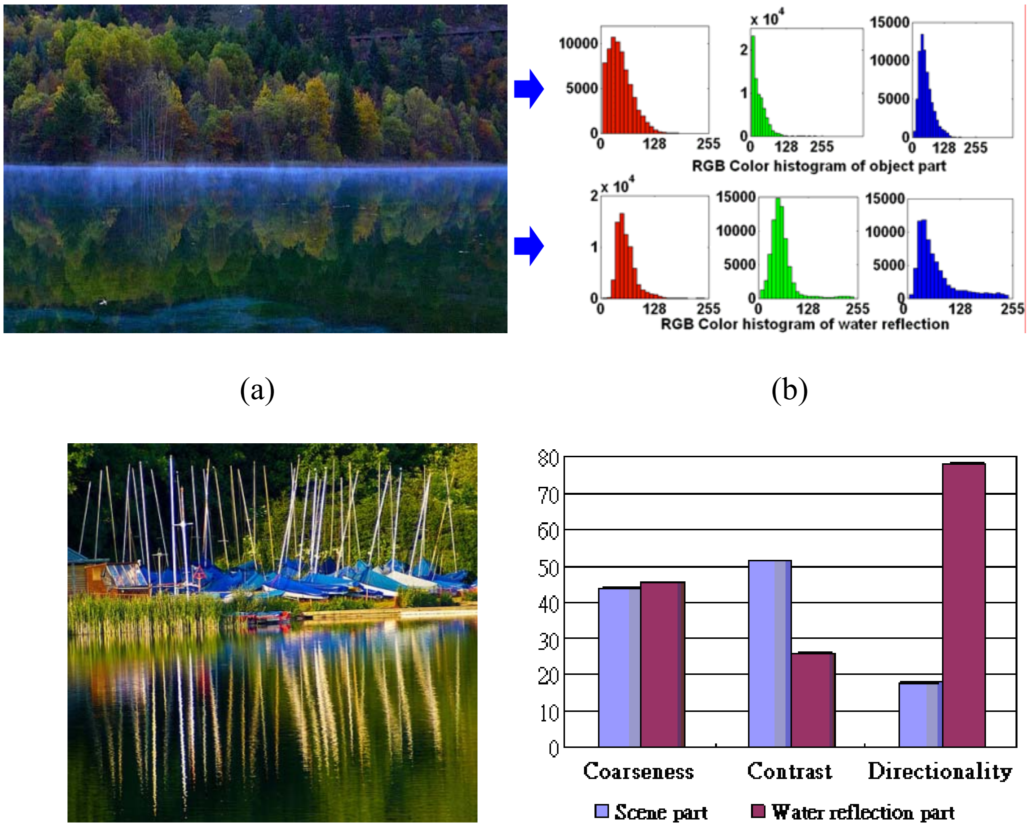
In global approaches, some algorithms are based on the global features, especially in Fourier domain. For example, Lucchese proposed an elegant approach to analyze the angular properties of an image in Fourier domain (Lucchese, 2004). Derrode and Ghorbel analyzed the symmetries of real objects by computing the Analytic fourier-mellin transform (AFMT) (Derrode & Ghorbel, 2004). From a perspective other than the Fourier domain, Friedberg and Gross et al. considered the entire contour at once when locating the axes of skewed symmetries (Friedberg, 1986), (Gross & Boult, 1994).

Because the use of local features is among the corner stones of modern computer vision, recent work starts emphasizing the use of local image features. The representative one is Scale-invariant feature transform (SIFT) descriptor. Loy et al. chose SIFT detection points as interesting salient points and took advantage of pairwise matching of their SIFT descriptors to detect the axis of symmetry (Loy & Eklundh, 2006). Some other existing work focused on the shape characteristics of symmetry (Prasad & Yegnanarayana, 2004). For example, local invariants were computed as single points on the curves (Vangool et al., 1995a & 1995b) or statistically compared pairs of contour points (Yuen & Chan, 1994) (Cham & Cipolla, 1995).

The advantage and disadvantage associated with local and global approaches are well reported in the image processing and computer vision literature. Local approaches can work even when certain parts of the curves are occluded or missing, while the global approaches in general severely suffer from occlusion. On the other hand, the global approaches are insensitive to the noise, while the local approaches are unstable and sensitive to noise unless a multi-scale approach is adopted (Mokhtarian, 1996).

2.4 Feature Space in Water Reflection Recognition

The key to address the difficulty in water reflection recognition is to find out the effective and robust feature descriptors. First, let us observe the distortion due to the water reflection in the most commonly used feature space. Figure 2.5 (a) and Figure 2.5 (b) demonstrate the color distortion of the forest after reflection. Obviously, much of the red information is lost. In Figure 2.5(d), three most important Tamura texture features from the scene part and water reflection part of Figure 2.5(c) are compared, which are the most popular features selected by psychological experiments (Tamura et al., 1978). There exist great differences of contrast and directionality between the original one and its reflection.



(a)

(b)

(c)

(d)

Figure 2.5. Feature distortion in water reflection images. (a) and (c) show the examples of water reflection images. (b) is the color histogram of object part and water reflection part in (a). (d) shows the texture features in the scene part and water reflection part of (a).

2.4.1 Limitation of existing feature space

Features from Fourier domain are the most commonly used for global approach of reflection symmetry detection and recognition. Their methods are based on the idea that Fourier transform preserves the symmetry of images in the Fourier domain. Let $I(\mathrm{x}), \mathrm{x} = [x\ y]^{\mathrm{T}} \in \mathbb{R}^2$ denote the scalar image of 2-D pattern. In (Lucchese, 2004), Lucchese proved that if an image having reflection symmetry with respect to the reflection axis $y = x \times \tan\alpha$, its Fourier transform $I(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^2$ has the same reflection symmetry with respect to the line $\omega_y = \omega_x \times \tan\alpha$. The difference between the original one and the reflection one will be much smaller than the difference between other parts. But due to the characteristics of the water part, this conclusion is not always true. Figure 2.6 (a) shows an example image with water reflection. Figure 2.6 (b) is the image with the correct reflection axis. We calculate the Fourier transform with this reflection axis. Based on Figure 2.6 (d) which is the Fourier transform $I(\boldsymbol{\omega})$ results of object part and water reflection part, we find the $I(\boldsymbol{\omega})$ does not have the reflection symmetry as expected. The average difference of object part and water reflection part is much larger than fake symmetry axis marked in Figure 2.6 (c).

For local approaches of reflection symmetry detection and recognition, SIFT descriptor is the most representative feature. As shown in Figure 2.7 (a), the desired result is that the SIFT saliency points are matched in pairs between the object and its reflection. Figure 2.7 (b) shows the real SIFT points detection and matching result using algorithm in (Loy & Eklundh, 2006). Obviously, it is difficult to recognize the water reflection by matching the SIFT points.

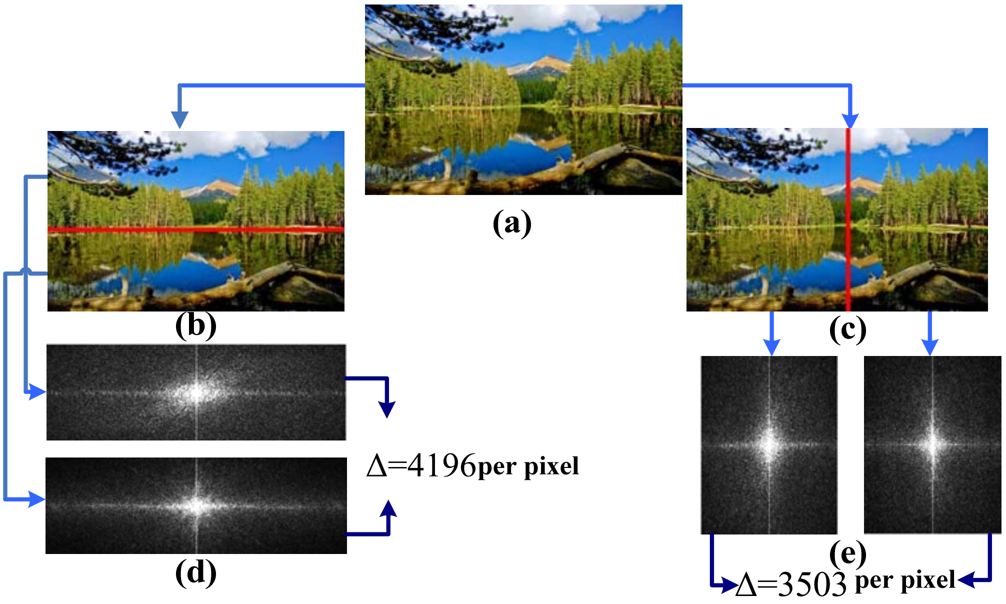
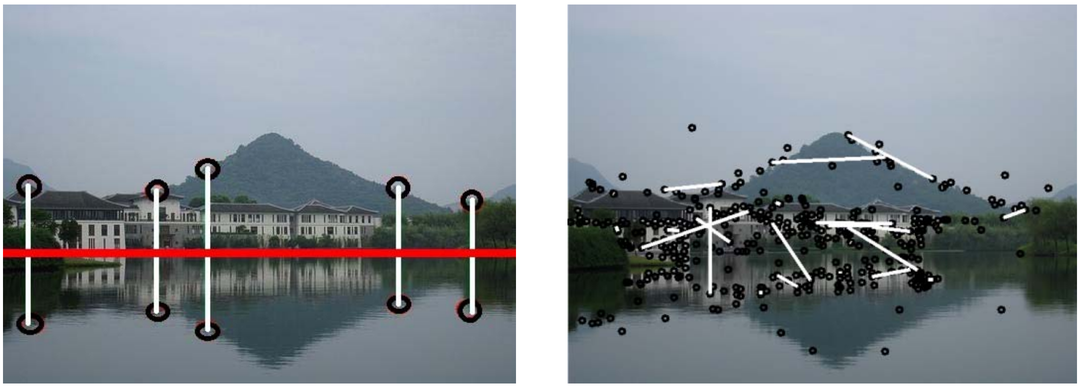


Figure 2.6. An example of features from Fourier domain. (a) Original image. (b) and (d) show the Fourier transform with real reflection axis. (c) and (e) show the Fourier transform with fake reflection axis.



(a) (b)

Figure 2.7. An example of SIFT saliency points detection and matching. (a) is the desired result. (b) is the real result of SIFT descriptor detection and matching.

2.4.2 Feature distortion in retina image formation

As we described in earlier sections, existing feature space utilized in symmetry detection and recognition is invalid to the task of water reflection recognition. We believe the main cause is the motion blur.

As a well-known degradation factor, motion blur is due to the relative motion of the sensor and the scene in finite exposure time (Flusser et al., 1996). To the case of water reflection, motion blur is a result of the interplay between the speed of oscillation of the surface waves and the camera's limited frame rate (Donate et al., 2006). It changes the image features needed for feature-based recognition techniques (Stern et al., 2000). The formation model for the motion blur is:

$$g(x,y) = I(x,y) * h(x,y) + n(x,y) \tag{2.1}$$

where $I(x,y)$ is the original image, $h(x,y)$ is the point spread function (psf), $n(x,y)$ is additive noise, and $g(x,y)$ represents the observed image. Assume that the translation motion function $M(t) = [M_x(t), M_y(t)]$ is known, $h(x,y)$ has the following form (2.2):

$$h(x,y) = \frac{1}{t_e}\int_{t_o}^{t_o+t_e} \delta[x - M_x(t), y - M_y(t)]dt \tag{2.2}$$

where the Dirac delta function describes the two-dimensional displacement function of the image during the exposure interval $(t_0, t_0 + t_e)$, where t_e denotes the exposure period, and $1/t_e$ is a normalizing factor.

Our eyes respond similar to a shutter speed of 1/30 second and the conventional cameras expose pictures 25 or 30 times per second. Although the fastest shutter speed available is much higher now, 1/30 second is still commonly selected in landscape photography. The average phase velocity of water is about 0.3m/sec. So in every expose, one particle of the water will shift about 10mm in average. Based on the principle of image formation, we could calculate that the motion in image is about 5 pixels in usual conditions. It is large enough to change the image features needed for feature-based recognition. Furthermore, motion blur causes a decay of the information and energy in high-frequency band. The change of high-frequency information in water reflection is one reason for the invalidity of existing global algorithms in Fourier domain (Lucchese, 2004).

To analyze the influence of the motion blur in water reflection, we need to have a fundamental understanding of water wave. Water wave could be considered as being composed of a great quantity of periodic progressive waves. Simply speaking, a periodic progressive wave is characterized by the amplitude A, wavelength λ, phase velocity V_p and period $T(T=\lambda/V_p)$, as illustrated in Figure 2.8. Actually, real water wave is exceedingly complex as it is also influenced by the depth of water, the velocity of wind, and so on.

The complex water wave problem in Figure 2.8 could be effectively simplified into a boundary value problem by Newman (Newman, 1977). According to the differential equation with the conditions at the boundaries (bottom boundary conditions and free surface boundary conditions), the small amplitude wave functions could be denoted as Eq. (2.3) in two dimensional x-z plane. In Eq. (2.3), ϑ is the velocity potential, g_r is the gravitational acceleration, and η is the wave profile which means the position of the water surface.

$$\begin{cases} \dfrac{\partial^2\vartheta}{\partial x^2}+\dfrac{\partial^2\vartheta}{\partial z^2}=0 & -H<z<\eta,-\infty<x<+\infty \\ \dfrac{\partial\vartheta}{\partial z}=0 & z=-H \\ \eta=-\dfrac{1}{g_r}\dfrac{\partial\vartheta}{\partial t}\Big|_{z=0} & \end{cases} \tag{2.3}$$

After solving the wave functions utilizing the method of variables separation, we could get the function of wave profile in Eq. (2.4) and phase velocity in Eq. (2.5). And the velocity of every point in the wave profile could be denoted as Eq. (2.6).

$$\eta=A\cos(2\pi x/\lambda-2\pi t/T) \tag{2.4}$$

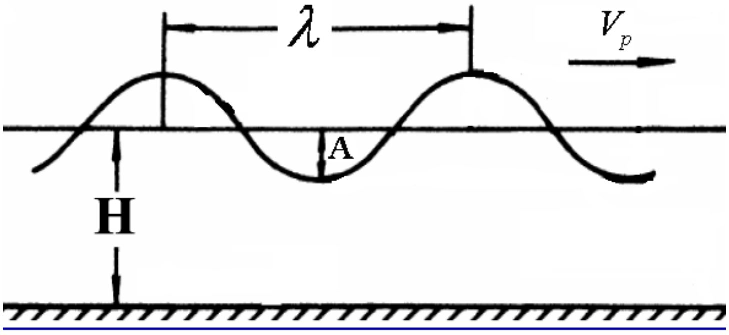


Figure 2.8. Sketch of a periodic progressive water wave in a fluid of mean depth H. Note that λ is the wavelength, and A is the amplitude of wave, and the wave translates with phase velocity V_p.

$$V_p = \sqrt{\frac{g_r L}{2\pi} \tanh \frac{2\pi \mathrm{H}}{\lambda}} \tag{2.5}$$

$$\begin{cases} v_x = 2\pi A \dfrac{\cosh 2\pi(z+\mathrm{H})/\lambda}{T \sinh 2\pi \mathrm{H}/\lambda} \cos(2\pi x/\lambda - 2\pi t/T) \\ v_z = 2\pi A \dfrac{\sinh 2\pi(z+\mathrm{H})/\lambda}{T \sinh 2\pi \mathrm{H}/\lambda} \sin(2\pi x/\lambda - 2\pi t/T) \end{cases} \tag{2.6}$$

Based on Eq. (2.4), we could conclude that the surface of water part has different offsets in position due to the water wave. The offset in position leads to the ineffectiveness of exiting symmetry algorithms based on the local features.

As we known, motion blur could be removed from images with the help of deconvolution, which is often adopted in the literature for motion blur detection and recognition. But the core idea in deconvolution is to calculate the point spread function, assuming that the velocity and direction of motion blur are unique (Qi et al., 2005) (Ji & Liu, 2008). Eq. (2.5) and Eq. (2.6) indicate obviously that the motion in the water is ubiquitous, and that the velocity of different position with different frequency is various too. Therefore, the necessary assumption of deconvolution methods is invalid. Thus, none of existing techniques is effective to this situation, which necessitates effective feature space to address the issues resulted from motion blur.

2.4.3 Invariant moment & Curvelet coefficient feature Space

As we described before, the key to water reflection recognition is effective feature space that is utilized to address the problem resulted from motion blur. Based on the characteristics of water reflection, the task of water reflection recognition could be separated into two parts, the first of which is to distinguish imperfect symmetry images and the second is to distinguish water reflection images from other imperfect symmetry images. Therefore, the proposed feature space has two components focus on these two requirements respectively.

The first component of proposed feature space is the motion blur invariant moments in low-frequency Curvelet space. This feature channel is utilized to distinguish imperfect symmetry images with other images.

Moment invariants were first introduced to the pattern recognition and image processing community in 1962 (Hu, 1962), when Hu employed the results of the theory of algebraic invariants and derived his seven famous invariants to the rotation of 2D objects. Since then, moment invariants have become one of the most important and most frequently used descriptors. There have been numerous papers on moment invariants to affine and projective transforms, to photometric changes and to linear filtering of an image.

Image moments are weighted averages (moments) of the image pixels' intensities, or functions of those moments, usually chosen to have some attractive property or interpretation (Abas & Ono, 2010). Compared with color histogram, the shift of moment due to the change of illumination will be minimal (Mandal et al. 1996), which also often happens in water part.

General moment M_{pq} of an image $I(x,y)$ is defined as:

$$M_{pq} = \iint_{D} p_{pq}(x,y) I(x,y) dxdy \tag{2.7}$$

where p,q are non-negative integers, $r=p+q$ is called the order of the moment, and $p_{pq}(x,y)$ is the polynomial basis function. The most common choice is a standard power basis $p_{pq}(x,y) = x^p y^q$ that leads to geometric moments:

$$m_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q I(x,y) dxdy \tag{2.8}$$

The central moments are defined as:

$$\mu_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\bar{x})^p (y-\bar{y})^q I(x,y) dxdy \tag{2.9}$$

where $\bar{x} = m_{10} / m_{00}$ and $\bar{y} = m_{01} / m_{00}$ are the components of the centroid. If $I(x,y)$ is a digital image, Eq. (2.8) and Eq. (2.9) are changed to Eq. (2.10) and Eq. (2.11).

$$m_{pq} = \sum_{x}\sum_{y} x^p y^q (x,y) \tag{2.10}$$

$$\mu_{pq} = \sum_{x}\sum_{y} (x-\bar{x})^p (y-\bar{y})^q I(x,y) \tag{2.11}$$

Moments η_{pq} where $p+q \geq 2$ can be constructed to be invariant to both translation and changes in scale by dividing the corresponding central moment by the properly scaled $(00)^{th}$ moment, using the following formula:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{(1+\frac{p+q}{2})}} \tag{2.12}$$

Based on (Flusser et al., 2009), the following four moment invariants could be proved invariant to linear motion convolution. Therefore, these moment invariants are also invariant to motion blur.

$$\begin{cases} IR_{m_1} = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\ IR_{m_2} = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\ IR_{m_3} = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ +(3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ IR_{m_4} = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ +(\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{cases} \tag{2.13}$$

The most important property of invariant features is invariance. Based on the analysis before, the moment invariants are satisfied with this property. In moment invariants feature space, the difference between the object part and the distorted part due to motion blur is not large. Another desirable property of invariant features is discriminability. Unfortunately, moment invariants are not useful to distinguish water reflection from imperfect symmetry.

Therefore, the second problem is how to distinguish water reflection images from other imperfect symmetry images. As we described before, motion blur causes a decay of the information and energy in high-frequency band. If our feature space effectively measures the existence of this phenomenon, we could distinguish water reflection images from imperfect images successfully. For this end, we utilize the Curvelet coefficients in high-frequency as one component of proposed feature descriptors. The Curvelet transform is a multiscale pyramid with many directions and positions at each length scale, and needle-shaped elements at fine scales. As the latest multi-directional & multi-scale transform, Curvelet was developed in an attempt to overcome inherent limitations of traditional multiscale representations such as wavelets (Cand`es & Donoho, 2004). Compared with wavelet transform, Curvelet transform has subtle capability to represent directional features in image (Feng et al., 2011).

For water reflection images, the object part contains a wealth of details in various directions, but the reflection part has a decay of the information and energy in high-frequency band. To describe the difference fully and accurately, our technique is based on the Curvelet coefficients in high-frequency band. In Curvelet transform, the work is throughout in two dimensions, i.e.,$\mathbb{R}^2$, with spatial variable $\mathbf{x}=(x,y)\in\mathbb{R}^2$, with the frequency domain variable ω, and with *r* and θ polar coordinates in the frequency-domain. The basic pair of windows includes the "radial window" *W*(*r*) with $r\in(1/2,2)$ and "angular window" *V*(*t*) with $t\in[-1,1]$. Then, the frequency window *U_a* is defined in the Fourier domain as follows:

$$U_a(r,\theta)=2^{-3a/4}W(2^{-a}r)V(\frac{2^{\lfloor a/2\rfloor}\theta}{2\pi}) \tag{2.14}$$

where $a=0,1,...$ is a scale parameter, $\lfloor a/2\rfloor$ is the largest integer below *a*/2. The support of U_a is a polar "wedge" defined by *W* and *V* which is applied with scale-dependent window widths in each direction.

Define the waveform $\varphi_a(\mathbf{x})$ by means of its Fourier transform $\hat{\varphi}_a(\boldsymbol{\omega})=U_a(\boldsymbol{\omega})$. $\boldsymbol{\omega}=(\omega_x,\omega_y)\in\mathbb{R}^2$ is utilized by letting $U_a(\omega_x,\omega_y)$ be the window defined in the polar coordinate system. The equispaced sequence of rotation angle is denoted as $\theta_v=2\pi\cdot2^{-\lfloor a/2\rfloor}\cdot v$, with the orientation parameter $v=0,1,...$ such that $0\le\theta_v\le2\pi$. And the sequence of translation parameter is denoted as $b=(b_x,b_y)\in\mathbb{Z}^2$. With these notations, the Curvelets are defined as function of $\mathbf{x}=(x,y)\in\mathbb{R}^2$ at scale 2^{-a}, orientation θ_v and position $\mathbf{x}_{a,v,b}$ by Eq. (2.15),

$$\varphi_{a,v,b}(\mathbf{x})=\varphi_a(\mathbf{R}_{\theta_v}(\mathbf{x}-\mathbf{x}_{a,v,b})) \tag{2.15}$$

where $\mathbf{x}_{a,v,b}$ is equal to $\mathbf{R}_{\theta_v}^{-1}(b_x\cdot2^{-a},b_y\cdot2^{-a/2})$, $\mathbf{R}_\theta$ is the rotation by θ radians as follows:

$$\mathbf{R}_{\theta} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{2.16}$$

So the Curvelet coefficient map $c_{a,v,b}(x,y)$ is then simply the inner product between an element $I(x,y) \in L^2(\mathbb{R}^2)$ of image and a Curvelet $\varphi_{a,v,b}$.

$$c_{a,v,b}(x,y) = < I(x,y), \varphi_{a,v,b} > \tag{2.17}$$

In digital Curvelet Transforms, similar with the continuous-Time Curvelet transform, U_a smoothly extracts frequencies near the dyadic corona $\{2^a \le r \le 2^{a+1}\}$ and near the angle $\{-\pi \cdot 2^{-a/2} \le \theta \le \pi \cdot 2^{-a/2}\}$. But due to the fact that the coronae and rotation are not especially adapted to Cartesian arrays, in digital Curvelet transform, the "Cartesian coronae" based on the concentric squares and shears are utilized. The basic tiling illustration of digital Curvelet transform is shown in Figure 2.9, in which the red part is the high-frequency spectral band with scale parameter a=10. This particular part is composed of 64 cartesian coronas, and every corona is corresponding to a specific direction v, v =1, 2, …, 64. The Curvelet coefficient map $c_{a,v,b}$ to every coronae could be calculated by Eq. (2.17). The absolute values of the coefficients indicate the strength of the information in specific direction.

In our proposed feature space, the Curvelet coefficients in high-frequency band to every coronae is denoted as $\mathbf{CC}_v$, $1 \le v \le 64$.

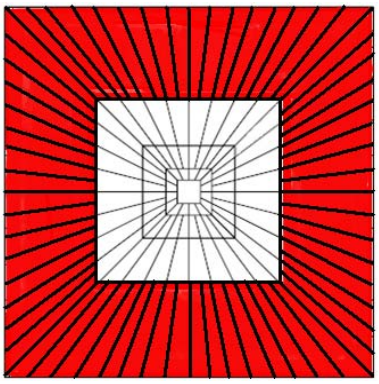


Figure 2.9. The basic tiling illustration of digital Curvelet transform.

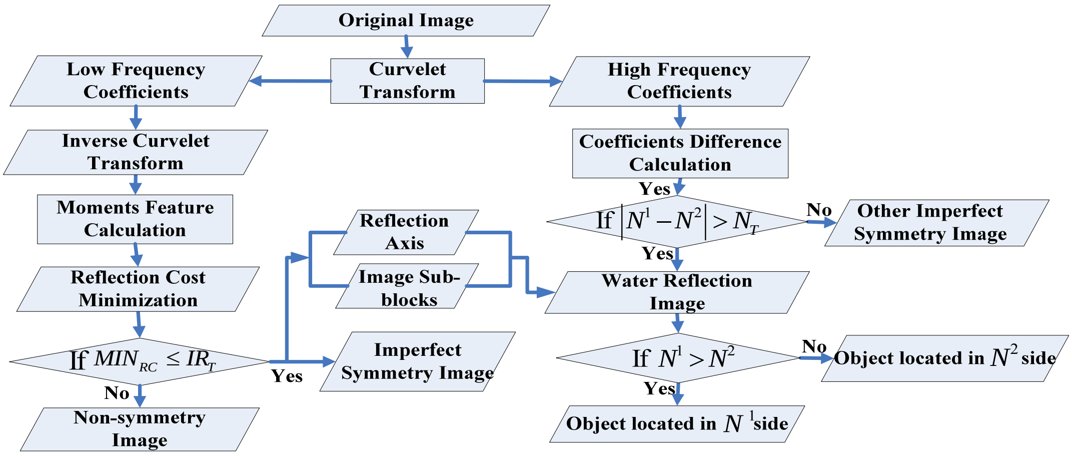

Original Image
Curvelet Transform
Low Frequency Coefficients
High Frequency Coefficients
Inverse Curvelet Transform
Coefficients Difference Calculation
Yes
If $|N^1 - N^2| > N_T$
No
Other Imperfect Symmetry Image
Moments Feature Calculation
Reflection Axis
Image Sub-blocks
Yes
Water Reflection Image
Reflection Cost Minimization
If $MIN_{RC} \leq IR_T$
Yes
Imperfect Symmetry Image
If $N^1 > N^2$
No
Object located in N^2 side
Yes
Object located in N^1side
No
Non-symmetry Image

Figure 2.10. The flowchart of proposed algorithm.

2.5 Recognition of Water Reflection Image

In Section 2.4, we have discussed the limitations of existing feature space and proposed Invariant moment & Curvelet coefficient (IMCC) feature space according to the characteristics of motion blur. Based on the feature space, in this section, we propose two effective sub-algorithms to recognize the water reflection image, including: Reflection cost minimization (RCM) and High-frequency Curvelet coefficients discrimination (HCCD).

Figure 2.10 presents the flowchart of the proposed algorithm. It includes two channels, the low-frequency and high-frequency Curvelet channels. To the first channel, the Curvelet transform is utilized to obtain the low frequency coefficients. We then calculate the moment invariants after using the inverse Curvelet transform on the low frequency coefficients. Based on the moment invariants, we minimize the reflection cost using Dynamic programming (DP) and distinguish the imperfect images from non-symmetry images. To the second channel, the high-frequency Curvelet coefficients are obtained by Curvelet Transform. According to the differences of the coefficients in the image sub-blocks located in both sides of reflection axis, water reflection and imperfect symmetry images are classified into two categories. Furthermore, the object part and the reflection part are then distinguishable from each other.

2.5.1 Reflection cost minimization

In this part, we introduce the sub-algorithm Reflection cost minimization (RCM) for imperfect symmetry recognition. The definition of reflection symmetry is given first.

Definition 1 A set $S \in R^n$ is reflection symmetric with respect to the vector (reflection axis) $< \cos\alpha_0, \sin\alpha_0 >$ with a reflection transform T_{D_K}, if $\forall \mathbf{x}_i \in S, \exists \mathbf{x}_j \in S$, *s.t*,

$$\mathbf{x}_j = T_{D_K}\mathbf{x}_i \tag{2.18}$$

where for $\mathbf{x}_i \in \mathbb{R}^2$, T_{D_K} is given by

$$T_{D_K}(x,y) = \begin{pmatrix} \cos 2\alpha_0 & \sin 2\alpha_0 & 0 \\ \sin 2\alpha_0 & -\cos 2\alpha_0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \tag{2.19}$$

So a centered image $I(\mathbf{x}), \mathbf{x} \in \mathbb{R}^2$, if it has the reflection symmetry with the reflection axis $< \cos\alpha_0, \sin\alpha_0 >$, obeys the Eq. (2.20).

$$I(\mathbf{x}) = I(T_{D_K}(\mathbf{x})), \forall \mathbf{x} \in \mathbb{R}^2 \tag{2.20}$$

In most conditions concerning imperfect reflection symmetry, Eq. (2.20) could not be strictly complied with imperfect symmetry, meaning that in water reflection case, $I(\mathbf{x}) \approx I(T_{D_K}(\mathbf{x})), \forall \mathbf{x} \in \mathbb{R}^2$.

Based on the analysis in Section 2.4, we transform the imperfect symmetry problem into an optimization problem based on the complex moment invariants feature space given by Eq. (2.21).

$$\begin{cases} \alpha_0^* = \arg\min |\sum_{i=1}^{4} IR_{m_i}(I(\mathbf{x})) - \sum_{i=1}^{4} IR_{m_i}(I(T_{D_K}(\mathbf{x})))| \\ \min |\sum_{i=1}^{4} IR_{m_i}(I(\mathbf{x})) - \sum_{i=1}^{4} IR_{m_i}(I(T_{D_K}(\mathbf{x})))| \leq IR_{thresh} \end{cases} \tag{2.21}$$

where IR_{thresh} is the threshold to distinguish imperfect symmetry images from non-symmetry images, and α_0^* is the tilt angle of the reflection axis.

In nature, water reflection or other imperfect symmetry often does not occur in the whole image and the reflection axis is usually not complete or straight. Taking these situations into consideration, we do some simplifications based on the optimized problem shown in Eq. (2.21). First, images are separated into M_s sub-images vertical to the supposed reflection axis direction DI_{α_0}. For every sub-image $I_j, 1 \leq j \leq M_s$, candidate reflection axis is denoted as $RA_{j,l}, 1 \leq j \leq M_s, 1 \leq l \leq H_{\alpha_0}$, where H_{α_0} is the height of the sub-image I_k. The sum difference of moment invariants $DF_{j,k,l}$ of two sub-block $I_{j,k,l}^1(\mathbf{x})$ and its reversed sub-block $I_{j,k,l}^2(\mathbf{x})$ located on both sides of line $RA_{j,l}$ is given in Eq. (2.22):

$$DF_{j,k,l} = \sum_{i=1}^{4}[IR_{m_i}(I^1_{j,k,l}(\mathbf{x})) - IR_{m_i}(I^2_{j,k,l}(\mathbf{x}))] \qquad (2.22)$$

where k is the height of sub-block that is above the threshold T_k.

Reflection axis distance $DS_{j,l}$ is utilized to measure the continuity of the adjacent reflection axis denoted as Eq. (2.23).

$$DS_{j,l} = \begin{cases} \left\| RA_{j.l} - RA_{j+1.l} \right\| / T_d + 1 & j \le M_s - 1 \\ 1 & j = M_s \end{cases} \qquad (2.23)$$

In this equation, $\left\| RA_{j.l} - RA_{j+1.l} \right\|$ is used to describe the vertical distance between the candidate reflection axis in adjacent sub-image I_j and I_{j+1}. T_d is the factor used to normalize the distance to a specified range.

Then we define the reflection cost RC in the current slide window SW_m, $1 \le m \le H_{\alpha_0} - W_{sw}$ which is decided by $DF_{j,k,l}$ and $DS_{j,l}$ in Eq. (2.24). The slide window SW_m with width W_{sw} is horizontal to the candidate reflection axis direction. The location of the centerline in SW_m is denoted as L_m and L_m=m+W_{sw}/2. The minimum of the reflection cost RC in all slide windows is denoted as MIN_{RC}. The optimized reflection axis, which is composed of RA_{j,l^*} in every sub-image I_j of slide window SW_m^* with the minimum of the reflection cost, is denoted as Eq. (2.25).

$$RC = \sum_{j=1}^{M_s}(DF_{j,k,l} \times DS_{j,l}), T_k \le k \le \frac{H_{\alpha_0}}{2},\ l \in SW_m,\ 1 \le m \le H_{\alpha_0} - W_{sw}, -\frac{\pi}{2} \le \alpha_0 \le \frac{\pi}{2} \qquad (2.24)$$

$$[\alpha_0^*, SW_m{}^*, RA_{1,l^*}, RA_{2,l^*}, \ldots RA_{M_s,l^*}] = \arg\min[\sum_{j=1}^{M_s}(DF_{j,k,l} \times DS_{j,l})] \qquad (2.25)$$

The aim of this optimization problem is to find MIN_{RC} and the optimal reflection axis with the corresponding sub-blocks in the image.

The optimization problem we described in Eq. (2.24) and Eq. (2.25) is similar as the one that is often solved by Dynamic programming (DP). DP is both a mathematical optimization method and a computer programming method. In both contexts DP refers to simplifying a complicated problem by breaking it down into simpler sub-problems in a recursive manner. We have a preprocessing work before DP to limit the number of candidate reflection axis $RA_{j,l}$ in every sub image I_j. Then we rank the differences of moment invariants $DF_{j,k,l}$, and only those $RA_{j,l}$ whose difference falls into the M_n minimum value are considered as the candidate reflection axis.

Then we define some basic concepts and variables in DP for water reflection problem. The *Stage variable* $K = j, 1 \le j \le M_s$ is used to describe the current stage or sub-image. The *State variable* λ_K in our algorithm $\lambda_K = RA_{j,l}$ is the candidate reflection axis in the sub-image I_K. The *decision variable* μ_K, in our case, is the choice of the candidate reflection axis $RA_{j+1,l}$ in the next sub image I_{j+1}. The *Transition Function* is defined as $\lambda_{K+1} = \mu_K$. The *Object function* is defined as Eq. (2.26) where v_K is the minimum of reflection cost in stage K denoted in Eq. (2.27).

$$V = \sum_{K=1}^{M_s} v_K(\lambda_K, \mu_K) \qquad (2.26)$$

$$v_K = \min[DF_{j,k,l} \times DS_{j,l}], K = j, T_k \le k \le \frac{H_{\alpha_0}}{2}, l \in SW_m, 1 \le m \le H_{\alpha_0} - W_{sw}, -\frac{\pi}{2} \le \alpha_0 \le \frac{\pi}{2} \quad (2.27)$$

The DP function is defined using Eq. (2.28),

$$\begin{cases} f_K(\lambda_K) = \min\{v_K(\lambda_K, \mu_K) + f_{K-1}(\lambda_{K-1})\} \\ \lambda_K \in \Lambda_K \qquad\qquad K = j, 1 \le j \le M_s \\ \lambda_{K+1} = \mu_K \end{cases} \qquad (2.28)$$

where $f_K(\lambda_K)$ is the minimum of the reflection cost in every stage K in current SW_m. Then we solve the Eq. (2.28) by positive sequence method to get the optimized policy in current slide window SW_m. After that, MIN_{RC} that is the minimum of all RC

in different slide windows and in different α_0 is calculated by Eq. (2.29).

$$MIN_{RC} = \min(f_{M_S}(\lambda_{M_s})), 1 \le m \le H_{\alpha_0} - W_{sw}, -\frac{\pi}{2} \le \alpha_0 \le \frac{\pi}{2} \tag{2.29}$$

Now, we compare the difference of computational complexity between exhaust algorithm and our DP algorithm. For simplicity, we only calculate that the computational complexity to find the optimization axis in direction α_0. To every image, if we utilize the exhaust algorithm to find the global optimization axis in Eq. (2.25), the complexity is: $O((W_{sw}^{M_s} + W_{sw}^{M_s} \times \log_2(W_{sw}^{M_s})) \times H_{\alpha_0})$. To the proposed DP algorithm, the complexity is: $O((M_s - 1) \times W_{sw}^{2} \times \log_2(W_{sw}^{2}) \times H_{\alpha_0})$. It is obvious that the total computational complexity of DP is much lower than that of exhaust algorithm.

2.5.2 High-frequency Curvelet coefficients discrimination

Water reflection is a special case of imperfect symmetry. The algorithm proposed to distinguish imperfect symmetry with non-symmetry is provided in Section 2.5.1. To further the proposal, the optimized reflection axis with the corresponding sub-blocks is obtained. In this part, we propose the sub-algorithm High-frequency Curvelet coefficients discrimination (HCCD) to distinguish water reflection images from imperfect symmetry images.

One important characteristic of motion blur is that it causes a decay of the information and energy in high-frequency band. Therefore, we focus on the high-frequency Curvelet coefficients to address distinguishability of water reflection images from other imperfect symmetry images.

After the Curvelet transform, the Curvelet coefficients in high-frequency band $\mathbf{CC}_v$, $1 \le v \le 64$ is calculated to every coronae. As we known, the absolute values of the coefficients indicate the strength of the information in specific direction. Therefore, for every direction, the differences of the absolute value between every optimized sub-blocks located in both sides of optimized reflection axis are calculated by Eq. (2.30).

$$\mathbf{DC}_{K,v} = \left|\mathbf{CC}_{K,v}^{1}\right| - \left|\mathbf{CC}_{K,v}^{2}\right| \qquad (2.30)$$

where $\mathbf{CC}_{K,v}^{1}$ and $\mathbf{CC}_{K,v}^{2}$ are denoted as the Curvelet coefficients in high-frequency band for direction *v* and sub-block pair I_K^1 and I_K^2 in stage *K*.

To every sub-block pair, we calculate the sum of the absolute value $sp_{K,v}$ and $sn_{K,v}$ in positive and negative part of $\mathbf{DC}_{K,v}$, respectively.

$$\begin{cases} sp_{K,v} = \mathrm{abs}[\sum_{m,n} DC_{K,v}(m,n)] & \text{if } DC_{K,v}(m,n) > 0 \\ sn_{K,v} = \mathrm{abs}[\sum_{m,n} DC_{K,v}(m,n)] & \text{if } DC_{K,v}(m,n) < 0 \end{cases} \qquad (2.31)$$

We count the number of positive Curvelet coefficients in object part and in its reflection part, just as Eq. (2.32),

$$\begin{cases} np_K = \sum_{v=1}^{64} [\varepsilon(\left|sp_{K,v}\right| - \left|sn_{K,v}\right|)] \\ nn_K = \sum_{v=1}^{64} [\varepsilon(\left|sn_{K,v}\right| - \left|sp_{K,v}\right|)] \end{cases} \qquad (2.32)$$

where $\varepsilon(n)$ is the unit step function, and $\varepsilon(n) = \begin{cases} 1 & \text{if } n \geq 0 \\ 0 & \text{if } n < 0 \end{cases}$. If the difference between np_K and nn_K is larger than T_n, just as Eq. (2.33), the image is water reflection. Otherwise, it is imperfect symmetry.

$$I(\mathbf{x}) \text{ is} \begin{cases} \text{water reflection image} & \text{if } \forall k: \left|np_K - nn_K\right| \geq T_n \\ \text{imperfect image} & \text{if } \forall k: \left|np_K - nn_K\right| < T_n \end{cases} \qquad (2.33)$$

Furthermore, in water reflection images, the comparison of np_K and nn_K is helpful to distinguish between the object part and the reflection part. Compared with the reflection part, the object part has more high-frequency information. Therefore, by Eq. (2.34), we could tell the object part from its reflection part easily.

$$\begin{cases} I^1 \text{ is the object part} & \text{if } \forall k : np_K > nn_K \\ I^2 \text{ is the object part} & \text{if } \forall k : np_K < nn_K \end{cases} \tag{2.34}$$

2.6 Performance Evaluation

In this section, we demonstrate the performance of our proposed technique on three experiments, including the classification of nature scene images with and without water reflection, the detection of reflection axis, and the retrieval of water reflection images. In our experiments, we set $M_s = 6$, $T_k = H_{\alpha_0}/5$, $T_d = H_{\alpha_0}/12$, $M_n = H_{\alpha_0}/4$, $T_n = 35$, $W_{sw} = H_{\alpha_0}/25$.

2.6.1 Experiment on water reflection image classification

In the first experiment, to evaluate the classification accuracy of the proposed technique, we construct a dataset including 50 images with water reflection and 50 nature scene images without water reflection. Figure 2.11 and Figure 2.12 present the thumbnails of images with and without water reflection respectively, all of which are utilized in the first experiment.

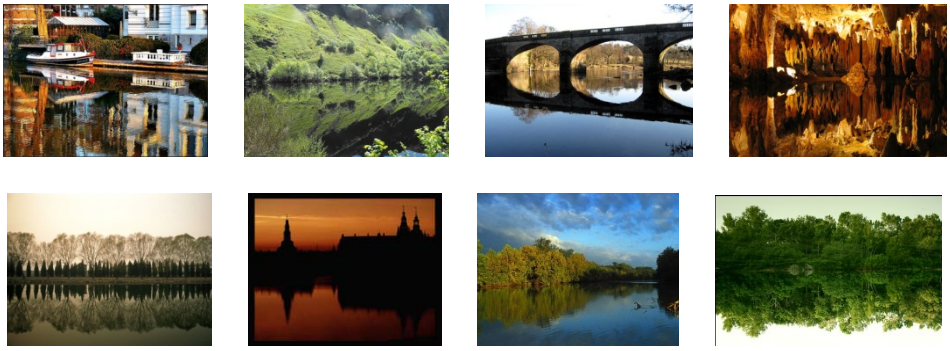


Figure 2.11. Example water reflection images in classification experiment.

Figure 2.12. Example images without water reflection in classification experiment.

We subdivide this dataset equally into five folders, and conduct 5-fold cross validations for the learning algorithms. Every time, we utilize one folder for testing, and the other four folders for training. If MIN_{RC} is below the threshold IR_{thresh} which is learnt by binary SVM classifier based on the training dataset, this image is classified as the water reflection image. The classification accuracy results are provided in Table 2.1. The results prove that our proposed technique based on IMCC features could effectively distinguish the water reflection images from other nature scene images. To evaluate the effectiveness of the proposed moment features, we also provide the classification accuracy using the same algorithm but based on color histogram. Here, WR stands for water reflection images and NWR stands for non water reflection images.

Table 2.1. Water reflection classification accuracy results.

Trail	1	2	3	4	5
WR based on moments	90%	90%	90%	100%	90%
NWR based on moments	80%	70%	80%	90%	80%
Overall based on moments	85%	80%	85%	95%	85%
WR based on color histogram	60%	70%	60%	70%	60%
NWR based on color histogram	50%	50%	70%	80%	70%
Overall based on color histogram	55%	60%	65%	75%	65%

2.6.2 Experiment on detection of the reflection axis

To compare with existing symmetry techniques, the detection experiment on

100 images with water reflection is carried out. The goal of this experiment is to detect the reflection axis. We first compare our technique with the representative technique (Loy & Eklundh, 2006), which utilized the SIFT detection points as interesting salient points and took advantage of pairwise matching of their SIFT descriptors to detect the axis of symmetry. We also compare ours with the only paper on detection of water reflection axis proposed by Zhang et al. based on the shape detector (Zhang, 2010). For the reflection axis detection experiment, the accuracy of SIFT algorithm (Loy & Eklundh, 2006) is 29%, and the accuracy of the Shape technique (Zhang, 2010) is 46%. Our proposed technique RCM achieves 87% accuracy. Examples are given in Figure 2.13 to illustrate the detection results of the three techniques.

It is obviously that our technique is more effective than the techniques of (Loy & Eklundh, 2006) and (Zhang, 2010). Due to the limitations of SIFT detectors and descriptors discussed in Section 2.4, it is predictable that the accuracy of the technique (Loy & Eklundh, 2006) is low. The technique (Zhang, 2010) utilized the flip invariant shape detector relying on the completeness of the shape. Unfortunately, water reflection is a complex and various phenomena often with incomplete and distorted shape in reflection part, which leads to the ineffectiveness of technique just as two examples in Figure 2.14.

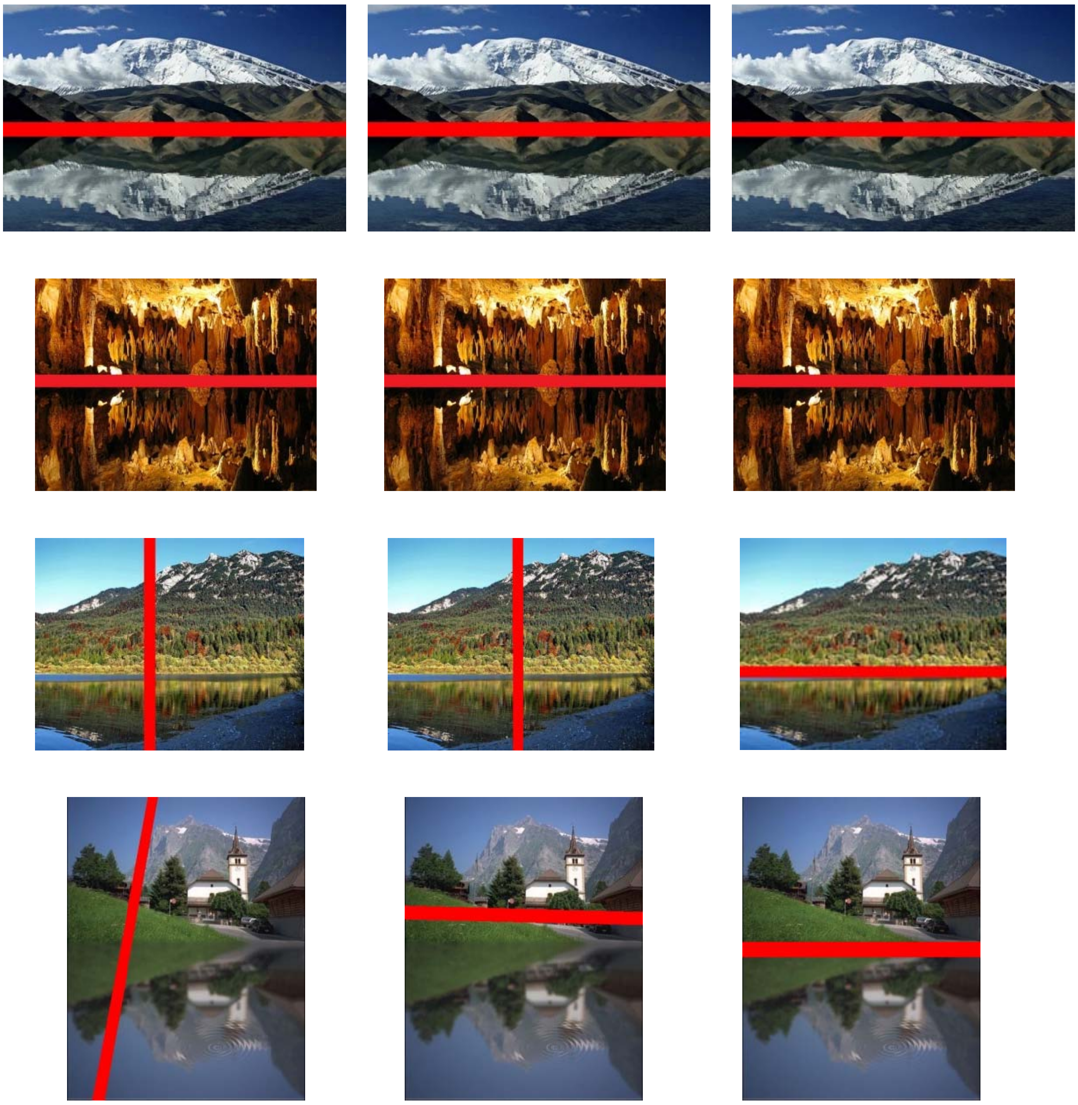


Figure 2.13. Performance comparison of reflection axis detection. First, second and third column are the results of Shape, SIFT and proposed RCM respectively.

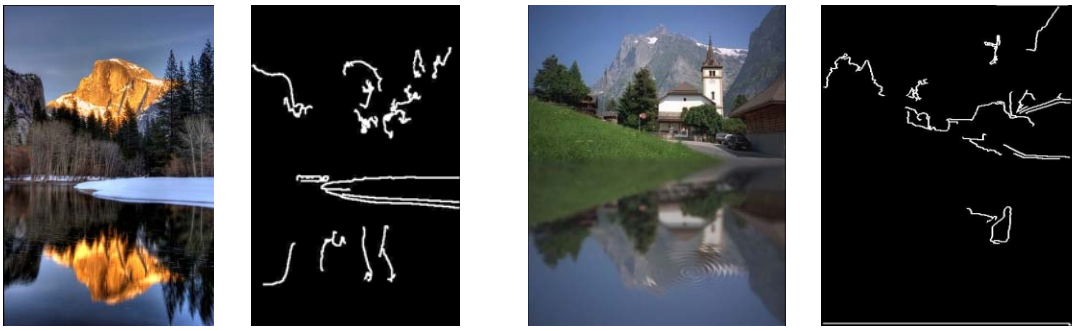


Figure 2.14. Examples of shape detection results of invariant shape technique.

2.6.3 Experiment on water reflection image retrieval

We then apply the proposed technique in text based image retrieval for evaluation. The textual query is “water reflection”, every image that is related to this concept is returned. The dataset is downloaded from Google and is composed of two parts. The first part is 50 images with water reflection, and the second part contains 10000 images without water reflection. Figure 2.15 shows the thumbnails of images with and without water reflection used in the retrieval experiment. Different from the nature scene images utilized in classification experiment, the images in the retrieval experiment are more diversified, and include imperfect symmetry images.

Many different measures for evaluating the performance of image retrieval systems have been proposed. In our experiment, we use four popular ones: precision, recall, Average Precision (AveP) and Normalized discounted cumulative gain (NDCG). Precision is defined as the fraction of the images retrieved that are relevant to the user's information need in the information retrieval system. Recall is the fraction of the images that are relevant to the query that are successfully retrieved.

$$\text{Precision} = \frac{N_{Relevant} \cap N_{Retrieved}}{N_{Retrieved}} \qquad (2.35)$$

$$\text{Recall} = \frac{N_{Relevant} \cap N_{Retrieved}}{N_{Relevant}} \qquad (2.36)$$

where $N_{Relevant}$ is the number of images which are relevant to the query and $N_{Retrieval}$ is the number of images that are finally retrieved out.

In the proposed technique, the Curvelet coefficients in high-frequency part are utilized to distinguish the water reflection image from the imperfect symmetry image. In this dataset, we first demonstrate the retrieval performance with or without the contribution of Curvelet coefficients part. Table 2.2 shows the Precision and Recall comparison results. The number of retrieval sample is from 10 to 50 with increments of 10. It is obviously that the high-frequency coefficients are helpful to achieve a better classification. Two examples of imperfect symmetry images which are correctly distinguished by Curvelet coefficients are given in Figure 2.16.

Precision and recall are single-value metrics based on the whole list of multimedia documents returned by the retrieval system. For systems that return a ranked sequence of images, it is also desirable to consider the order in which the returned images are presented. Average precision emphasizes ranking relevant images higher and is computed in Eq. (2.37) at the point of each of the relevant images in the ranked sequence,

$$\mathrm{AveP}@\,p = \frac{\sum_{i=1}^{p} (P_i \times Rel_i)}{N_{Relevant}} \tag{2.37}$$

where p is the rank position, Rel_i is a binary function on the relevance of a given rank, and P_i is the precision at a given cut-off rank where $N_{RR}(i)$ is the number of relevant retrieved images of rank i or less:

$$P_i = \frac{N_{RR}(i)}{i} \tag{2.38}$$

The premise of DCG is that relevant documents appearing lower in a search result list should be penalized, as the graded relevance value is reduced logarithmically with a proportion to the position of the result. The DCG accumulated at a particular rank position p is defined as Eq. (2.39). For a query, the normalized discounted cumulative gain, or NDCG, is computed as Eq. (2.40), where IDCG is the ideal DCG at position p. Figure 2.17 shows the results for AveP and NDCG of our proposed technique.

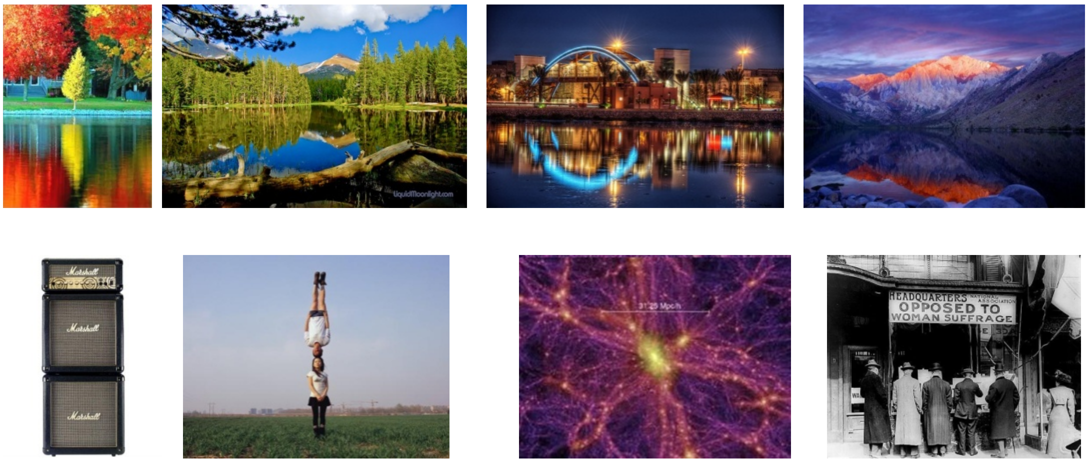


Figure 2.15. Example images with and without water reflection in retrieval experiment.

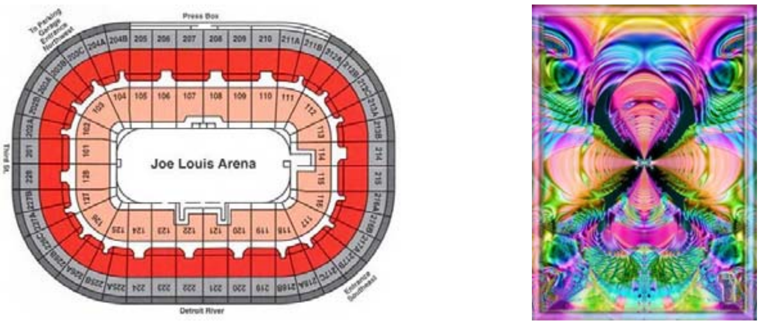


Figure 2.16. Examples of imperfect symmetry images that are correctly distinguished by Curvelet coefficients.

Table 2.2. Water reflection precision and recall results.

Retrieval Number	10	20	30	40	50
Precision without Curvelet	70%	70%	73%	73%	72%
Recall without Curvelet	14%	28%	44%	58%	72%
Precision with Curvelet	90%	90%	90%	83%	86%
Recall with Curvelet	18%	36%	54%	66%	86%

$$\mathrm{DCG@}p = Rel_1 + \sum_{i=2}^{p} \frac{Rel_i}{\log_2 i} \qquad (2.39)$$

$$\mathrm{NDCG} @ p = \frac{\mathrm{DCG} @ p}{\mathrm{IDCG} @ p} \qquad (2.40)$$

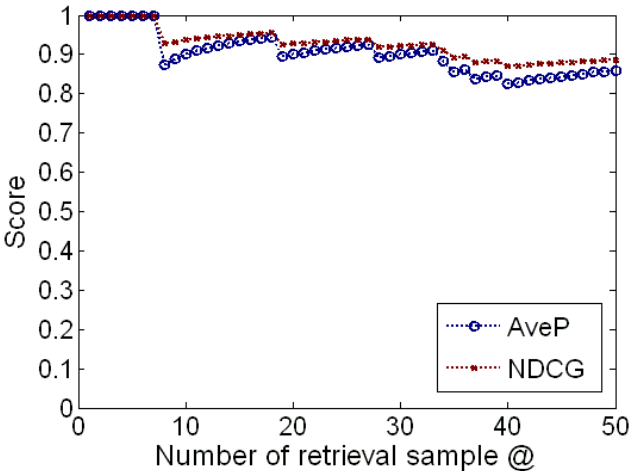
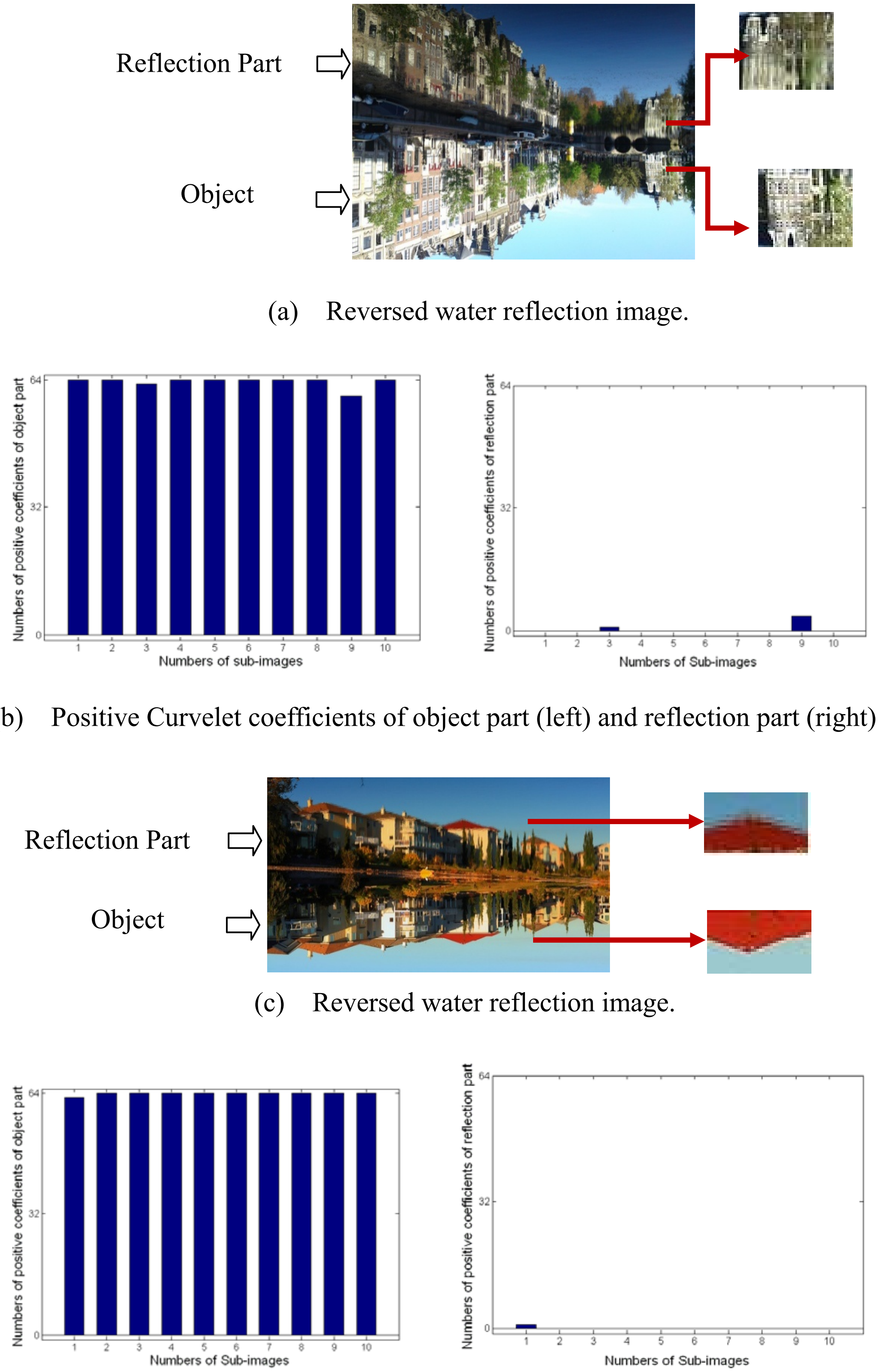


Figure 2.17. AveP and NDCG results of the retrieval experiment.

2.6.4 Discussion

According to the three experiments and a series of evaluations, our proposed feature descriptors and techniques are proven effective in water reflection detection and recognition. Moreover, our technique has an additional advantage in detecting which part is the object in water reflection images. Figure 2.18 gives two sample images from our dataset, both of which have been uploaded upside down. Due to the similarity of object and reflection parts, such case is even difficult to be detected by human eyes in the original image size. As described previously, the object part tends to have larger Curvelet coefficients in high-frequency band. Thus a comparison of Curvelet coefficients located on the both sides of the reflection axis could help determine the object part easily and correctly.



(a) Reversed water reflection image.

(b) Positive Curvelet coefficients of object part (left) and reflection part (right).

(c) Reversed water reflection image.

(d) Positive Curvelet coefficients of object part (left) and reflection part (right).

Figure 2.18. Object part and reflection part determined by Curvelet coefficients.

2.7 Summary

In this chapter, we propose a novel feature space Invariant moment & Curvelet coefficient (IMCC) based on the characteristics of retina image formation for the task of object detection and recognition. IMCC is composed of motion blur invariant moments in low-frequency Curvelet space and of Curvelet coefficients in high-frequency Curvelet space. An effective and efficient algorithm is then constructed based on IMCC and applied to the water reflection recognition and reflection axis detection. Experiments and evaluation all confirm the effectiveness of our technique, which is more reliable and successful compared with existing feature space and algorithms.

Chapter 3 Attention Allocation for Image Saliency Detection and Quality Assessment

3.1 Overview

This chapter proposes a novel attention model called Bi-directional saliency map (BSMP) (Zhong et al., 2010a) (Zhong et al., 2011b) for the task of image saliency detection and quality assessment.

Construction of saliency map in multimedia data is useful for applications in multimedia like object segmentation, object recognition and quality assessment. In this chapter, our novel attention model integrates bottom-up saliency features and top-down targets information together. In bottom-up direction, the proposed Gabor & Curvelets based saliency map relies on 2D Gabor and Curvelet transforms. Compared with the traditional bottom-up model based on DOG and wavelets, our model takes advantage of Garbor transforms's spatial localization and Curvelet transform's edge and directional information. In top-down direction, the proposed Target based saliency map relying on the center priority and target location. Compared with existing top-down model emphasizing the importance of center or some empirical high-level features, such as face detection region, our model concerns about the cognitive understanding by referring the tag information.

Empirical validations on standard dataset demonstrate the effectiveness of the bottom-up saliency detection, and top-down saliency detection. Furthermore, we apply Bi-directional saliency map (BSMP) for the image quality assessment task. Our technique outperforms the representative blurriness metrics.

3.2 Related Work

As a cognitive process of selectively concentrating on one aspect of the environment while ignoring other things, attention has been referred to as the allocation of processing resources (Anderson, 2004). In human vision system (HVS), to encode detailed visual information, eyes need to be moved so that this area is focused on the visual locations in which are interested (Styles, 2005). As the most famous attention model, saliency map is proposed to measure of conspicuity and calculate the likelihood of a location to attract attention (Koch & Ullman, 1985).

Owing to the models of image saliency provide predictions about which regions are likely to attract observers' attention (Parkhurst et al., 2002), automatic detection of visually salient regions is useful in different multimedia applications. These applications include content aware resizing (Avidan, & Shamir, 2007), quality assessment (Zhong et al., 2010), segmentation (Fukuchi et al., 2009), object detection and object recognition (Yu et al., 2010).

There are two different kinds of processing in attention, bottom-up and top-down processing. In existing works on visual saliency detection, most of them focus on the bottom-up processes of HVS. Typically, multiple low-level visual features such as intensity, color, orientation, texture and motion are extracted at multiple scales. After a feature map is computed, they are normalized and combined into a master saliency map that represents the saliency of each pixel (Itti & Koch, 2000). Nearly all of existing bottom-up models were inspired by the theories from biology, psychology and neuropsychology (Wang et al., 2000). Among them, the most famous one was proposed by (Itti et al., 1998). They developed the center surround structure akin to on-type and off-type visual receptive field. In recent years, more proposed work simulated the multi-scale and multi-orientation function of primary visual cortex V1, Achanta et al. detected the saliency map with a Difference of Gaussians (DOG) model to describe the spatial properties of visual regions (Achanta et al., 2009). Gabor filters and Log-Gabor wavelets are utilized to explore the salient features such as spatial localization, spatial frequency characteristics in (Harel et al., 2007) and (Wang et al., 2000) respectively.

The DOG, wavelets and Gabor transforms are prevalent in saliency map construction in recent years; nevertheless, both of them have some inherent drawbacks. The DOG is a wavelet mother function of null total sum which approximates the Mexican Hat wavelet by subtracting a wide Gaussian from a narrow Gaussian. Compared with DOG, wavelets have the ability to capture the scale-space information in details. But wavelets are ill-suited for detecting or providing a compact representation of intermediate dimensional structures, for example, wavelets are very crude in representing directional features. The principal motivation to use Gabor transforms is biological relevance that the receptive field is oriented and has characteristic spatial frequencies. But due to the elimination of spectra overlap, ‘holes’ are created in the spectra plane of Garbor transforms, which causes loss of spectral information, especially the edge and fine directional information. As the latest multi-directional & multi-scale transform, Curvelet transforms have subtle capability to resolve directional feature than wavelet transform and improved ability to represent edges and other singularities along curves.

Due to the difficulty in refining the goal of attention in natural images (Fergus et al., 2003), little work about saliency detection simulates the top-down processing of human visual system. In (Judd et al., 2009), Judd et al. collected eye tracking data and utilized the dataset to train a model of saliency based on low, middle and high-level image features. From their work, the bottom-up saliency model does not match the actual eye movements. On the contrary, some high-level image features such as detection results are more related with the human's attention allocation. But how to integrate the semantic-level information into the top-down processing is still a problem. To address the problem, we propose a novel top-down processing method to obtain the semantic-level information with the aid of rich tag information from Internet. Now, many web applications, such as Flickr, allow users to upload photos with their own annotated tags, which generally indicate the objects users concerned or the targets they took photos. In our proposed model, the targets information from given tags are integrated into the top-down processing.

Image quality assessment is a criterion able to score (on a scale) the quality of a tested image which may have been distorted. Assessing the quality of images automatically in agree with human's judgment, is the requirement of objective image quality assessment. Objective image assessment can be divided into three categories: full-reference, no-reference (blind quality assessment), and reduced-reference. In many practical applications, however, the reference image cannot be available, and a no-reference approach is desirable. Most existing no-reference quality assessment techniques focus on measuring the sharpness of the images based on some characteristics of human visual system (HVS) (Ferzli, & Karam, 2009) (Varadarajan & Karam, 2008) (Sadaka et al., 2008). Recently, attention modeling is considered as an effective method to extract important information and evaluate quality (Sadaka et al., 2008). Therefore, we extend the proposed saliency map model into the no-reference image quality assessment task.

In this chapter, we propose a novel approach for saliency detection in natural images, and present a novel image quality assessment metric under the guidance of Bi-directional saliency map (BSMP). In bottom-up direction, the proposed method in this chapter takes advantage of 2D Gabor's spatial localization and Curvelet transforms' edge and directional information. In top-down direction, in order to simulate the top-down processes in human visual system, we consider the semantic meaning from tag information and the influence of center bias into our model. The Bi-directional saliency map combined with Just noticeable blur (JNB) algorithm (Ferzli & Karam, 2009) to get the image quality assessment.

The rest of chapter is organized as follows. Section 3.3 presents a novel saliency map construction method Bi-directional saliency map (BSMP) in detail. The image quality assessment metric based on BSMP discusses in Section 3.4. Experimental results are given in Section 3.5 and the chapter is concluded in Section 3.6.

3.3 Saliency Map Detection

3.3.1 Saliency detection using bottom-up 2D Garbor and Curvelet transform

Gabor filter, named after Dennis Gabor, is a linear filter used in image processing. In the spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave. The 2D Gabor function in space domain is defined as below:

$$g(x_1,x_2)=c_s(x_1,x_2)g_s(x_1,x_2) \tag{3.1}$$

where $c_s(x_1,x_2)$ is a complex sinusoid, known as the carrier, and $g_s(x_1,x_2)$ is a 2D Gaussian function, known as the envelope. $c_s(x_1,x_2)$ is denoted as Eq. (3.2).

$$c_s(x_1,x_2)=\exp[2\pi iF_g(x_1\cos\omega_g+x_2\sin\omega_g)] \tag{3.2}$$

where $F_g=\sqrt{u_g^2+v_g^2}$, $\omega_g=\tan^{-1}(v_g/u_g)$, i.e. $u_g=F_g\cos\omega_g$ and $v_g=F_g\sin\omega_g$. And (u_g,v_g) are the spatial frequencies of the sinusoid carrier in Cartesian coordinates.

The Gaussian envelope $g_s(x_1,x_2)$ is given as follows:

$$g_s(x_1,x_2)=\exp[-\pi a(x_1-x_{1g})_t^2-\pi b(x_2-x_{2g})_t^2] \tag{3.3}$$

where a,b are the scales of the two axis in the Gaussian envelope, (x_{1g},x_{2g}) is the location of the peak of the Gaussian envelope. The rotation and translation transformation is denoted as below, where θ_g is the rotation angle of the Gaussian envelope.

$$\begin{cases}(x_1-x_{1g})_t= & (x_1-x_{1g})\cos\theta_g+(x_2-x_{2g})\sin\theta_g\\(x_2-x_{2g})_t= & -(x_1-x_{1g})\sin\theta_g+(x_2-x_{2g})\cos\theta_g\end{cases} \tag{3.4}$$

A special member of the emerging family of multiscale geometric transforms is the Curvelet transform. It was developed in an attempt to overcome inherent limitations of traditional multiscale representations such as wavelets (Cand`es & Donoho, 2004). The Curvelet transform is a multiscale pyramid with many directions and positions at each length scale, and needle-shaped elements at fine scales. In Section 2.4.3, the Curvelet transform is introduced in detail. Here, we directly utilize it to build the Curvelet feature map.

Since visual neurons are often excited by one color and inhibited by opponent color, we choose preattentive features as the red/green (RG), blue/yellow (BY) color and intensity. The 2D Gabor and Curvelet transform functions $g(x_1, x_2)$ and $\varphi(x_1, x_2)$ as we described before are utilized to build the feature map of 2D Gabor $F_G(x_1, x_2)$ and the feature maps $F_C(x_1, x_2)$ of Curvelet as follows:

$$F_G(x_1, x_2) =< f(x_1, x_2), g > \tag{3.5}$$

$$F_C(x_1, x_2) =< f(x_1, x_2), \varphi > \tag{3.6}$$

The feature maps $F_G(x_1, x_2)$ and $F_C(x_1, x_2)$ are then across-scale combined and normalized into activation maps within Gabor and Curvelet channels based on the commonly used combination and normalization method in (Itti et al., 1998) (Harel et al., 2006). The activation maps are the components to simulate the bottom-up processing of attention.

3.3.2 Saliency detection using top-down target information

The target information of the image is acquired with the help of corresponding tags. To avoid the interference of irrelevant or trivial tags, we use a lexicon to remove all tags that do not belong to the 'physical entity' group. WordNet is a lexical database for the English language (Miller, 1995). Based on the function of finding the hypernym of words, WordNet can be utilized as the lexicon to remove irrelevant or trivial tags. So the tag information is automatically transformed to the target information and then is used to calculate the saliency regions.

According to the research in (Judd et al., 2009), the center priority is an important feature to represent semantics of the image because human photographers tending to place objects of interest in the center of photographs. Therefore, we also integrate the center priority as a part of top-down information into our model.

3.3.3 Bottom-up and top-down channels integration

The unsupervised Bi-directional saliency map (BSMP) framework is illustrated in Figure 3.1. In bottom-up channel of this framework, we utilize the 2D Gabor filters and Curvelet transforms to build the feature maps. In top-down channel of this framework, the center bias and target detection are utilized to construct the feature maps and activation maps. These activation maps are linearly combined together.

The supervised Bi-directional saliency map (BSMP) framework is illustrated in Figure 3.2. Here, instead of combining the top-down and bottom-up activation maps linearly, SVM is utilized to learn the saliency map model based on the eye fixation points. To choose the positively and negatively labeled pixels for saliency model learning by SVM, a ground truth map according with the contrast sensitivity research is built (Wang & Bovik, 2001). The function of contrast sensitivity as a function of pixel position (*x*,*y*) is given by:

$$S_c = \frac{e_2 \ln(\frac{1}{CT_0})}{\alpha[e_2 + \tan^{-1}(\frac{d(x,y)}{Lv})]} \tag{3.7}$$

where α is spatial frequency decay constant; e_2 is half resolution eccentricity constant (degrees); L is the image width (measured in pixels), v is the viewing distance from viewer to computer screen (measured in image width). $d(x,y)$ is the distance from (*x*,*y*) to the fixation point. The ground truth map *I* is created by convolving the function of contrast sensitivity over top N fixation locations for all M users.

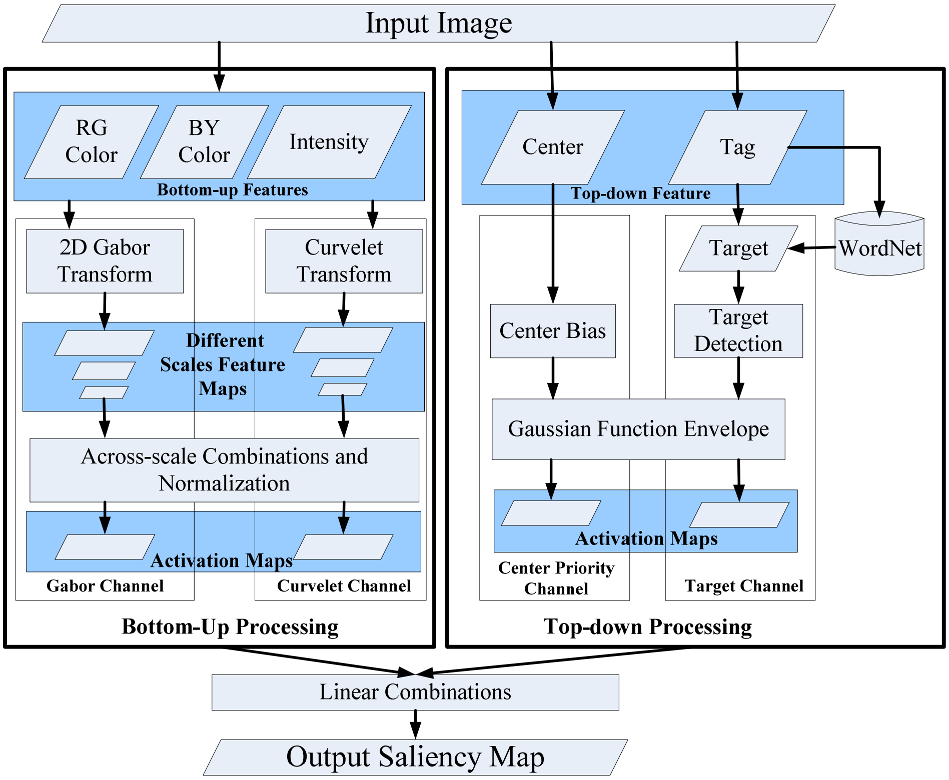


Figure 3.1. Framework of BSMP under unsupervised framework.

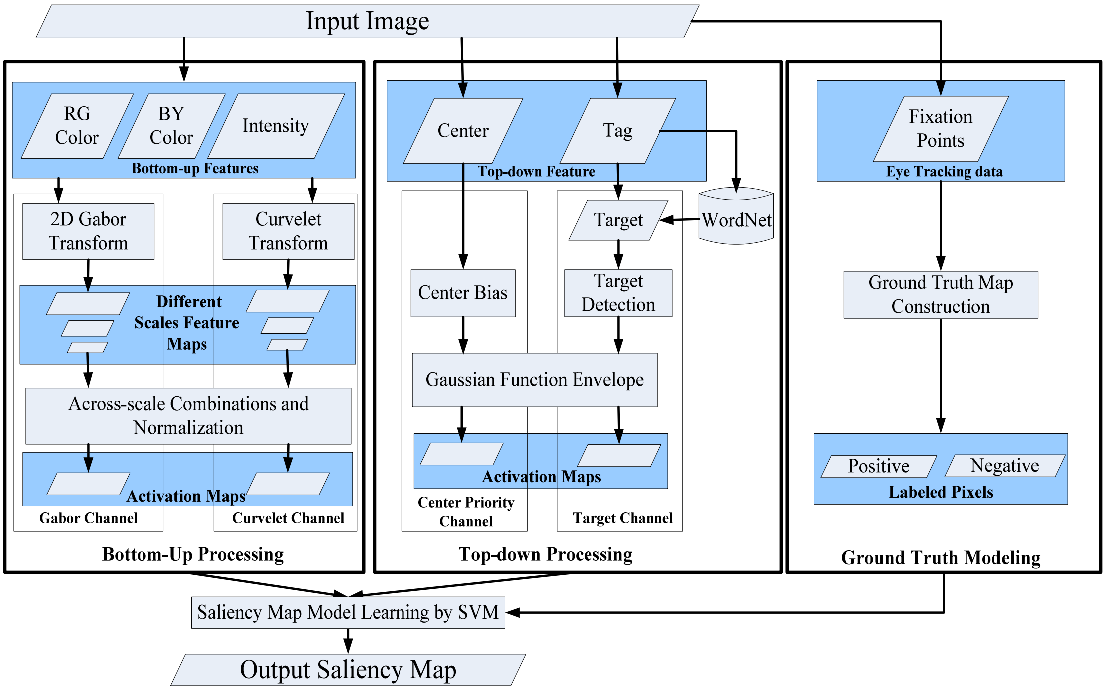


Figure 3.2. Framework of BSMP under supervised framework.

$$g(x,y)=\sum_{i=1}^{M}\sum_{j=1}^{N}\delta_{i,j}(u-f_x(i,j),v-f_y(i,j))\otimes\frac{1}{e_2+\tan^{-1}(\frac{d(x-u,y-v)}{Lv})} \tag{3.8}$$

$$I(x,y)=g(x,y)/\max_{x,y}(g(x,y)) \tag{3.9}$$

$$d(x,y)=\sqrt{x^2+y^2} \tag{3.10}$$

After generating the ground truth map, we choose the saliency locations as positively labeled pixels and the non-saliency locations as negatively labeled ones to learn the Bi-directional saliency map (BSMP) by SVM.

3.4 Image Quality Assessment based on Saliency Map

In this part, we combine the saliency guidance to measure the blurriness of image. Firstly, the input image is divided into 64×64 blocks. Then the JNB metric (Ferzli & Karam, 2009) is used to calculate the blurriness in every local edge block. The perceived blur distortion within an edge R_b is given by:

$$D_{R_b}=(\sum_{e_i\in R_b}\left|\frac{W(e_i)}{W_{JNB}(e_i)}\right|^{\beta})^{\frac{1}{\beta}} \tag{3.11}$$

where β, which sets as 3.6 in (Ferzli & Karam, 2009). It is chosen to increase the correspondence of Eq. (3.11) with the experimentally determined psychometric function. $W(e_i)$ is the measured width of the edge and $W_{JNB}(e_i)$ is the JNB width which depends on the local contrast C which is defined as the magnitude of the difference between the maximum and minimum intensities. W_{JNB} is modeled as follows:

$$W_{JNB}=\begin{cases}5 & C\le 50\\ 3 & C>50\end{cases} \tag{3.12}$$

Then the saliency guidance combined with blurriness in each block is utilized to assess the image quality. Blur distortion in saliency regions D_s and non-saliency regions D_{ns} are defined as Eq. (3.13), where R_{bs} is the salient part in R_b, and $Num(\cdot)$ is to calculate the number of pixels in corresponding region.

$$D=\begin{cases}D_s=(\sum_{R_b}\left|D_{R_b}\right|^{\beta})^{\frac{1}{\beta}} & Num(R_{bs})/Num(R_b)>\alpha_{thresh}\\ D_{ns}=(\sum_{R_b}\left|D_{R_b}\right|^{\beta})^{\frac{1}{\beta}} & Num(R_{bs})/Num(R_b)\le\alpha_{thresh}\end{cases} \tag{3.13}$$

The proposed objective sharpness metric is given by:

$$S=\alpha_s\cdot(\frac{L_s}{D_s})+(1-\alpha_s)\cdot(\frac{L_{ns}}{D_{ns}}) \tag{3.14}$$

where L_s, L_{ns} are the numbers of saliency blocks and non-saliency blocks in the image. α_s is the weight of saliency part to the sharpness metric.

3.5 Performance Evaluation

In this section, we evaluate the performance of our saliency detection model and corresponding image quality assessment metric. The saliency detection model is evaluated on a public image dataset with collected eye tracking data on 1003 images (Judd et al., 2009). This dataset is the largest dataset with eye tracking data which is popularly utilized in saliency map construction.

The first experiment is to compare the bottom-up processing part of BSMP with four other bottom-up saliency detection models, including basic Itti's model (Itti et al., 1998), Graph model based on Gabor filter (Harel et al., 2006), DOG model

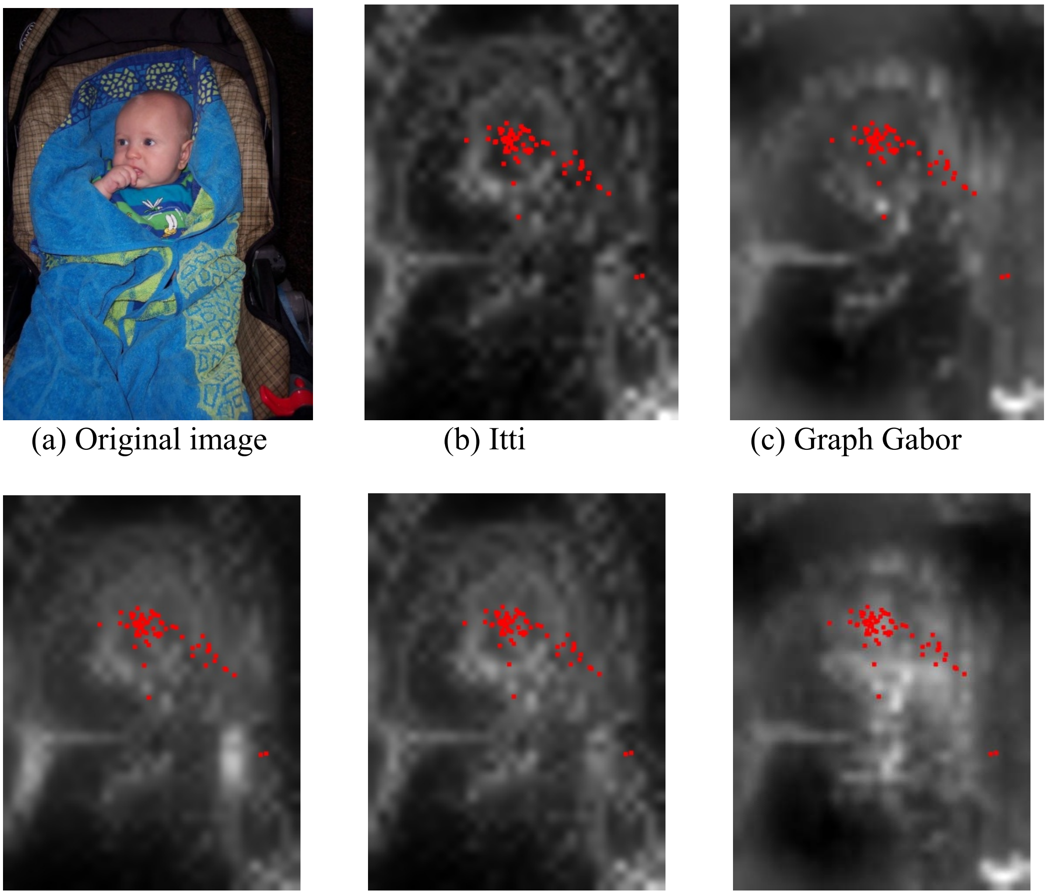
(Achanta et al., 2009) and 2D Log-Garbor model (Wang et al., 2010).

In the second experiment, the top-down processing and the bottom-up processing will be integrated together. The saliency detection performance is discussed under the unsupervised and supervised learning framework.

The third experiment provides the performance for the image quality assessment task. The representative evaluation metrics of Classical JNB (Ferzli & Karam, 2009), Saliency weighted JNB (Sadaka et al., 2008), and JNB with edge refinement (Varadarajan & Karam, 2008) will be compared with proposed metric.

3.5.1 Experiment on bottom-up saliency detection

We firstly demonstrate the experimental results on public image dataset with eye tracking data (Judd et al., 2009) based on bottom-up processing, any features extracted based on the top-down processing are not considered into our model. Firstly, the comparison of the saliency maps constructed by our technique with other techniques is shown in Figure 3.3. Whiter means the corresponding value in saliency map is higher. The comparison algorithms include: Itti (Itti et al., 1998), Graph Gabor (Harel et al., 2006), DOG (Achanta et al., 2009), and Log-Gabor (Wang et al., 2010) and proposed Bottom-up BSMP saliency map. Although all techniques focus on salient changes to capture attention, the saliency regions detected by our model concentrates more on to the baby's face with more fixation points. Therefore, our method can predict where human look more efficiently and more accurately.



(a) Original image (b) Itti (c) Graph Gabor

(d) DOG (e) 2D Log-Garbor (f) Bottom-up BSMP

Figure 3.3. The comparison of saliency detection results. (a) Original image. From (b) to (f) is the saliency maps with human fixations marked as red dots. (b) Itti's saliency map, (c) Graph Gabor saliency map, (d) DOG saliency map, (e) 2D Log-Garbor saliency map, (f) Bottom-up BSMP saliency map.

Furthermore, the ROC areas of these techniques are shown in Table 3.1. We could easily observe from Table 3.1 that our model has the largest ROC area and achieves the best overall performance. The comparison algorithms include: Itti (Itti et al., 1998), Graph Gabor (Harel et al., 2006), DOG (Achanta et al., 2009), and Log-Gabor (Wang et al., 2010) and proposed Bottom-up BSMP saliency map.

Table 3.1. ROC area comparison based on bottom-up models.

Model	Itti	Graph Gabor	DOG	Log-Gabor	**Bottom-up BSMP**
ROC area	0.6736	0.6820	0.6816	0.6874	**0.6990**

3.5.2 Experiment on bi-directional saliency detection

The center bias is a common phenomenon when humans look in natural scenes, which has been considered into top-down attention model. After integrating the center bias into the proposed model and all other compared models, we could obtain the average ROC areas over all users and all images in Table 3.2. The comparison algorithms include: Itti (Itti et al., 1998), Graph Gabor (Harel et al., 2006), DOG (Achanta et al., 2009), and Log-Gabor (Wang et al., 2010) and Bottom-up BSMP with center bias. From Table 3.2, it can be seen that center bias could improve the performance greatly. The performance improvement of proposed technique is statistically significant ($p<0.05$).

The target information is another important component in top-down processing; we add the face detection channel as target channel into our model to obtain the saliency map. The ROC area increases from 0.8180 to 0.8281 in the images with human. But to the images without human, the false alarm of the detector will lead to the performance have a 1.08% reduction, from 0.8086 to 0.7999. Therefore, this result proves that the reliable tag information is useful to build a better saliency map even in an unsupervised learning fashion.

Under the supervised learning framework, in every training image, we randomly choose 30 saliency pixels as positively labeled data from the 10% most salient locations and 30 non-saliency pixels as negatively labeled data from the 10% least salient locations. The experiment results show that 84.156% saliency points detected by our model are inside and 76.344% non-saliency points are outside the saliency regions judged by human. This performance is better than the results from existing saliency map modeling methods (Judd et al., 2009).

Table 3.2. ROC area comparison with center bias.

Model	Itti	Graph Gabor	DOG	Log-Gabor	**Bottom-up BSMP with center bias**
ROC area	0.7763	0.8169	0.8095	0.8176	**0.8200**

3.5.3 Experiment on image quality assessment

For image quality evaluating, we use the images from Flickr with the tag “people”. People or its hyponym of words are the most popular tags in Flickr and more than 4 million images in Flickr use this tag. We randomly select 160 images with the tag or the hypernym of tag of “people”. We averagely partition the images into eight groups blurred with eight different 7×7 Gaussian masks of σ values 0.8, 1.6, 2.0, 2.4, 3.2, 4.0, 4.8 and 5.6. For each displayed image, fourteen subjects are asked to rate the quality of the images in terms of perceived blurriness using a scale from 1 to 5 corresponding to “Very annoying”, “Annoying”’, “Slightly Annoying”, “Perceptible but not annoying”, and “Imperceptible”, respectively.

We provide the correlation analysis between the objective measures and the Mean opinion scores (MOS). In our model, we set the parameters $\alpha_{thresh}=1/3$, $\alpha_s=0.8$, and $\alpha_{ns}=0.2$. Table 3.3 shows the comparison of different sharpness metrics in terms of the Nonlinear Pearson, Spearman, Mean absolute error (MAE) and Root mean squared error (RMS) coefficients after nonlinear regression. Obviously, our technique has better results under most of cases. Except RMS, the performance of our technique is better than other techniques. The performance of our technique does not achieve best performance on RMS due to RMS is sensitive to some extreme prediction values. Generally speaking, integrating the saliency model will improve the performance. However, if the saliency model doesn’t consist with the real fixation points, it may hurt the performance. It is the reason that the performance of Saliency weighted JNB (Sadaka et al., 2008) and JNB with edge refinement (Varadarajan & Karam, 2008) is even lower than Classical JNB (Ferzli & Karam, 2009).

Table 3.3. Evaluation of the proposed image quality metric performance.

	Nonlinear Pearson	Spearman	MAE	RMS
Proposed metric	**0.914**	**0.860**	**0.173**	0.250
JNB	0.885	0.815	0.219	0.292
Saliency weighted JNB	0.863	0.801	0.317	**0.232**
JNB with edge refinement	0.618	0.466	0.387	0.494

3.6 Summary

In this chapter, we propose an effective saliency detection model Bi-directional saliency map (BSMP) and extend it into the image quality assessment task. In bottom-up direction, we utilize Garbor transforms based on its spatial localization ability and Curvelet transforms based on its better directional and edges representation ability. In top-down direction, target information and center priority are integrated into the model to provide the semantic meaning for saliency map modeling.

In experimental results, our bottom-up BSMP has the highest ROC area in comparison with other three state-of-the-art bottom-up techniques. With top-down processing channel, proposed Bi-directional saliency map achieves the best performance. Moreover, in no-reference image quality evaluation task, our metric exhibits increased correlation with perceived quality of subjects.

Chapter 4 Perceptual Modeling for Image Annotation

4.1 Overview

This chapter proposes a novel label to region assignment technique called Fuzzy-based contextual-cueing label propagation (FCLP) based on the characteristics of perceptual processing in human visual system (Zhong et al. 2010b) (Zhong et al., 2012a).

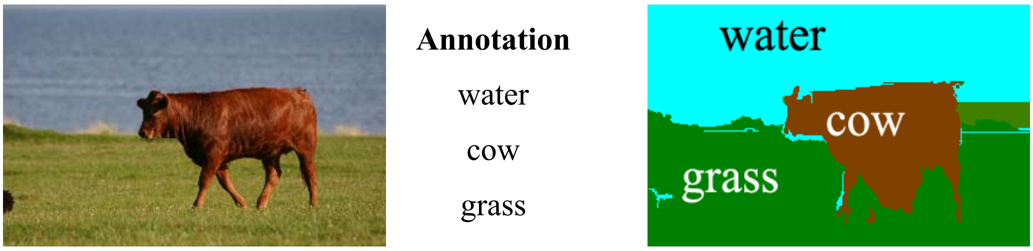
Label to region assignment (LRA) can be defined as the automatic assignment of the image-level annotations to the precise regions of the image. With LRA techniques, the tedious manual region-level annotations can be substituted (Liu et al., 2009). Furthermore, LRA is useful to achieve reliable content-based image retrieval which is one of the ultimate questions to multimedia content analysis. Hence, this chapter proposes a novel region level annotation technique to this challenging task.

Fuzzy-based contextual-cueing label propagation (FCLP) broadly includes four stages. First, an image is over-segmented into a set of image patches and the visual features are extracted to represent the image patch. In the second step, fuzzy representation and logic are utilized to model the spatial invariants in contextual cueing, especially the position information and spatial topological relationships. Third, labels are propagated inter and intra images in visual and contextual cueing space, respectively. Finally, the Fuzzy C-means clustering based on *k*-nearest neighbor (KNN-FCM) is used to segment images into semantic regions with corresponding annotations. Experiments on two public datasets are used to evaluate the effectiveness of proposed technique FCLP and other representative techniques.

4.2 Background and Motivation

LRA is the assignment of the image-level annotations to the precise regions within the given image. For example, Figure 4.1 (a) is an image with three image-level annotations, including “water,” “cow,” and “grass.” The aim of LRA is to segment the given image to several regions and associate image-level annotations to the corresponding semantic regions as shown in Figure 4.1 (b).

Early work on LRA is known as the simultaneous image segmentation and object recognition. With the unsupervised learning framework, Cao et al. handled images with single object and clean background (Cao & Li, 2007). With the supervised learning framework, J. Li et al. focused on special object recognition and image segmentation for images in sports domain (Li et al., 2009). The WordNet is utilized to refine the image-level annotations to improve the recognition accuracy. The proposed technique demonstrated impressive performance for the images of badminton, bocce, croquet, polo, rock climbing, rowing, sailing, snowboarding, and *etc*.



(a) (b)

Figure 4.1. Example of LRA. (a) An image with given image-level annotations (b) The label to region assignment result.

The latest work for real-world applications demonstrated good performance for natural images. X. Li et al. proposed Bi-layer sparse coding and label propagation technique for LRA (Liu et al., 2009). The basic idea of their work is that the regions in different images with the common annotation are more likely to have similar visual features. Their techniques demonstrated distinguished performance for images with multiple objects or with complex background.

However, visual similarity does not work for all the cases in region level annotation. Figure 4.2 (a) shows an ordinary image with sky, sea and sand. Similar with (Liu et al., 2009), we use Scale-invariant feature transform (SIFT) descriptors as the features of the uniform sampled points shown in Figure 4.2 (b). It is obvious the values of the local features from different data points are similar. Figure 4.2 (c)-(f) compare the SIFT feature descriptors from four random points, and two data points are from sky while two data points are from sea. Obviously, the difference of SIFT descriptors between (d) and (e) is even smaller than the difference between (d) and (c). Therefore, if only relying on the visual similarity based on SIFT descriptors, we cannot distinguish sky and sea. Actually, the problem is not resulted from SIFT descriptor. One interesting observation is that human has no problem in distinguishing sea and sky easily. In real life, human have seen similar views or pictures many times in environment, so that they have formed the prior knowledge that the sky is generally above the sea. Such kind of prior knowledge was formally defined as contextual cueing by psychologists, the manner in which human brains gather information by incidentally learned associations between target locations and spatial configurations (Chun & Jiang, 1998).

In this chapter, we try to provide more semantic understandings of images for LRA task with the aid of contextual cueing. Different from contextual information, such as, the synchronized or unsynchronized logs and texts associated with multimedia data, which have been widely utilized to understand videos (Jiang et al., 2009), contextual cueing is seldomly studied by multimedia content analysis society. One possible reason is that the bivalent set theory causes semantics loss in describing contextual cueing, such as position information and spatial topologic. In classical bivalent set theory, the membership of elements in a set is assessed in binary terms according to a bivalent condition —either belongs or does not belong. By contrast, fuzzy set theory permits the gradual assessment of the membership of elements in a set with the aid of a membership function valued in the real unit interval [0, 1]. To address this problem, we utilize fuzzy representation and fuzzy logic to substitute the classical bivalent set theory, to model and infer the relationships of the semantic regions within an image.

We introduce the contextual cueing with fuzzy theory in Section 4.3. A novel LRA technique FCLP is proposed in Section 4.4. Section 4.5 provides the performance comparisons on two public datasets. The chapter is closed with conclusion in Section 4.6.

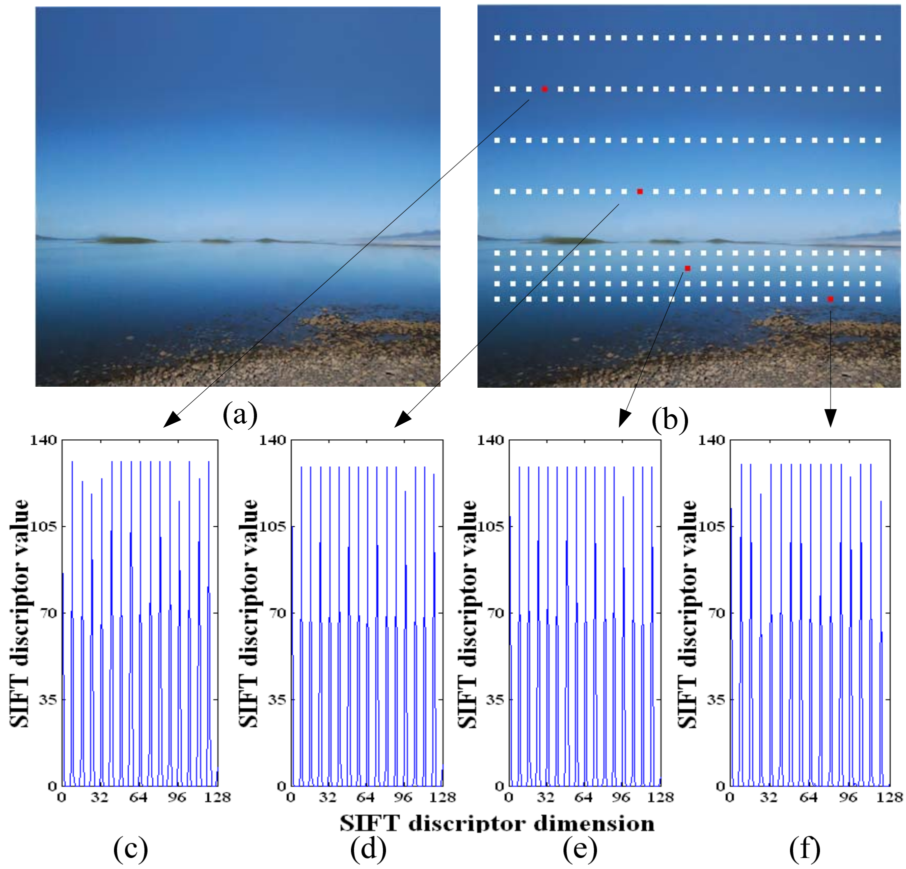
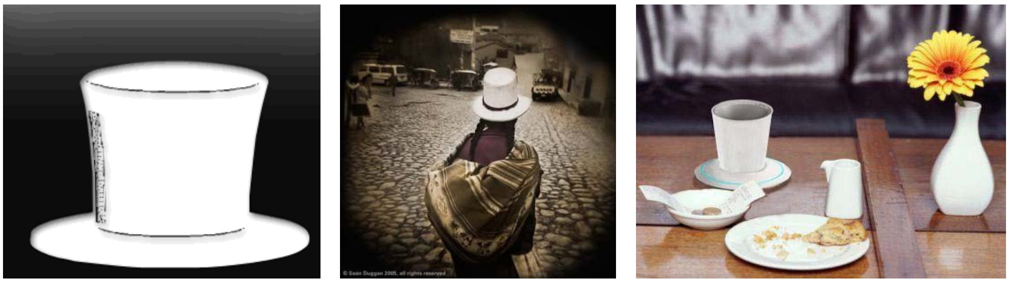


Figure 4.2. Example of the difficulty of distinguishing sky and sea only based on visual feature. (a) The original image. (b) The original image with 200 data points. (c) – (f) 128 SIFT features of four random points in original image.

4.3 Contextual Cueing with Fuzzy Theory

The semantics of each object depends on the context it is regarded within. In 1998, Chun and Jiang originally proposed the concept of contextual cueing in psychology (Chun & Jiang, 1998). It refers to the manner in which human brain gathers information from the visual elements and their surroundings. Generally, the information is acquired incidentally from past experiences of regularities of the visual world, and gradually forms the knowledge about spatial invariants. The convictive support to the rationality of contextual cueing comes from the behavior experiments in psychology (Biederman et al., 1974) (Potter, 1976) (Intraub, 1981) (Davenport & Potter, 2004). It is found that in behavior experiment, subjects have the ability to extract semantic information from presentations as brief as 80 ms before saccading all portions of picture, and even before the recognition of individual objects (Davenport & Potter, 2004). How is the semantic information extracted so rapidly? The contextual cueing makes a significant contribution of facilitating object recognition by reducing the possibilities of locations that need to be considered. Furthermore, from the research in neuroscience, some brain areas of human visual system have the function of contextual cueing. For example, orbital Prefrontal cortex (PFC) is involved in producing guesses and expectations, and the projections from PFC can be directly connected to area Inferior temporal (IT) and to the amygdale (Bar, 2004). Hence, contextual cueing is useful to address the semantic gap in the LRA. Five types of spatial invariants are thought to be important in contextual cueing (Biederman et al., 1982):

1) Probability: the likelihood that certain objects will be present in a scene
2) Co-occurrence: the likelihood that certain objects will be present together
3) Size: the familiar relative size of objects
4) Position: the typical positions of some objects in some scenes
5) Spatial topological relationship: left of, right of, above, below, surround, inside, and *etc*.



(a) (b) (c)

Figure 4.3. Example of how contextual cueing is used to resolve ambiguity for the recognition. (a) An example of ambiguous object: hat or cup? (b) A hat on the head (c) A cup surrounding some dishware.

The spatial invariants can guide the visual attention, speed the visual elements search, and help the object recognition (Chun, 2000). In cases where recognition cannot be accomplished quickly based only on the appearance attributes, contextual information can provide more relevant input for the recognition (Bar, 2004). Figure 4.3 is an example to illustrate how contextual cueing works to resolve ambiguity for object recognition. It is very difficult to distinguish which the object is in Figure 4.3 (a), a cup or a hat, because their appearances are very similar. But in Figure 4.3 (b), it is easy to recognize the target object as a hat when the related object (head) appears below it. In Figure 4.3 (c), the surrounding dishware is helpful to disambiguate the identity of the object.

Based on these theories and evidences in psychology and neuroscience, this chapter intends to integrate the contextual cueing in label region assessment to provide a human-like judgement of the image. Two issues should be addressed in this chapter. One important issue is how to model the acquired knowledge of contextual cueing, i.e., spatial invariants. The second issue is how to model the formation of contextual cueing, i.e., the learning process of knowledge acquirement. Here, fuzzy membership is used to quantize the degree of truth for these spatial invariants. Moreover, we use fuzzy reasoning to imitate human's learning process of contextual cueing.

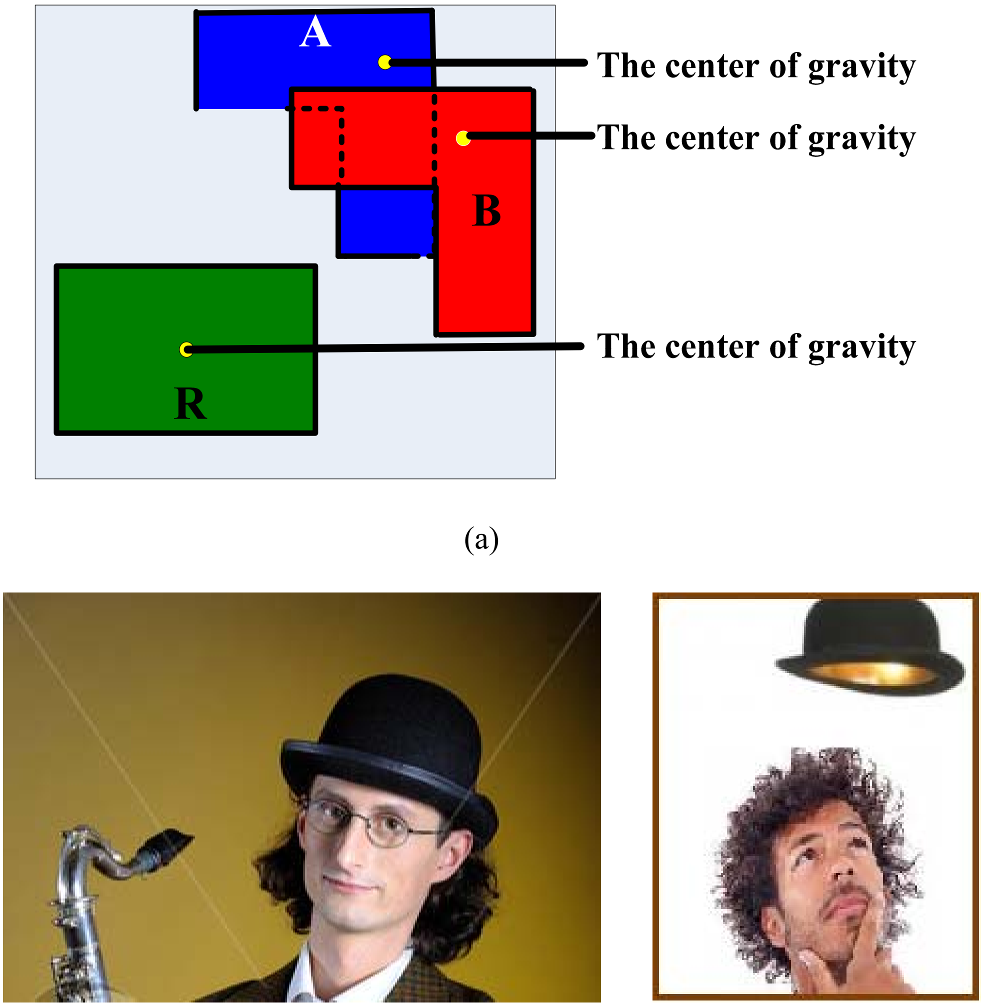
For modeling of spatial invariants, the probability and co-occurrence information are available with regard to the problem of LRA, because in LRA, the image-level annotations are known in advance. And the object size has been considered in (Liu et al., 2009). Some studies known as simultaneous image segmentation and object recognition, have demonstrated performance improvement even only one or two kinds of spatial invariants are considered (Yuan et al., 2007)

(Galleguillos et al., 2008).

However, little work has been conducted to explore spatial invariants for LRA problem of natural images, mainly due to the difficulty of modeling position information and spatial topological relationships for the images with multiple objects and complex background. For example, in Figure 4.4(a), to the reference object *R*, object *A* and *B* are located in similar position. But all parts of object *A* are above *R*, while some parts of object *B* are below *R*. How can we describe this kind of difference? Furthermore, even minor difference of position information and spatial topological relationships will influence the object recognition. In Figure 4.4 (b), the objects both are located above the human's head in two images, but one is hat, another is lamp. How can we represent this position information and spatial topological relationships? In classical bivalent set theory, one object cannot be above and below another object at the same time. Therefore, we do not consider the classical bivalent set theory here due to its semantic loss. We must find a method to address the difficulty in modeling position information and spatial topological relationships for multiple objects.

Fuzzy theory is thought to be a powerful tool to model position information and spatial topological relationships. Different from the probability-based theory that measures the likelihood an event occurs, fuzzy theory measures the degree that an event occurs. For example, if we only use the center of gravity to represent an object in Figure 4.4 (a), the degree of position *top*, *middle* and *bottom* is about 0.76, 0.24 and 0. The position of object *A* could be denoted as $R_{Position}^{A} = \{\frac{0.76}{top}, \frac{0.24}{middle}, \frac{0}{bottom}\}$. The details of representation and calculation of spatial invariants will be discussed in Section 4.4.2.

In this chapter, aiming to model the spatial invariants of contextual cueing, we utilize fuzzy theory to measure the degree of the truth. Specifically, we use fuzzy membership to quantize the degree of truth for spatial invariants. And we use the fuzzy logic and fuzzy reasoning to make decision based on the fuzzy membership.



(a)

(b)

Figure 4.4. Topological relationship for object recognition. (a) Example of topological relationship (b) Example of topological relationship for object recognition.

4.4 Fuzzy Based Contextual Cueing Label Propagation

In this chapter, a novel Fuzzy-based contextual-cueing label propagation (FCLP) technique is proposed for label to region assignment task. As shown in Figure 4.5, first, an image is over-segmented into atomic image patches. Then, visual features and spatial invariants are extracted from each patch. Color and texture, as visual features, are coded based on two Bag-of-words (BOW) codebooks. Size, fuzzy position and fuzzy spatial topological relationship are utilized as the spatial invariants. Labels are propagated inter images using Bi-layer sparse coding based on the visual features. According to the similarity calculation of spatial invariants between the segmented patches and the acquired concepts, labels are propagated intra images. Finally, the post processing via KNN-FCM assign the given image-level annotations to corresponding image regions.

The algorithm overview corresponding to Figure 4.5 is given in Section 4.4.1. The details of FCLP will be discussed from Section 4.4.2 to Section 4.4.5.

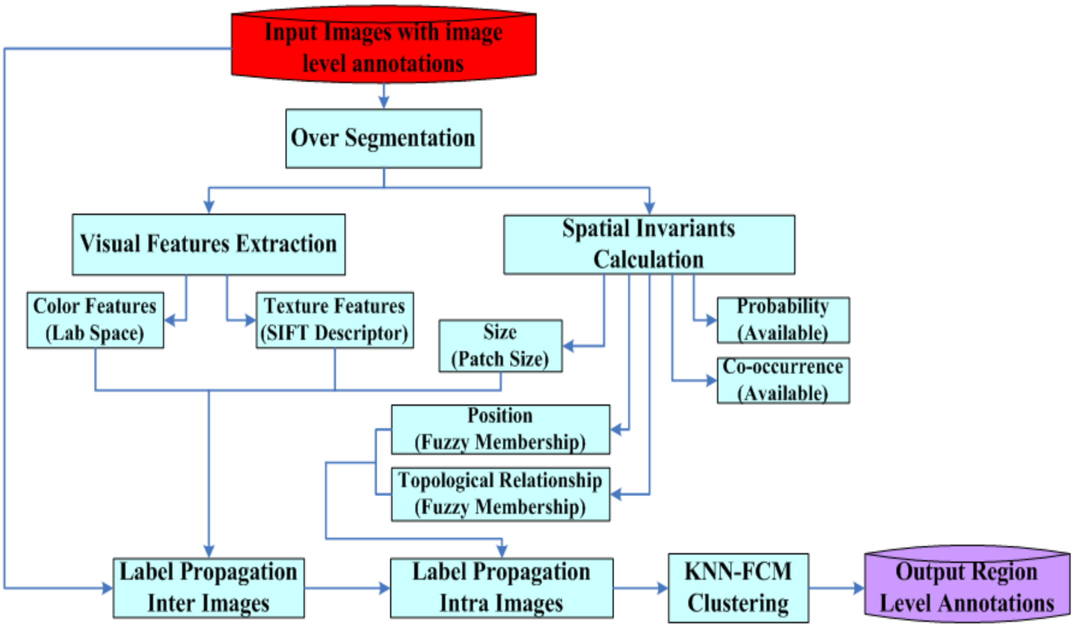

Input Images with image level annotations
Over Segmentation
Visual Features Extraction
Spatial Invariants Calculation
Color Features (Lab Space)
Texture Features (SIFT Descriptor)
Size (Patch Size)
Probability (Available)
Co-occurrence (Available)
Position (Fuzzy Membership)
Topological Relationship (Fuzzy Membership)
Label Propagation Inter Images
Label Propagation Intra Images
KNN-FCM Clustering
Output Region Level Annotations

Figure 4.5. Sketch of FCLP technique.

4.4.1 Algorithm overview

The overview algorithm is described as follows:

Step 1: All images are over-segmented into atomic image patches to ensure the segmented patches are involved within an object/concept;

Step 2: Visual features and spatial invariants are extracted from each patch based on Section 4.4.2;

Step 3: Labels are propagated inter images using Bi-layer sparse coding based on Section 4.4.3;

Step 4: Calculate the spatial invariants of concepts including the fuzzy position and the fuzzy spatial topological relationship based on Section 4.4.4;

Step 5: Calculate the similarity of spatial invariants between the segmented patches and the acquired concepts;

Step 6: Labels are propagated intra images based on Algorithm 4.1;

Step 7: KNN-FCM is utilized to generate cluster, those patches with in a same cluster are merged to form a semantic region. The final region-level label is set as the one with the largest value in the label vector based on Section 4.5.

4.4.2 Image representation

Two different kinds of features are utilized to present the image, visual features and spatial invariants. In pre-processing stage, we advocate segmenting images into multiple segmentations. Like the existing techniques (Cao & Li, 2007) (Li et al., 2009) (Liu et al., 2009), to ensure each segmented patch be involved within an object/concept, we utilize a modified version of a graph-based over-segmentation algorithm (Felzenszwalb & Huttenlocher, 2004). This method incrementally merges smaller-size patches with similar appearances and with small minimum spanning tree weights. Different with the original over-segmentation algorithm, we initialize each pixel as one atomic patch. Then, we use the color features to describe the appearance of an initial image patch and apply the over-segmentation algorithm (Felzenszwalb & Huttenlocher, 2004) for merging smaller patches into larger ones.

Visual features

After the over-segmentation stage, the feature representation is obtained for each atomic patch. Each atomic patch is described by using Bag-of-words (BOW) features generated by color (in *Lab* space) and texture features (SIFT descriptor). Within each region, a number of interest points are detected by SIFT detector. In some small patches, SIFT detector would not detect any interest points. In this case, N_s points are randomly picked from each patch as the selected points. A codebook includes two parts are obtained for the selected points and the region appearance by unsupervised *k*-means clustering. One is obtained based on the SIFT descriptor of the selected points. Another is obtained based on the *Lab* color of the selected points and the average *Lab* value of whole region. The visual feature of an atomic patch $\mathrm{x}_{i,j}$ in image $\mathrm{x}_i, i=1,...N$ could be denoted as an *m*-dimensional descriptor feature $\mathrm{x}_{i,j} \in \mathbb{R}^m$, $j=1,2,...,n_i$, where *N* is the number of image dataset, and n_i is the number of patches in image x_i.

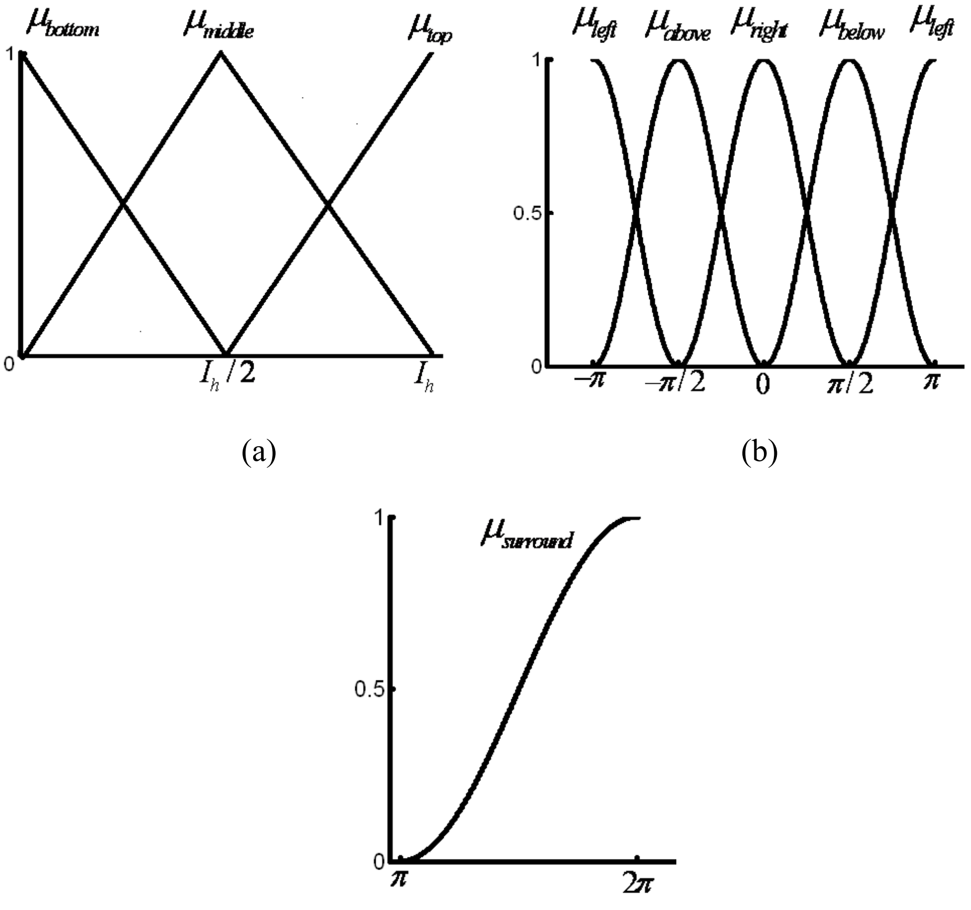
Spatial invariance

For the modeling of spatial invariants, the probability and co-occurrence information are available with regard to LRA task, because the image-level annotations are known in advance. And the object size has been considered in (Liu et al., 2009). Hence, we mainly discuss two kinds of spatial invariants here, including: fuzzy position and spatial topological relationship.

To represent these spatial features in each atomic patch, the location of the patch needs to be determined at first. Based on the balance of the computational complexity and the representational ability, we choose the center of gravity and the contour points as typical points to represent each image patch.

Fuzzy position is used to represent the typical positions in images. Based on the research by Torralba et al., vertical location is very important for image understanding and one object category is likely to be within a horizontal section of the image (Torralba et al., 2006). Therefore, we use the vertical location “top,” “middle,” and “bottom,” to characterize the position of object. The fuzzy membership of position is defined in Figure 4.6 (a) by a commonly used triangular function, where I_h is the height of the image. The fuzzy membership of position is calculated on the typical points. The average fuzzy value of these points is defined as the fuzzy position of patch *j* which is denoted to be

$R^{P}_{Position}(j)=\{\frac{\mu^{P}_{top}(j)}{top},\frac{\mu^{P}_{middle}(j)}{middle},\frac{\mu^{P}_{bottom}(j)}{bottom}\}$. To distinguish with the fuzzy membership for concept, Letter "*P*" is used as a superscript to indicate it is the fuzzy membership for patch.



(a) (b)

(c)

Figure 4.6. Fuzzy membership for the position and the spatial relationship. (a) Fuzzy membership for the position "top," "middle," "bottom." (b) Fuzzy membership for the spatial relationships in four directions. (c) Fuzzy membership for the spatial relationship "surround."

The spatial topological relations are influenced by the direction and distance between two patches. Existing work defined the primitive spatial topological relations involving direction and distance information by using terms such as "right of,", "left of,", "below,", "far below,", "above,", "far above," "surround," and "inside.".

For image x_i, we first calculate the angle θ made by the line passing through the typical points' pairs $p_i(j_1, j_2)$ belonging to patch x_{i,j_1} and x_{i,j_2} ($j_1 = 1,...n_i, j_2 = 1,...n_i, j_1 \neq j_2$). Then, the dominant angle θ_d between two patches is determined based on the angle histograms of contour points in two patches. Based on the dominant angle θ_d, the membership functions of four spatial topological relationships between x_{i,j_1} and x_{i,j_2} are defined in Figure 4.6 (b) (Miyajima & Ralescu, 1994). The membership function of the fuzzy set "surround" is given by Eq. (4.1) and Figure 4.6 (c), where θ_r is the range of angle θ. Taking the distance of the center of gravity into consideration, the membership function of "far above" is defined by Eq. (4.2), where g_{j_1} and g_{j_2} are the center of gravity in patch x_{i,j_1} and x_{i,j_2}, respectively.

$$\mu_{surround}(\theta_r) = \begin{cases} \cos^2(\theta_r / 2) \text{ if } \pi \leq \theta_r \leq 2\pi \\ 0 \quad \text{otherwise} \end{cases} \tag{4.1}$$

$$\mu_{far\ above}(j_1, j_2) = \begin{cases} 1 \wedge (1.5 \times \|g_{j1} - g_{j2}\|_2 / I_h) \text{ if } \mu_{above}(j_1, j_2) > 0 \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{4.2}$$

The opposite relationship pairs of the fuzzy membership is denoted as Eq. (4.3).

$$\begin{cases} \mu_{left}(j_1, j_2) = \mu_{right}(j_2, j_1) \\ \mu_{above}(j_1, j_2) = \mu_{below}(j_2, j_1) \\ \mu_{surround}(j_1, j_2) = \mu_{inside}(j_2, j_1) \end{cases} \tag{4.3}$$

The fuzzy spatial topological relationship between patch j_1 and j_2 in image x_i is defined as $R^{PP}_{Spatial}(j_1, j_2) \in F(P \times P)$ of $P \times P$=$\{(\mathrm{x}_{ij_1}, \mathrm{x}_{ij_2}) | \mathrm{x}_{ij_1} \in P \wedge \mathrm{x}_{ij_2} \in P\}$,

$R^{PP}_{Spatial}(j_1, j_2)$ could be denoted as follow:

$$\begin{aligned} &R^{PP}_{Spatial}(j_1, j_2) \\ &= \{\frac{\mu^{PP}_{right}(j_1, j_2)}{right}, \frac{\mu^{PP}_{left}(j_1, j_2)}{left}, \frac{\mu^{PP}_{below}(j_1, j_2)}{below}, \frac{\mu^{PP}_{far\ below}(j_1, j_2)}{far\ below}, \\ &\frac{\mu^{PP}_{above}(j_1, j_2)}{above}, \frac{\mu^{PP}_{far\ above}(j_1, j_2)}{far\ above}, \frac{\mu^{PP}_{surround}(j_1, j_2)}{surround}, \frac{\mu^{PP}_{inside}(j_1, j_2)}{inside}\} \end{aligned} \tag{4.4}$$

4.4.3 Label propagation inter images

Bi-layer sparse coding proposed by Liu et al. is utilized to construct a linear combination between image patches with the same annotation in visual feature space (Liu et al., 2009). Label propagation inter images is based on the linear combination coefficients obtained by Bi-layer sparse coding. In label propagation inter image, the label is propagated from the candidate region to the selected patches of the remaining images and vice versa. Supposed that $\mathrm{z}_i \in \mathbb{R}^{\mathrm{n}_k}$ indicates the annotation vector, n_k is the total number of image-level annotations. The binary element $\mathrm{z}_i(k)$ takes 1 if the i^{th} image contains the k^{th} annotation and 0 otherwise. Firstly, the patch-level annotation vector $\{z_{i,j}\}$, $\mathrm{z}_i \in \mathbb{R}^{\mathrm{n}_k}$ is initialized with annotation vector $\mathrm{z}_i \in \mathbb{R}^{\mathrm{n}_k}$ of image x_i. Then for every candidate patch j_1 from image x_{i_1}, if $\hat{\alpha}_{i_1,j_1,i_2,j_2} > 0$, z_{i_2,j_2} and z_{i_1,j_1} are updated by Eq. (4.5), where $i_1, i_2 = 1,...N;\ j_1, j_2 = 1,...n_i$. And β_{i_2,j_2} is the weight coefficient calculated based on the size of the j_1 th atomic patch and normalized by the image size of the i_1 th image.

$$\begin{cases} z_{i_2,j_2} = z_{i_2,j_2} + \hat{\alpha}_{i_1,j_1,i_2,j_2} \\ z_{i_1,j_1} = z_{i_1,j_1} + \sum_{i_2=1}^{N}\sum_{j_2=1}^{n_i}(\hat{\alpha}_{i_1,j_1,i_2,j_2} \times \hat{\beta}_{i_2,j_2}) \end{cases} \tag{4.5}$$

After label propagation inter images, the patch-level annotation vector $\{z_{i,j}\}$ is updated. Based on $\{z_{i,j}\}$, we define $w_{i,j,k}^{0} = z_{i,j,k} / \overset{n_k}{\underset{k'=1}{\vee}} (z_{i,j,k'})$, which is the initial membership value of the patch x_{ij} to label k.

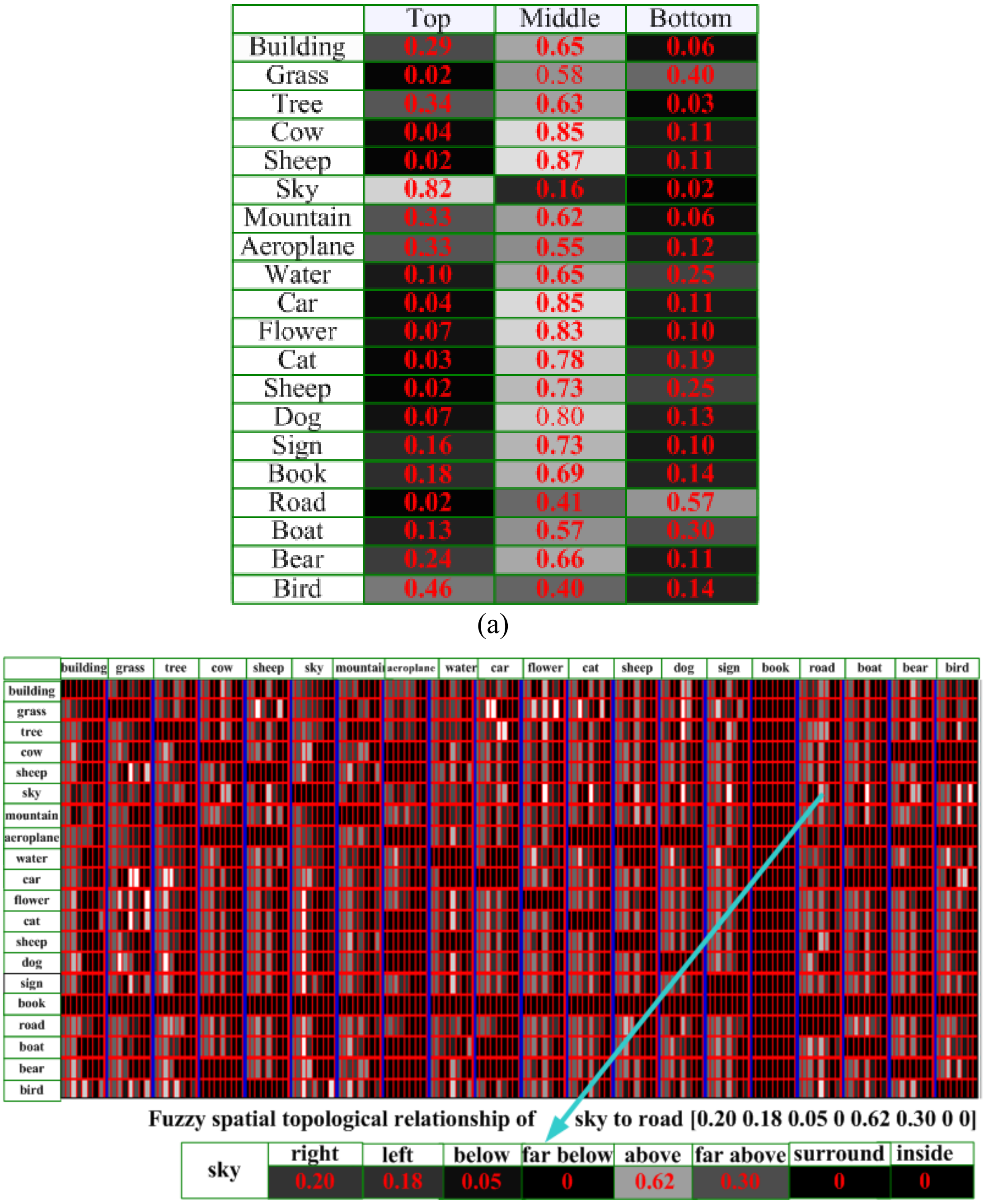
4.4.4 Label propagation intra images

By utilizing fuzzy logic reasoning, labels are propagated intra images in semantic space based on the similarity comparison of contextual cueing features between the patch and concept definition from the common knowledge.

In order to obtain the images from the knowledge base for learning the contextual cueing spatial invariants, we firstly query in Google Image by the given image-level annotations and the pairs of annotations in the MSRC and COREL databases. Then we calculate the fuzzy position of different objects and the fuzzy spatial topological relationship between different objects. The fuzzy membership for the common knowledge is calculated in the same way as the fuzzy membership for patches defined in 4.4.2. Fuzzy position of concept $R_{\boldsymbol{Position}}^{C}(k) = \{\frac{\mu_{\boldsymbol{top}}^{C}(k)}{top}, \frac{\mu_{\boldsymbol{middle}}^{C}(k)}{middle}, \frac{\mu_{\boldsymbol{bottom}}^{C}(k)}{bottom}\}$ is constructed to represent the common position of concept k. Fuzzy spatial topological relationship of concept $R_{\boldsymbol{Spatial}}^{CC}(k_1, k_2) \in F(C \times C)$ is constructed to represent the common spatial relationship between concept k_1 and k_2. $R_{\boldsymbol{Spatial}}^{CC}(k_1, k_2)$ is given as follow:

$$
\begin{aligned}
&R^{CC}_{\boldsymbol{Spatial}}(k_1,k_2)\\
&=\{\frac{\mu^{CC}_{\boldsymbol{right}}(k_1,k_2)}{right},\frac{\mu^{CC}_{\boldsymbol{left}}(k_1,k_2)}{left},\frac{\mu^{CC}_{\boldsymbol{below}}(k_1,k_2)}{below},\frac{\mu^{CC}_{\boldsymbol{far\ below}}(k_1,k_2)}{far\ below},\\
&\frac{\mu^{CC}_{\boldsymbol{above}}(k_1,k_2)}{above},\frac{\mu^{CC}_{\boldsymbol{far\ above}}(k_1,k_2)}{far\ above},\frac{\mu^{CC}_{\boldsymbol{surround}}(k_1,k_2)}{surround},\frac{\mu^{CC}_{\boldsymbol{inside}}(k_1,k_2)}{inside}\}
\end{aligned}
\tag{4.6}
$$

The statistical results of the fuzzy membership are shown in Figure 4.7(a) and Figure 4.7(b) based on the manual annotation of objects in 5000 Google images. Higher the variable is, whiter the color is demonstrated.



	Top	Middle	Bottom
Building	0.29	0.65	0.06
Grass	0.02	0.58	0.40
Tree	0.34	0.63	0.03
Cow	0.04	0.85	0.11
Sheep	0.02	0.87	0.11
Sky	0.82	0.16	0.02
Mountain	0.33	0.62	0.06
Aeroplane	0.33	0.55	0.12
Water	0.10	0.65	0.25
Car	0.04	0.85	0.11
Flower	0.07	0.83	0.10
Cat	0.03	0.78	0.19
Sheep	0.02	0.73	0.25
Dog	0.07	0.80	0.13
Sign	0.16	0.73	0.10
Book	0.18	0.69	0.14
Road	0.02	0.41	0.57
Boat	0.13	0.57	0.30
Bear	0.24	0.66	0.11
Bird	0.46	0.40	0.14

(a)

Fuzzy spatial topological relationship of sky to road [0.20 0.18 0.05 0 0.62 0.30 0 0]

	right	left	below	far below	above	far above	surround	inside
sky	0.20	0.18	0.05	0	0.62	0.30	0	0

(b)

Figure 4.7. Fuzzy membership for the position and the spatial relationship on Google images. (a) Fuzzy membership for position “top,” “middle,” “bottom.” (b) Fuzzy membership for eight spatial relationships. Fuzzy memberships are transformed into 0 to 255 and higher the fuzzy membership is, whiter the color is shown.

With the aid of the knowledge of the position and spatial topological relationships of every concept, we use fuzzy logic reasoning to compare the similarity of those two contextual cueing relationships between $R_{Position}^{P}$, $R_{Spatial}^{PP}$,

$R^{C}_{Position}$, $R^{CC}_{Spatial}$. The similarity measurement of the position to label k_1 in the patch x_{i,j_1} is calculated by Eq. (4.7). Here, "$\circ$" is the fuzzy inner product operator.

$$\begin{aligned}&S^{PC}_{Position}(j_1,k_1)\\&=[R^{P}_{Position}(j_1)\circ R^{C}_{Position}(k_1)]/\vee[R^{P}_{Position}(j_1)\vee R^{P}_{Position}(k_1)]\\&=\vee[R^{P}_{Position}(j_1)\wedge R^{C}_{Position}(k_1)]/\vee[R^{P}_{Position}(j_1)\vee R^{P}_{Position}(k_1)]\end{aligned}\tag{4.7}$$

The similarity measurement of the spatial topological relationships of the patch x_{i,j_1} to label k_1 with patch x_{i,j_2} to the label k_2 is calculated by Eq. (4.8).

$$\begin{aligned}&S^{PC}_{Spatial}(j_1,j_2,k_1,k_2)\\&=R^{PP}_{Spatial}(j_1,j_2)\circ R^{PP}_{Spatial}(k_1,k_2)/\vee[R^{PP}_{Spatial}(j_1,j_2)\vee R^{CC}_{Spatial}(k_1,k_2)]\\&=\vee[R^{PP}_{Spatial}(j_1,j_2)\wedge R^{PP}_{Spatial}(j_1,j_2)]/\vee[R^{PP}_{Spatial}(j_1,j_2)\vee R^{CC}_{Spatial}(k_1,k_2)]\end{aligned}\tag{4.8}$$

Based on the similarity measurement of the position $S^{PC}_{Position}(j_1,k_1)$ and the spatial topological relationships $S^{PC}_{Spatial}(j_1,k_1)$, the fuzzy membership w_{i_1,j_1,k_1} of patch x_{i,j_1} to label k_1 is updated as follow:

$$\begin{aligned}&w^{n}_{i_1,j_1,k_1}=\lambda\times w^{n-1}_{i_1,j_1,k_1}\times S^{PC}_{Position}(j_1,k_1)+\\&(1-\lambda)\times w^{n-1}_{i_1,j_1,k_1}\times\frac{1}{n_{i_1}-1}\times\sum_{j_2=1,j_2\neq j_1}^{n_{i_1}}\{\bigwedge_{k_2=1,k_2\neq k_1}^{n_k}[w^{n-1}_{i_1,j_2,k_2}\times S^{PC}_{Spatial}(j_1,j_2,k_1,k_2)]\}\end{aligned}\tag{4.9}$$

Algorithm 4.1 summarizes the label propagation intra images procedure.

```
Input:  Initial membership value $w^{0}_{i_1,j_1,k_1}$;
Fuzzy membership variable $R^{P}_{Position}$, $R^{PP}_{Spatial}$, $R^{C}_{Position}$, $R^{CC}_{Spatial}$;
The tolerance factor $\varepsilon_{\min}$; Weight parameter $\lambda$.
Output: Fuzzy label membership vector $w^{out}_{i_1,j_1,k_1}$

1: while  The difference of the fuzzy membership with current iteration
and last adjacent iteration $\sum_{j_1=1}^{n_i}\sum_{k_1=1}^{N_c}\| w^{n}_{i_1,j_1,k_1} - w^{n-1}_{i_1,j_1,k_1} \|_F$ is larger than the
threshold $N_c \times \mathrm{n}_i \times \varepsilon_{\min}$
       for $j_1$=1,…, n$_i$;
       for $k_1$ = 1,…, $N_c$;
       for $j_2$ = 1,…, n$_i$;
       for $k_2$=1,…, $N_c$
       do
2:     Calculate the similarity of fuzzy position by Eq. (4.7)
3:     Calculate the similarity of spatial topological relationship by Eq. (4.8)
       end for $k_2$;
       end for $j_2$;
4:     Update the fuzzy label membership of every patch $w^{n}_{i_1,j_1,k_1}$ by Eq. (4.9)
       end for $k_1$
5:     Normalize the fuzzy label membership $w^{n}_{i_1,j_1,k_1}$, $w^{n}_{i_1,j_1,k_1} = w^{n}_{i_1,j_1,k_1} / \sum_{k_1=1}^{n_k} w^{n}_{i_1,j_1,k_1}$
       end for $j_1$;
       end while
6:     Assign the membership of current iteration to the output $w^{out}_{i_1,j_1,k_1} = w^{n}_{i_1,j_1,k_1}$
```

Algorithm 4.1: Label Propagation Intra Image

4.4.5 Post processing based on KNN-FCM clustering

In the post processing stage, we use KNN-FCM (Zahid et al., 2001) to segment the images into semantic regions and associate them with corresponding image-level annotations. Firstly, the initial cluster centers are computed by KNN. Then FCM algorithms are utilized to generate cluster $F_{k'}$, $k' = 1,\ldots K_i$ with its center $c_{i,k'}$, where K_i is the number of annotations in image x_i. In the end, those patches within a same cluster are merged together to form a semantic region. The final region-level label is set as the one with the largest value in the label vector just as follow:

$$l_{i,j_1,k_1} = \begin{cases} 1 \text{ if } \max_{1 \le k' \le m_i} (c_{i,k'}) = c_{i,k_1}, \mathrm{x}_{i,j} \in F_{k'} \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{4.10}$$

4.5 Performance Evaluation

The evaluation of the label to region assignment is based on the region-level ground truth of each image. To demonstrate the performance of our proposed technique, we report the performance comparisons on two public datasets, COREL Stock Photo CDS and MSRC. Both of these datasets have been utilized to evaluate the performance of LRA task in (Liu et al., 2009). The quantitative label-to-region assignment accuracy measures as the percentage of pixels with agreement between the assigned label and ground truth.

Two kinds of techniques are compared here with FCLP. One is a series of Binary support vector machines (BSVM) based algorithms with different maximal patch size, namely, SVM1: 150 pixels, SVM2: 200 pixels, SVM3: 400 pixels, and SVM4: 600 pixels. BSVM is implemented based on the lib-SVM library and the Gaussian Radial Basis Function kernel is used here by setting the kernel parameter to be 1. The others are two latest LRA techniques of label propagation with one-layer sparse coding and Bi-layer sparse coding (Liu et al., 2009). For the proposed techniques, the parameter λ, $T_{\max}$ and $\varepsilon_{\min}$, actually shows stable performance under different values. In our experiments, we set $\lambda = 0.7$, $\varepsilon_{\min} = 0.1$, $N_{\max} = 50$, $N_s = 5$, and the dimension of the BOW feature vector *m* = 628, including 500 dimensions and 128 dimensions for SIFT and LAB color descriptors respectively.

4.5.1 Experiment on COREL

COREL dataset is one of the popular dataset in the community of image retrieval and image recognition. Following the collection strategy in (Yuan et al., 2007) (Liu et al., 2009), we randomly select 150 images from COREL-1000 and manually annotate the ground-truth, which contains 8 categories: grass, cow, mountain, sky, bear, water, tree, and building. Table 4.1 shows the accuracy comparison of SVM-based algorithms, one-layer and Bi-layer LRA, and proposed technique FCLP. The detailed comparisons of individual objects are illustrated in Figure 4.8. Obviously, FCLP achieves the best performance under all the cases.

Table 4.1. LRA accuracy comparisons on COREL.

Method	Accuracy
SVM1	0.29
SVM2	0.32
SVM3	0.34
SVM4	0.33
One Layer	0.51
Bi Layer	0.62
FCLP	**0.70**

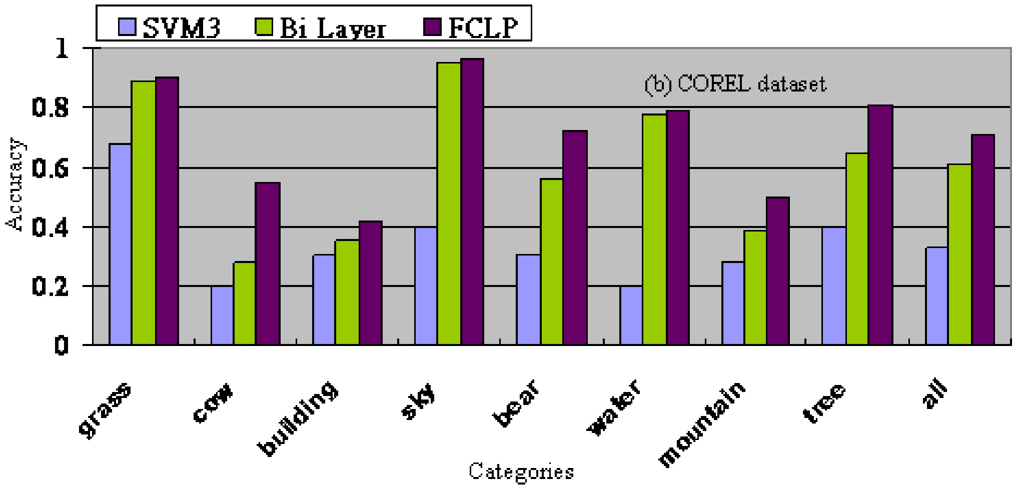


Figure 4.8. LRA accuracies on COREL dataset. The horizontal axis shows the name of each label/annotation and the vertical axis provides the accuracy.

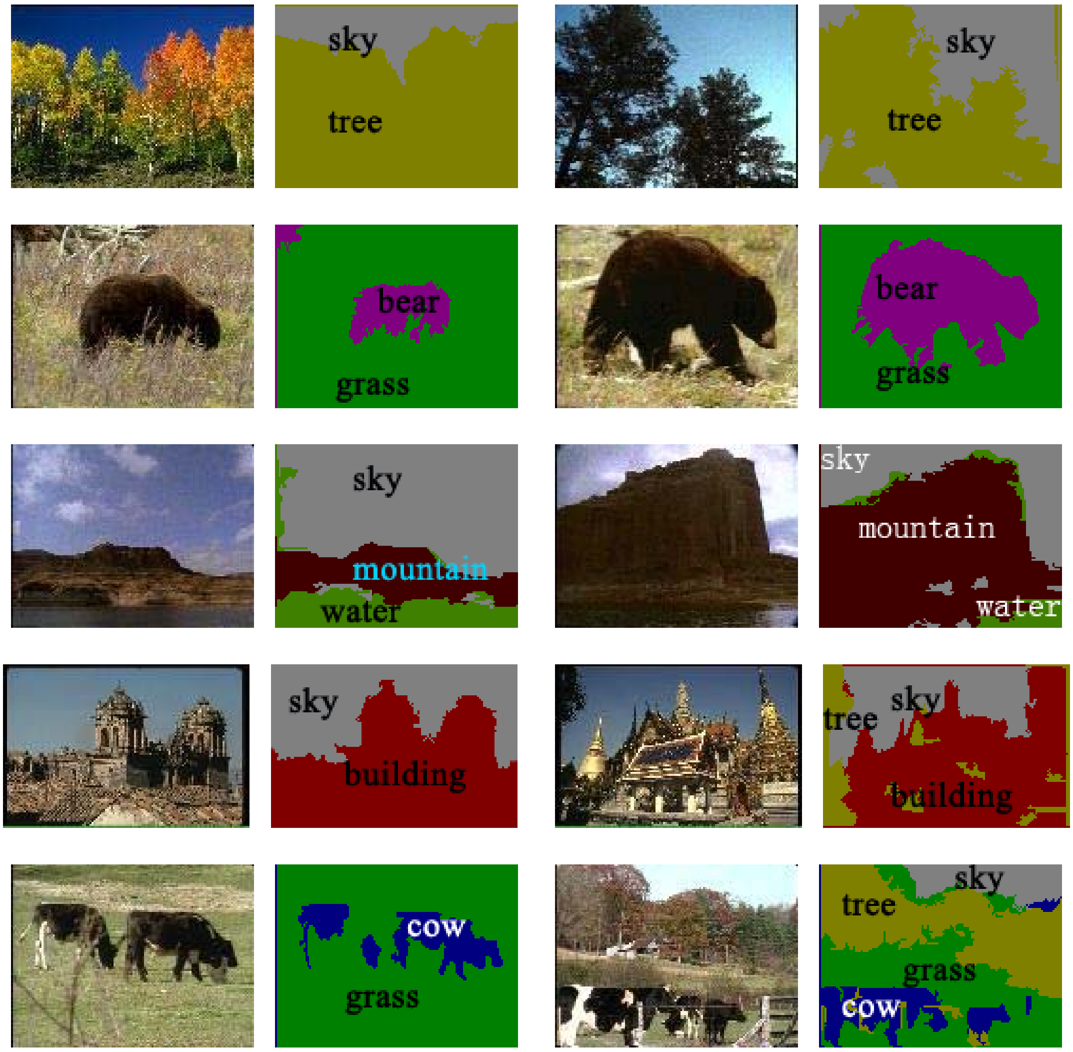


Figure 4.9. Examples of LRA results on COREL dataset. Each color is denoted as one class of localized region.

Figure 4.9 provides some examples of results on COREL dataset, covering all eight categories of regions. Compared with previous work, our proposed technique shows much higher accuracy on the objects with relatively explicit position information or spatial topological relationship with other objects, even their appearance is similar, such as sky, airplane, road, car, and boat. Furthermore, all existing techniques suffer from the performance decreasing if more objects appear in the image. But FCLP may benefit from it because more objects provide more contextual cueing information to help us understand the whole image.

4.5.2 Experiment on MSRC

MSRC dataset contains 591 images from 23 categories with image-level annotations and the ground-truth of region-level annotations. There are about 3 labels on average for each image. Similar with previous work on this dataset (Liu et al., 2009), we remove the images with only single annotation or infrequent annotation. This gives rise to 380 images with totally 18 categories: building, grass, tree, cow, boat, sheep, sky, mountain, aeroplane, water, bird, book, road, car, flower, cat, sign, and dog. Table 4.2 shows the accuracy comparisons of a series of SVM-based algorithms, one-layer and Bi-layer LRA, and the proposed FCLP. FCLP performs much better than all other algorithms. The detailed comparison results are illustrated in Figure 4.10.

Figure 4.11 demonstrates some samples of proposed LRA results on MSRC dataset, from simple images with only one or two objects to complex images with four or five different objects. Compared with images in COREL dataset, the objects are more varied in MSRC dataset. Obviously, our technique shows effectiveness in images with multiple objects even some of their appearances are similar, such as grass and tree, building and airplane.

Moreover, we demonstrate some interesting observations of the performance comparison in Figure 4.12. Figure 4.12 (a) is the image with given image-level annotations of sky, building, tree and road. Since the regions of sky and road have very similar SIFT features, Bi-layer technique assigns the road annotation to the sky region. Moreover, this error influences the further region segmentation and assignment as shown in Figure 4.12 (b). With the aid of fuzzy position and fuzzy spatial topological memberships shown in Figure 4.7, FCLP technique recognizes the sky region correctly in Figure 4.12(c). Another example is shown in Figure 4.12 (d), which contains sky, road, tree and car. Bi-layer technique assigns the road annotation to the sky region again. Furthermore, Bi-layer technique is less effective for handling the categories for foreground objects (Liu et al., 2009). For example, the car was not segmented and recognized correctly because of the small size in Figure 4.12 (e). FCLP also considers the sizes of the objects, but as shown in Figure 4.12 (f), the car is correctly assigned. The key is how to use contextual cueing appropriately in image content analysis. As mentioned previously in Section 4.3, five types of spatial invariants, such as size and position, are thought to be important in contextual cueing, so if only one or two spatial invariants are utilized, the image analysis result may over-emphasize certain aspect of contextual cueing. In this case, fuzzy theory successfully demonstrates the effectiveness in modeling human's understandings of the visual world in label to region assignment task.

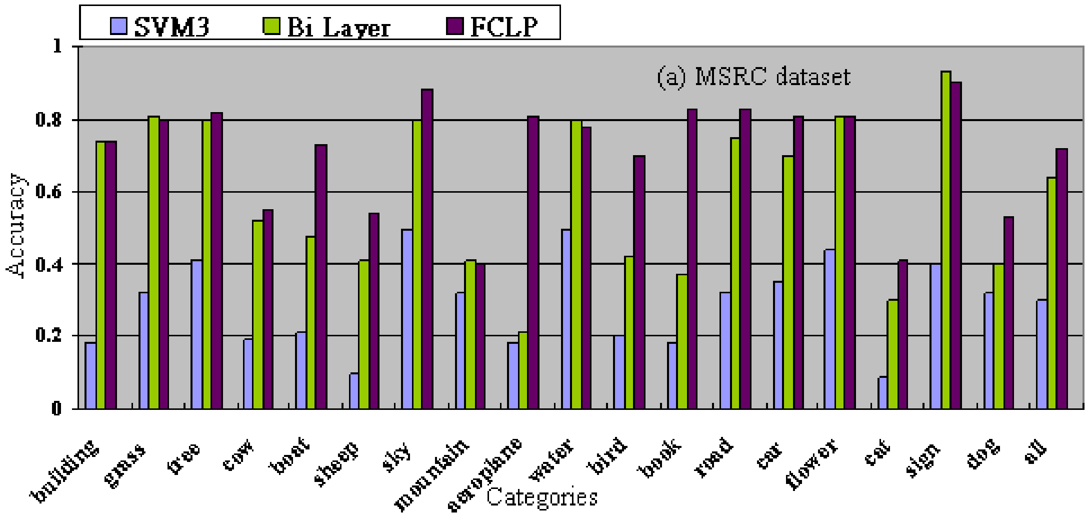


Figure 4.10. Detailed LRA accuracies for MSRC dataset. The horizontal axis shows the name of each label/annotation and the vertical axis represents the accuracy.

Table 4.2. LRA accuracy comparisons on MSRC.

Method	Accuracy
SVM1	0.24
SVM2	0.22
SVM3	0.27
SVM4	0.25
One Layer	0.54
Bi Layer	0.65
FCLP	**0.72**

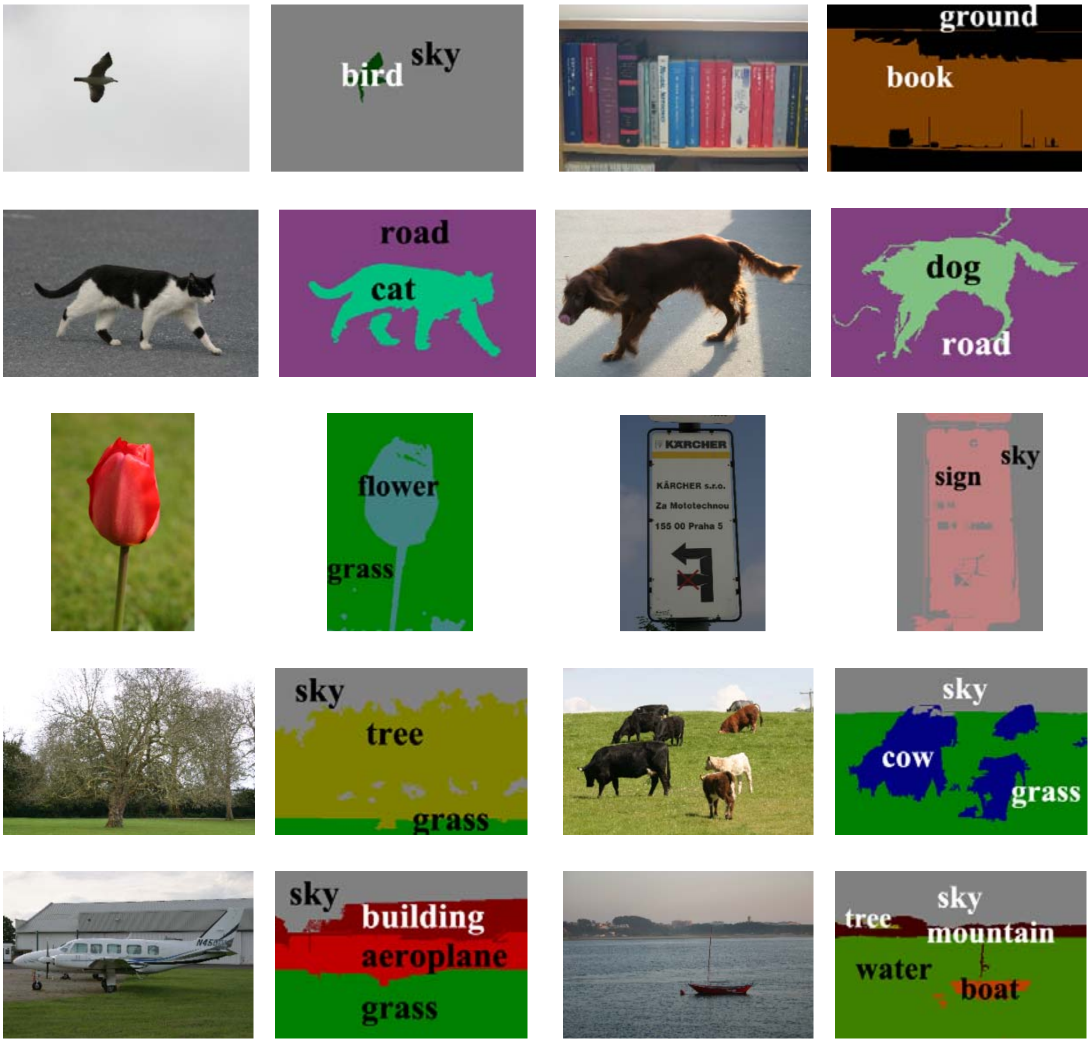


Figure 4.11. Examples of LRA results on MSRC dataset. Each color is denoted as one class of localized region.

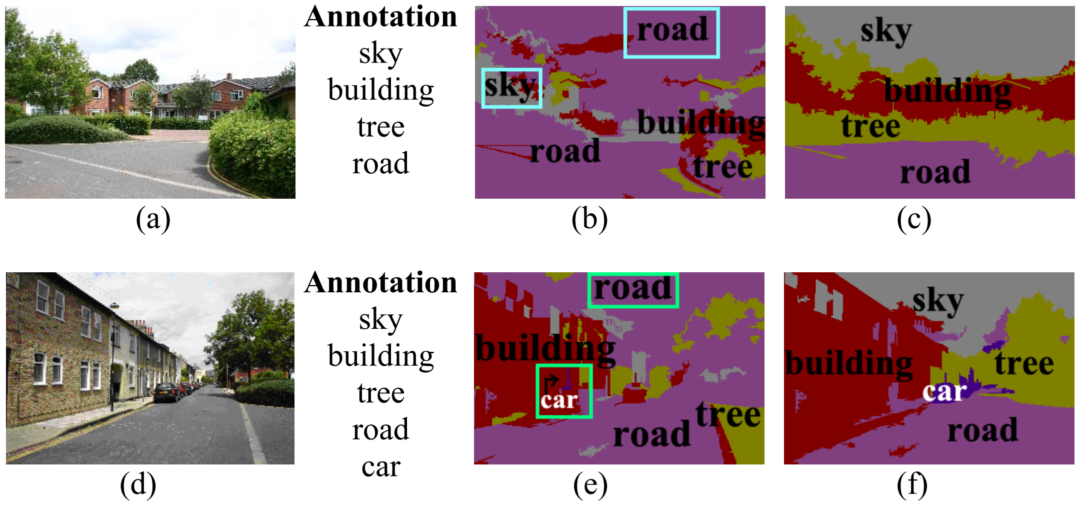


Figure 4.12. Comparisons of LRA results. (a) An image with the annotations of sky, building, tree, road (b) Bi-layer result (c) FCLP result (d) An image with annotations of sky, building, tree, car, and road. (e) Bi-layer result (f) FCLP result.

4.6 Summary

Aims to the ultimate problem of image understanding, region level annotation is an important task in multimedia content analysis which makes the commercial multimedia search engine possible. This chapter proposes a novel FCLP technique for LRA problem to natural images. In our proposed technique, we integrate contextual cueing, which demonstrates impressive performance when objects have similar visual appearances, can effectively improve the semantic understanding of the images. Fuzzy theory is utilized to describe the contextual cueing knowledge to fill the semantic gap between human judgment and computable features using mathematical model. It demonstrates especially for the position information and the topological relationship. Moreover, FCLP inherits the merits of label propagation methods, which reduces the training cost by taking advantage of the similarity among the data with common labels. The experiments on two public datasets demonstrate that the proposed technique achieves obvious performance improvement of LRA for the images with multiple objects and complex background.

Chapter 5 Visual Cortex Simulation for Image Classification

5.1 Overview

Image classification is a well-known classical problem in multimedia content analysis. This chapter proposes a novel deep learning model called Bilinear deep belief networks (BDBN) for image classification (Zhong et al., 2011c) (Zhong et al., 2012b). Unlike previous image classification models, BDBN aims to provide human-like judgment by referencing the architecture of the human visual system and the procedure of intelligent perception. Therefore, the multi-layer structure of the cortex and the propagation of information in the visual areas of the brain are realized faithfully. Unlike most existing deep models, BDBN utilizes a bilinear discriminant strategy to simulate the "initial guess" in human object recognition, and at the same time to avoid falling into a bad local optimum. To preserve the natural tensor structure of the image data, a novel deep architecture with greedy layer-wise reconstruction and global fine-tuning is proposed. To adapt real-world image classification tasks, we develop BDBN under a semi-supervised learning framework, which makes the deep model work well when labeled images are insufficient.

Comparative experiments on four standard classification datasets show that the proposed algorithm outperforms both representative classification models and existing deep learning techniques. Furthermore, in face image dataset, our model is able to automatically abstract and emphasize the important facial features and patterns which are consistent with the human's attention map. The success of this approach suggests a plausible proof for a class of neurobiological models for different multimedia tasks.

5.2 Introduction

A long-time goal for multimedia and computer vision has been to build system that achieves human-level classification and recognition performance (Serre et al., 2007). Image classification, which aims to understand the semantic meaning of visual information and determine the category of the images according to some predefined criteria, has been extensively studied for more than fifteen years (Moosmann et al., 2008). Existing image classification methods can be roughly divided into two broad families of approaches: parametric and nonparametric classifiers. Parametric classifiers, also known as learning-based classifiers, require an intensive training phase of the classifier parameters, e.g., the parameters of SVM (Kumar & Sminchisescu, 2007), Boosting (Opelt et al., 2004), fragments and object parts (Yang et al., 2009), decision trees (Bosch et al., 2007), web graphs (Mahajan & Slaney, 2010), hierarchical classification models (Tsai et al., 2010). To date, the leading image classifiers are parametric classifiers, particularly SVM-based methods. Nonparametric classifiers make their classification decisions directly on the data, and require no training of parameters (Boiman et al., 2008). Recently, in the literature on multimedia, many papers focused on the specific applications; for instance, landmark image classification (Xian et al., 2010), sports genre & view type classification (Li et al., 2009), age images classification (Chu et al., 2010) and affective images classification (Machajdik & Hanbury, 2010) (Valenti et al., 2010).

But today, even for the best artificial classification systems, image classification for real-world applications remains a well-known challenge. One interesting observation is that humans, even children, do not have any difficulty in classifying images. Before the age of 25 months, children have already developed the ability to recognize novel three-dimensional objects fast (Wallis & Bülthoff, 1999). Driven by this observation, researchers in the fields of cognitive science and neuroscience have conducted pioneering work on modeling the human visual cortex using computational models for multimedia content analysis tasks.

Pioneering attempts of using the visual neuroscience in computer vision have been limited to early vision for deriving stereo algorithms, e.g. (Marr & Poggio, 1979), or justifying the use of DoG (derivative-of-Gaussian) filters and Gabor filters (Jones, & Palmer, 1987). Current work of computational neuroscience models resembled the simple S units and complex C units of visual cortex with hierarchical architecture (Fukushima, 1980) (LeCun et al., 1998) (Riesenhuber & Poggio, 1999), (Serre et al., 2007). In recent year, deep learning, which models the learning tasks using the architecture composed of multiple layers of parameterized nonlinear modules, has attracted more and more attention because of its impressive performance in various visual data analysis tasks. Although these visual cortex-like computational models achieved notable success in modeling the human visual system, the limitations of them are also obvious. Most of them only utilize some individual characters of human visual cortex, hence lack of systematic construction and development of machine visual system.

This chapter proposes a novel image classification framework by imitating human's visual cortex systematically. According to the findings in neuroscience, we first summarize the characters of human cortex as follows:

1) The neocortex has a complex multi-layer hierarchy (Lee & 2003). The laminar structure and a multi-layer illustration of the neocortex. The neocortex can be roughly divided into six functionally distinct layers from Molecular layer I to Multiform layer VI. Layer IV in the primary visual cortex (V1) is further divided into four layers, labeled 4A, 4B, 4Cα, and 4Cβ. Therefore, dozens of cortical layers are involved in generating even the simplest vision (Leuba & Kraftsik, 1994).

2) Bi-directional information propagates in human visual cortex. Our visual systems contain multi-layer generative models in which top-down connections (feedback connectons) can be used to generate low-level features of images from high-level representations, and bottom-up connections (feedforward connections) can be used to infer the high-level representations that would have generated an observed set of low-level features (Hinton, 2007). Single cell recordings and the reciprocal connectivity between cortical areas (Felleman & Van Essen, 1991) both suggest a hierarchy of progressively more complex features in which each layer can influence the layers below it.

3) In the primary visual cortex, all the way through the optic tract to a nerve position is a direct correspondence from an angular position in the field of view of the eye, just like a matrix.

4) There exist two peaks of activation in the visual cortex areas. With regard to object recognition, the early peak is related to the activation of an "initial guess" based on the discriminative knowledge that has been acquired, while the late peak reflects the post-recognition activation of conceptual knowledge related to the recognized object (VanRullen & Thorpe, 2001).

5) The complexity and the invariance of object representation in images increases along with the layer increases. The invariance includes the position, scale,

viewpoint, illumination conditions and the noises transformations (Serre et.al, 2007) (Fu et al., 2012).

6) The response function of the simple cells in V1 is similar with Gabor functions (Gabor, 1946). The input image is first analyzed by a multidimensional array of simple S1 units which correspond to the classical simple cells of Hubel and Wiesel found in the primary visual cortex (V1) (Hubel & Wiesel, 1962). And S1 units take the form of Gabor functions.

7) Visual cortex mainly consists of two pathways: ventral stream and dorsal stream, the former is involved in the identification of objects, while the latter is linked to the localization of objects (Serre, 2006).

To simulate these characters of human visual cortex, we propose a novel bilinear deep belief network (BDBN) for image classification. BDBN utilizes deep architecture composed of multiple layers of parameterized nonlinear modules, to imitate the laminar structure of neocortex. To simulate the bi-directional information propagation in visual cortex, BDBN uses the greedy layer-wise reconstruction proposed by deep belief networks (DBN), the most representative deep learning model. However, most DBN based techniques unfold the image to vector before inputting to the deep architecture, which destroys the natural second order tensor structure of the image. Moreover, it is not consistent with human's visual perception. To satisfy the third characters of human visual cortex, BDBN proposes a novel 2D deep architecture. The input layer and all hidden layers in BDBN are constructed by a set of second-order planes, which are fully connected with the adjacent ones until the output layer, a vector to indicate the label of the images. Based on this new deep architecture, we propose a novel deep learning algorithm with three stages: bilinear discriminant initialization, greedy layer-wise reconstruction, and global fine-turning.

The rationale for three-stage learning comes from the phenomenon of two peaks of activation described as the fourth character of human visual cortex. In most existing deep models, “post activation” is modeled by the fine-tuning stage, but the “initial guess” process is neglected. In our model, two peaks of activation and the propagation of information in the visual cortex are faithfully realized. We model the peak activation of the “initial guess” by preserving the disciminant information of the labeled data to the greatest extent. Most existing deep models initialize the parameter space in a random manner and gradually approximate a locally optimal solution by learning. Unfortunately, a bad initial parameter space may lead to a poor local optimum and thus seriously affect the following learning procedure. To address this problem, we utilize a bilinear discriminant strategy to construct a second-order plane from the lower layer. The symmetrically weighted connections between these two adjacent layers are used as the initial parameter space for further learning. Moreover, the discriminant-based “initial guess” brings an additional advantage to the meaningful architecture. Currently, the number of neurons in each layer is fixed and pre-defined intuitively. In our model, the size of the deep architecture is determined based on the optimum dimension for retaining the discriminant information.

Although BDBN doesn't provide special design to address the fifth and the sixth characters of human visual cortex, the encouraging fact is that the experimental results of real image classification tasks have demonstrated that BDBN has shown good consistency to these characters. For the fifth character, after the model is learnt based on the training data, the weights of the first layer in BDBN are oriented, Gabor-like and resemble the receptive fields of V1 simple cells. To the sixth character, the BDBN has improvement in noise invariance with increased layers. For the seventh characters of two pathway stream, it is not directly relevant to image classification task, so we will not address it in this chapter. In our future work of label to region assignment by deep learning model, this character plays an important role.

Last but not least, we develop our deep model under a semi-supervised learning framework because of the insufficiency of the labeled images in real-world applications. Relying on the efforts of experienced human annotators, labeled instances are often difficult, expensive, or time consuming to obtain (Zhu, 2006). By contrast, with the growing availability of a large number of images from photo-sharing sites such as Flickr, abundant unlabeled data are available (Gross et al. 2008). Moreover, semi-supervised learning framework is also consistent with human's daily learning experience.

The remainder of this chapter is organized as follows. We first introduce the related work on deep learning in Section 5.3. A novel deep architecture and a new deep learning algorithm are introduced in Section 5.4. Section 5.5 shows the performance of the proposed techniques in real image classification tasks and Section 5.6 concludes this chapter.

5.3 Related Work on Deep Learning

Different from shallow learning models, deep learning is about learning multiple levels of representation and abstraction that helps to make sense of data. Besides evidence from neuroscience, some theoretical analyses from machine learning also provide support for the argument that deep models are more compact and expressive than shallow models in representing most learning functions, especially highly variable ones. Many empirical validations support the argument that deep architectures have shown promise performance in solving hard learning problems (Larochelle, et al., 2007). Theoretical analysis also indicates that compared with shallow circuits, such as a typical Support vector machines (SVM), deep architectures are more efficient because they can represent most common functions, especially highly-variable learning functions compactly and effectively.

Unfortunately, it is difficult to learn the parameters of deep architectures with multiple hidden layers containing trainable weights at all levels. Back propagation, a well-known computationally efficient model for multilayer neural networks, also suffers from the problems of insufficient labeled data, high computational cost, and poor local optima when working under a deep model (Hinton, 2007). To reduce the difficulty of deep learning, Hinton and Salakhutdinov propose a densely-connected, directed belief nets with multiple hidden layers, called Deep belief networks (DBN), which partitions the learning procedure to two stages: abstract input information layer by layer and fine-tune the whole deep network to the ultimate learning target (Hinton, et al., 2006) (Salakhutdinov, et al., 2007). DBN pairs each feed-forward layer with a feed-back layer that attempts to reconstruct the input of the layer from the output. Such layer-wise generative models are implemented by a family of Restricted boltzmann machines (RBMs) (Smolensky, 1986). After a greedy unsupervised learning to each pair of layers, the lower-level features are progressively combined into more compact high-level representations. In the second stage, the whole deep network is refined using a contrastive version of the "wake-sleep" algorithm via a global gradient-based optimization strategy. Owing to this two-stage fast greedy learning, DBN exhibits notable performance in image retrieval (Hörster & Lienhart 2008), image annotation (Wang et al., 2010), audio event classification (Ballan, et al. 2009).

In recent years, deep convolutional architectures have been attracting an increasing amount of attention because of their ability to preserve the space structure and resistance to small variations in the data (Lee et al., 2009) (Taylor et al., 2010). As early as in 1989, LeCun et al. proposed deep convolutional networks that used a feature detection layer followed by a feature pooling layer as the basic module, and that was trained to minimize the overall loss for classification (LeCun et al., 1989). While the convolutional nets are deep, i.e., including a series of multiple detection/pooling modules, they do not seem to suffer from the convergence problems that plague deep fully-connected neural nets (Bengio & LeCun, 2007). Similar with DBN, Deep convolutional networks (DCNN) has no distinct feature extractor and classifier. All of the layers in DCNN are trained from data in an integrated fashion. Currently, DCNN has been successfully used to extract spatial features (Memisevic & Hinton, 2010) and spatial-temporal features (Taylor et al., 2010) in different applications, such as image classification (Jarrett et al., 2009) (Mahajan & Slaney, 2010) and human action recognition (Xu et al., 2013).

5.4 Bilinear Deep Belief Networks

In this section, we propose a novel learning framework based on Bilinear deep belief networks (BDBN). The learning procedure of our Bilinear deep belief networks, is demonstrated in Section 5.4.1. The bilinear discriminant initialization stage is discussed in Section 5.4.2. Section 5.4.3 contains details of the greedy layer-wise reconstruction. The global fine-tuning process of the whole deep network is described in Section 5.4.4. The algorithm of BDBN and an attention modeling method based on BDBN is given in Section 5.4.5.

5.4.1 Learning procedure of BDBN

Let X be a set of data samples as shown below:

$$X = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k, \ldots, \mathbf{X}_K] \tag{5.1}$$

where $\mathbf{X}_k$ is a sample datum in the image space $\mathbb{R}^{I \times J}$ and K is the number of sample data. Let Y be a set of labels corresponding to X, which can be seen as:

$$Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k, \ldots, \mathbf{y}_K] \tag{5.2}$$

And $\mathbf{y}_k$ is the label vector of $\mathbf{X}_k$ in $\mathbb{R}^C$, where C is the number of classes.

$$y_k^c = \begin{cases} 1 & \text{if } \mathbf{X}_k \in c\text{th class} \\ 0 & \text{if } \mathbf{X}_k \notin c\text{th class} \end{cases} \tag{5.3}$$

Based on the given training set, the aim in image classification is to learn a mapping function from the image set *X* to the label set *Y*, and then classify the new coming data points according to the learned mapping function.

To address the problem of image classification, we propose a novel bilinear deep learning technique BDBN. Figure 5.1 shows the architecture of BDBN. A fully interconnected directed belief network includes input layer H^1, hidden layer H^2,…, H^N, and one label layer La at the top. The input layer H^1 has $I \times J$ units, and this size is equal to the dimension of the input features. In our model, we use the pixel values of sample datum $\mathbf{X}_k$ as the original input features. In the top, the label layer has C units, which is equal to the number of classes. The search of the mapping function from *X* to *Y* is transformed to the problem of finding the optimum parameter space θ^* for the deep architecture.

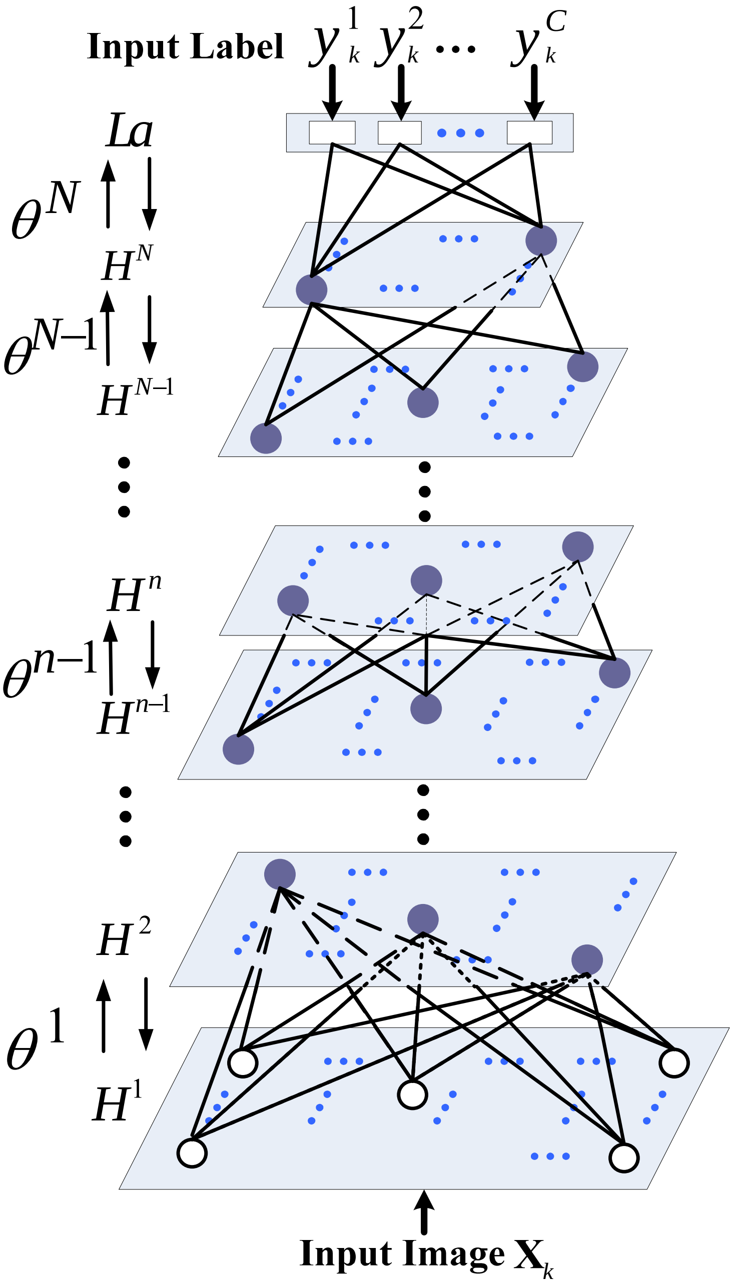


Figure 5.1. Architecture of the Bilinear deep belief networks.

The learning procedure of our proposed BDBN is listed below:

1) The strategy of Bilinear discriminant projection is utilized to construct a projection to map the original data into a discriminant bilinear subspace.

2) The initial symmetrically weighted connections are constructed between adjacent layers according to the "initial guess" based on the discriminant

information. The size of the deep architecture is determined automatically based on the optimum dimension to retain the discriminant information.

3) After the architecture of the next layer is determined, the parameter space is refined by the greedy layer-wise information reconstruction using RBMs as building blocks.
4) Repeat the first to third stages until the parameter space θ in all N layers is constructed.
5) In the "post activation" stage, the whole deep model is fine-tuned to minimize the classification error based on backpropagation.

5.4.2 Bilinear discriminant initialization

In this subsection, we introduce the Bilinear discriminant projection (BDP), which is used to extract the discriminant information from the original image datasets.

Given the labeled training data points $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_L \in \mathbb{R}^{I \times J}$, without unfolding the input data to vectors, BDP aims to find two projection matrices $\mathbf{U} \in \mathbb{R}^{I \times P}$ and $\mathbf{V} \in \mathbb{R}^{J \times Q}$ such that the latent representation $\mathbf{TX}_1, \mathbf{TX}_2, ..., \mathbf{TX}_L \in \mathbb{R}^{P \times Q}$ can be obtained by $\mathbf{TX}_s = \mathbf{U}^T \mathbf{X}_s \mathbf{V}$ (s = 1, ..., L), just as depicted in Figure 5.2.

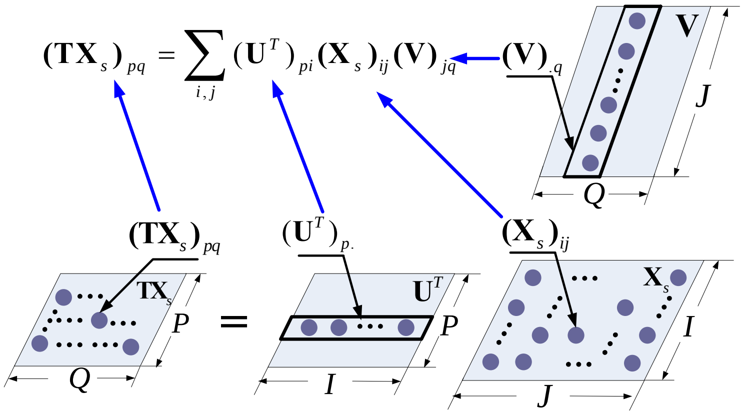


Figure 5.2. Latent representation with projection matrices $\mathbf{U}$ and $\mathbf{V}$.

In order to preserve the discriminant information in the learning procedure, the objective function of BDP could be represented as follows:

$$\begin{aligned} &\arg\max_{\mathbf{U},\mathbf{V}} \; J(\mathbf{U},\mathbf{V}) = \sum_{s,t=1}^{L} \| \mathbf{U}^T(\mathbf{X}_s - \mathbf{X}_t)\mathbf{V} \|^2 \, (\alpha \mathbf{B}_{st} - (1-\alpha)\mathbf{W}_{st}) \\ &s.t. \; \mathbf{U}^T\mathbf{U} = \mathbf{I}_P, \;\; \mathbf{V}^T\mathbf{V} = \mathbf{I}_Q \end{aligned} \tag{5.4}$$

where $\alpha \in [0,1]$ is the parameter used to balance the between-class weights $\mathbf{B}_{st}$ and the within class weights $\mathbf{W}_{st}$, which are defined as follows (Yan et al., 2001)(Sugiyama, 2007):

$$\mathbf{B}_{st} = \begin{cases} \dfrac{1}{n_d} - \dfrac{1}{n_c}, & \text{if } \mathbf{y}_s^c = \mathbf{y}_t^c = 1, \\ \dfrac{1}{n_d}, & \text{else,} \end{cases} , \mathbf{W}_{st} = \begin{cases} \dfrac{1}{n_c}, & \text{if } \mathbf{y}_s^c = \mathbf{y}_t^c = 1, \\ 0, & \text{else,} \end{cases} \tag{5.5}$$

where $\mathbf{y}_s^c$ denotes the class label of datum point $\mathbf{X}_s$, n_d is the number of data points in all classes and n_c is the number of data points in class c, where $c \in \{1, ..., C\}$.

By simultaneously maximizing the distances between data points from different classes and minimizing the distances between data points from the same class, the discriminant information is preserved to the greatest extent in the projected feature space. Optimizing $J(\mathbf{U},\mathbf{V})$ is a non-convex optimization problem with respect to the projection matrices $\mathbf{U}$ and $\mathbf{V}$. However, solving $\mathbf{U}$ (or $\mathbf{V}$) with fixed $\mathbf{V}$ (or $\mathbf{U}$) is a convex optimization problem. Let $\mathbf{E}_{st} = \alpha \mathbf{B}_{st} - (1-\alpha)\mathbf{W}_{st}$, with the fixed **V**. The optimal **U** is composed of the first *P* eigenvectors of the following eigendecomposition problem:

$$\mathbf{D}_{\mathrm{V}}\mathbf{u} = \boldsymbol{\lambda}\mathbf{u} \tag{5.6}$$

where $\mathbf{D}_{\mathrm{V}} = \sum_{st} \mathbf{E}_{st} (\mathbf{X}_s - \mathbf{X}_t) \mathbf{V}\mathbf{V}^T (\mathbf{X}_s - \mathbf{X}_t)^T$. Similarly, with the fixed **U**, the optimal **V** is composed of the first *Q* eigenvectors of the following eigendecomposition problem:

$$\mathbf{D}_{\mathrm{U}}\mathbf{v} = \boldsymbol{\lambda}\mathbf{v} \tag{5.7}$$

where $\mathbf{D}_{\mathrm{U}} = \sum_{st} \mathbf{E}_{st} (\mathbf{X}_s - \mathbf{X}_t)^T \mathbf{U}\mathbf{U}^T (\mathbf{X}_s - \mathbf{X}_t)$.

Therefore, we can alternately optimize **U** (with a fixed **V**) and **V** (with a fixed **U**). The above steps monotonically increase $J(\mathbf{U},\mathbf{V})$ and since the function is upper bounded, it will converge to a critical point with transformation matrices **U**, **V**.

The sizes of *P* and *Q* are determined by the number of positive eigenvalues in $\mathbf{D}_{\mathrm{V}}$ and $\mathbf{D}_{\mathrm{U}}$, respectively, since adding the eigenvectors corresponding to the nonpositive eigenvalues will not increase $J(\mathbf{U},\mathbf{V})$ in Eq. (5.4). As a result, the original dimension $I \times J$ is automatically reduced into $P \times Q$.

5.4.3 Greedy layer-wise reconstruction

The sample data set $\mathbf{X}$ is inputted to the deep architecture as the input layer H^1 to construct an RBM with the first hidden layer H^2.

The energy of the state $(\mathbf{h}^1, \mathbf{h}^2)$ in the first RBM is:

$$\begin{aligned} & E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right) \\ & = -(\mathbf{h}^1\mathbf{A}^1\mathbf{h}^2 + \mathbf{b}^1\mathbf{h}^1 + \mathbf{c}^1\mathbf{h}^2) \\ & = -\sum_{i=1,j=1}^{i\le I,j\le J} \sum_{p=1,q=1}^{p\le P^2,q\le Q^2} h_{ij}^1 A_{ij,pq}^1 h_{pq}^2 - \sum_{i=1,j=1}^{i\le I,j\le J} b_{ij}^1 h_{ij}^1 - \sum_{p=1,q=1}^{p\le P^2,q\le Q^2} c_{pq}^1 h_{pq}^2 \end{aligned} \quad (5.8)$$

where $\theta^1 = \left(\mathbf{A}^1,\mathbf{b}^1,\mathbf{c}^1\right)$ are the model parameters between the input layer H^1 and first hidden layer H^2. $A_{ij,pq}^1$ is the symmetric interaction term between the input unit (i,j) in H^1 and the hidden unit (p,q) in H^2. b_{ij}^1 is the $(i,j)^{\text{th}}$ bias of layer H^1 and c_{pq}^1 is the $(p,q)^{\text{th}}$ bias of layer H^2. $I\times J$ is the number of units in H^1, while $P^2\times Q^2$ is the number of units in H^2. Therefore, the first RBM has the following joint distribution:

$$P\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right) = \frac{1}{Z}e^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)} = \frac{e^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)}}{\sum_{\mathbf{h}^1}\sum_{\mathbf{h}^2} e^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)}} \quad (5.9)$$

where Z is the normalization constant. The probability of the model assigned to $\mathbf{h}^1$ in H^1 is:

$$P\left(\mathbf{h}^1\right) = \frac{1}{Z}\sum_{\mathbf{h}^2} e^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)} = \frac{\sum_{\mathbf{h}^2} e^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)}}{\sum_{\mathbf{h}^1}\sum_{\mathbf{h}^2} e^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)}} \quad (5.10)$$

And the log-likelihood of $P\left(\mathbf{h}^1\right)$ is:

$$\log P\left(\mathbf{h}^1\right) = \log\sum_{\mathbf{h}^2} e^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)} - \log\sum_{\mathbf{h}^1}\sum_{\mathbf{h}^2} e^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)} \quad (5.11)$$

Gibbs sampling from an RBM proceeds by sampling $\mathbf{h}^2$ given $\mathbf{h}^1$, then sampling $\mathbf{h}^1$ given $\mathbf{h}^2$, and so on. The conditional distributions over input state $\mathbf{h}^1$ in layer H^1 and hidden state $\mathbf{h}^2$ in layer H^2 are given by the logistic functions Eq. (5.12) and Eq. (5.13), where $\sigma(x) = \frac{1}{1+\exp(-x)}$.

$$p\left(\mathbf{h}^2 \,|\, \mathbf{h}^1\right)=\prod_{p,q} p\left(h^2_{pq} \,|\, \mathbf{h}^1\right), p\left(h^2_{pq}=1 \,|\, \mathbf{h}^1\right)=\sigma\left(\sum_{i=1,j=1}^{i\le I,j\le J} h^1_{ij} A^1_{ij,pq}+c_{pq}\right) \quad (5.12)$$

$$p\left(\mathbf{h}^1 \,|\, \mathbf{h}^2\right)=\prod_{i,j} p\left(h^1_{ij} \,|\, \mathbf{h}^2\right), p\left(\mathrm{h}^1_{ij}=1 \,|\, \mathbf{h}^2\right)=\sigma\left(\sum_{p=1,q=1}^{p\le P^2,q\le Q^2} A^1_{ij,pq} h^2_{pq}+b_{ij}\right) \quad (5.13)$$

Denote $\mathbf{h}^2(t)$ for the t^{th} of $\mathbf{h}^2$ sample from the chain, starting at $t=0$ with $\mathbf{h}^1\mathbf{(0)}$, which is the input observation for the RBM, and $(\mathbf{h}^2(t),\mathbf{h}^1(t))$ for $t \longrightarrow \infty$ is a sample from the Markov chain. Therefore, we can calculate the derivative of Eq. (5.11) with respect to the parameter $\theta^1=\left(\mathbf{A}^1,\mathbf{b}^1,\mathbf{c}^1\right)$ below:

$$\frac{\partial \log p(\mathbf{h}^1(0))}{\partial \theta^1}=-\sum_{\mathbf{h}^2(0)} p(\mathbf{h}^2(0)\,|\,\mathbf{h}^1(0))\frac{\partial E(\mathbf{h}^2(0),\mathbf{h}^1(0))}{\partial \theta^1}+\sum_{\mathbf{h}^2(t)}\sum_{\mathbf{h}^1(t)} p(\mathbf{h}^2(t),\mathbf{h}^1(t))\frac{\partial E(\mathbf{h}^2(t),\mathbf{h}^1(t))}{\partial \theta^1} \quad (5.14)$$

The idea of the Contrastive divergence algorithm (Hinton, 2002) using the difference between two Kullback-Liebler divergences is to take *t* small (typically $t=1$) to run the chain for only one step. When $t=1$, the derivative to the model parameter $\mathbf{A}^1$ can be obtained by Eq. (5.15),

$$\begin{aligned}\frac{\partial \log P(\mathbf{h}^1(0))}{\partial \mathbf{A}^1}&=-\sum_{\mathbf{h}^2(0)} P(\mathbf{h}^2(0)\,|\,\mathbf{h}^1(0))\frac{\partial E(\mathbf{h}^2(0),\mathbf{h}^1(0))}{\partial \mathbf{A}^1}+\sum_{\mathbf{h}^2(1)}\sum_{\mathbf{h}^1(1)} P(\mathbf{h}^2(1),\mathbf{h}^1(1))\frac{\partial E(\mathbf{h}^2(1),\mathbf{h}^1(1))}{\partial \mathbf{A}^1} \\ &=<\mathbf{h}^1(0)\mathbf{h}^2(0)>_{data}-<\mathbf{h}^1(1)\mathbf{h}^2(1)>_{recon}\end{aligned} \quad (5.15)$$

where $\langle\cdot\rangle_{data}$ denotes an expectation with respect to the data distribution and $\langle\cdot\rangle_{recon}$ denotes the "reconstruction" distribution of data after one step. This leads to a simple learning rule for performing the stochastic steepest ascent in the log probability of the training data in Eq. (5.16) and Eq. (5.17).

$$A^1_{ij,pq}=\vartheta A^1_{ij,pq}+\vartriangle A^1_{ij,pq} \quad (5.16)$$

$$\vartriangle A^1_{ij,pq}=\varepsilon_{\mathbf{A}}(<h^1_{ij}(0)h^2_{pq}(0)>_{data}-<h^1_{ij}(1)h^2_{pq}(1)>_{recon}) \quad (5.17)$$

Other parameters in the θ^1 update function can be calculated in a similar manner.

$$b^1_{ij} = \vartheta b^1_{ij} + \triangle b^1_{ij} = \vartheta b^1_{ij} + \varepsilon_{\mathbf{b}}(h^1_{ij}(0) - h^1_{ij}(1)) \tag{5.18}$$

$$c^1_{pq} = \vartheta c^1_{pq} + \triangle c^1_{pq} = \vartheta c^1_{pq} + \varepsilon_{\mathbf{c}}(h^2_{pq}(0) - h^2_{pq}(1)) \tag{5.19}$$

where ϑ is the momentum and $\varepsilon_{\mathbf{A}}$, $\varepsilon_{\mathbf{b}}$, $\varepsilon_{\mathbf{c}}$ are the learning rate of model parameters $\mathbf{A}$, $\mathbf{b}$, and $\mathbf{c}$.

As far as we know, all existing deep learning models determine the structure, such as the sizes of the hidden layers, based on intuition. In our proposed model, we intend to provide a more meaningful architecture by integrating the determinative information from labeled data. To integrate discriminative information obtained from Bilinear discriminant projection for classification, we have two procedures: determining the sizes of hidden layers and calculating the discriminative initial symmetrically weighted connections.

As described before, we find a bilinear projection that can automatically reduce the original dimension $I \times J$ to $P \times Q$ through the transformation matrices $\mathbf{U}^1$ and $\mathbf{V}^1$. As a result, the number of neurons in layer H^2 is determined by the row and column size of the transformation matrices $\mathbf{U}^1$ and $\mathbf{V}^1$.

$$P^2 = row(\mathbf{U}^1)\,,\ Q^2 = column(\mathbf{V}^1) \tag{5.20}$$

Furthermore, in existing deep learning models, the weights of the symmetrical connections $\mathbf{A}$ are initialized to small random values chosen from a zero-mean Gaussian with a standard deviation of about 0.01. Differently from them, we set the discriminative transformation parameters obtained from the Bilinear discriminant projection as the initial weights of the symmetrical connections by Eq. (5.21).

$$A^1_{ij,pq}(0) = (\mathbf{U}^1_{ip})^T \mathbf{V}^1_{jq} \tag{5.21}$$

The above discussion is the greedy layer-wise abstraction for the first layer H^1

with its next adjacent layer H^2. Similar operations can be performed on the higher layer pairs.

5.4.4 Global fine-tuning

Above, we use the greedy layer-by-layer algorithm to learn a deep model with the help of discriminant information obtained from Bilinear discriminant projection. In this section, we use backpropagation through the whole deep model to fine-tune the parameters $\theta = [\mathbf{A}, \mathbf{b}, \mathbf{c}]$ for optimal reconstruction.

In the greedy layer-by-layer information abstraction stage, a global search has been performed for a sensible and good region in the whole parameter space. Therefore, before proceeding to the process of fine-tuning, we have already constructed a good data concept extraction model. In our model, backpropagation is utilized to adjust the entire deep network to find good local optimum parameters $\theta^* = [\mathbf{A}^*, \mathbf{b}^*, \mathbf{c}^*]$ to effectively classify the data. In this stage, the learning algorithm is used to minimize the classification error $[-\sum_l \mathbf{y}_l \log \widehat{\mathbf{y}_l}]$, where $\mathbf{y}_l$ and $\widehat{\mathbf{y}_l}$ are the correct label and the output label value of labeled sample datum $\mathbf{X}_l$ in X^L.

5.4.5 Algorithm and discussion

In this section, we firstly provide the detailed procedure of the BDBN in Algorithm 5.1. Then, we give some discussion about the how to construct attention model based on BDBN.

```
Input:  Training data $X$, Labeled samples $X^L$, Corresponding labels set $Y$
        Number of layers $N$, Number of epochs $E$,
        Number of labeled data $L$, Parameter $\alpha$,
        Between-class weights $\mathbf{B}_{st}$, Within class weights $\mathbf{W}_{st}$
        Initial bias parameters $\mathbf{b}$ and $\mathbf{c}$, Momentum $\vartheta$
        learning rate $\varepsilon_{\mathbf{A}}$, $\varepsilon_{\mathbf{b}}$, $\varepsilon_{\mathbf{c}}$
Output: Optimal parameter space $\theta^* = [\mathbf{A}^*, \mathbf{b}^*, \mathbf{c}^*]$

for $n = 1,\ldots, N$ do
    for $e = 1,\ldots, E$ do
        if $n = 1$
            $T^n = X^L$
        else
            for $l = 1,\ldots, L$ do
                $\mathbf{T}_l^n = \sigma(\mathbf{T}_l^{n-1} A^{n-1} + c^{n-1})$
            end for
        end if
        while not convergent do
            $\mathbf{D}_{\mathbf{V}} = \sum_{st} \mathbf{E}_{st} (\mathbf{X}_s - \mathbf{X}_t) \mathbf{V}\mathbf{V}^T (\mathbf{X}_s - \mathbf{X}_t)^T$
            $\mathbf{D}_{\mathbf{U}} = \sum_{st} \mathbf{E}_{st} (\mathbf{T}_s^n - \mathbf{T}_t^n)^T \mathbf{U}\mathbf{U}^T (\mathbf{T}_s^n - \mathbf{T}_t^n)$
            Fix V/U, compute U/ V by solving $\mathbf{D}_{\mathbf{V}}\mathbf{u} = \boldsymbol{\lambda}\mathbf{u} \,/\, \mathbf{D}_{\mathbf{U}}\mathbf{v} = \boldsymbol{\lambda}\mathbf{v}$
        end while
        Determine the size of next layer $P^{n+1} = row(\mathbf{U}^n)$, $Q^{n+1} = column(\mathbf{V}^n)$
        Compute initial weights of the connections $A_{ij,pq}^n(0) = (\mathbf{U}_{ip}^n)^T \mathbf{V}_{jq}^n$
        Calculate the state of the next layer
            $p\left(h_{pq}^{n+1} = 1 \mid \mathbf{h}^n\right) = \sigma(\sum_{i=1,j=1}^{i \le P^n, j \le Q^n} h_{ij}^n A_{ij,pq}^n + c_{pq}^n)$, $p\left(\mathrm{h}_{ij}^n = 1 \mid \mathbf{h}^{n+1}\right) = \sigma(\sum_{p=1,q=1}^{p \le P^{n+1}, q \le Q^{n+1}} A_{ij,pq}^n h_{pq}^{n+1} + b_{ij}^n)$
        Update the weights and biases
            $A_{ij,pq}^n = \vartheta A_{ij,pq}^n + \varepsilon_{\mathbf{A}} (< h_{ij}^n(0) h_{pq}^{n+1}(0) >_{data} - < h_{ij}^n(1) h_{pq}^{n+1}(1) >_{recon})$
            $b_{ij}^1 = \vartheta b_{ij}^1 + \varepsilon_{\mathbf{b}} (h_{ij}^1(0) - h_{ij}^1(1))$, $c_{pq}^1 = \vartheta c_{pq}^1 + \varepsilon_{\mathbf{c}} (h_{pq}^2(0) - h_{pq}^2(1))$
    end for
end for
Calculate optimal parameter space $\theta^* = \arg\min_{\theta} [-\sum_l \mathbf{y}_l \log \hat{\mathbf{y}}_l]$
```

Algorithm 5.1: Bilinear deep belief networks

As a cognitive process of selectively concentrating on one aspect of the environment while ignoring other things, attention has been referred to as the allocation of processing resources. Construction of attention map is useful for applications in multimedia like object segmentation, object recognition and so on. BDBN automatically extracts and emphasizes the important areas of image just like the attention allocation. Therefore, it is natural to extend the proposed model BDBN to construct a computational attention map. In our model, the first RBM can be utilized to construct the attention model as Figure 5.3. To every neuron in the input layer, the weight value to the one in the first hidden layer is calculated as feature map. Then, the weight value of every neuron is normalized and combined into an attention map. In the experiment part, the attention map based on BDBN will be compared with other computational attention maps.

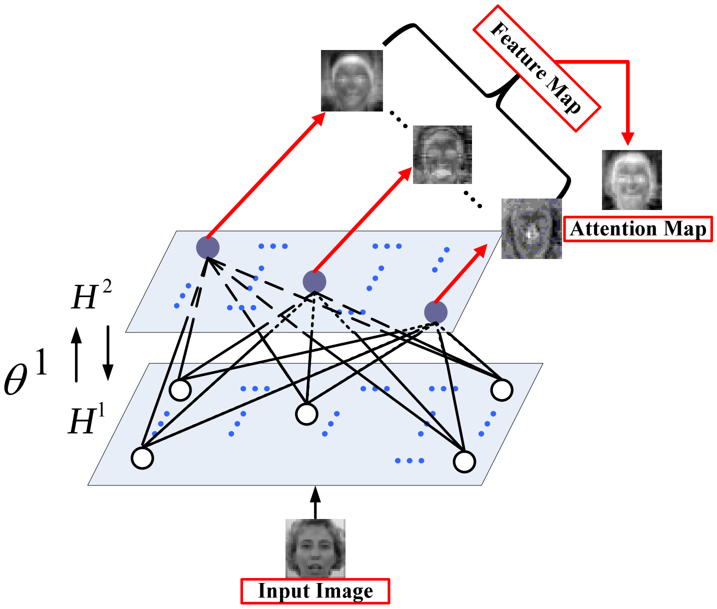


Figure 5.3. Construct attention map by first RBM in Bilinear deep belief networks.

5.5 Performance Evaluation

In this section, five standard datasets with different kinds of visual data are used to demonstrate the performance of the proposed BDBN. The first dataset is a standard large handwritten digits dataset MNIST, containing 70,000 images with 10 classes (LeCun et al., 1998). The second dataset is the Caltech101, a standard dataset for image classification, which includes images of 100 different objects plus a background category (Li et al., 2004). In this chapter, we use images from the first five categories. The third dataset is the Urban and Natural Scene. This dataset is composed of 2,688 color images with eight categories (Oliva & Torralba, 2001). The fourth dataset is the CMU pose, illumination, and expression (PIE) dataset (Sim & Baker, 2003). The fifth dataset is the BioID face dataset which consists of 1521 gray level images collected contains 23 subjects (Jesorsky et al., 2001).

For simplicity, we set the balance weight α as 0.5 in our experiments. For parameters such as the learning rate and the momentum in the deep learning model, we simply follow the general setting of previous work on deep learning (Bengio et al., 2006), although a more careful choice may lead to better performance. For example, in greedy layer wise learning, the number of epochs is fixed at 30 and the learning rate η is equal to 0.1.The initial momentum ϑ is 0.5. After five epochs, the momentum is set to 0.9. In the fine-tuning stage, the method of conjugate gradients is utilized and three line searches are performed in each epoch until convergence.

We compare the performance of BDBN with other representative classifiers, including *k*-nearest neighbor (KNN), Support vector machines (SVM) (Boser et al., 1992), Transductive SVM (TSVM) (Collobert et al., 2006), Neural network (NN)

(Mitchell, 1997), EmbedNN (Weston et al., 2008), Semi-DBN (Bengio et al., 2006), DBN-rNCA (Salakhutdinov & Hinton, 2007.), DDBN (Zhou et al., 2010), and DCNN (Jarrett et al., 2009). KNN, a typical nonlinear classifier, is always used as the baseline for comparisons of performance. In this chapter, we set k equal to 3. SVM and NN are two powerful classifiers. EmbedNN is the semi-supervised version of NN with deep architecture, which strengths the discriminative information in the fine-tuning stage. Semi-DBN, DBN-rNCA, and DDBN are the semi-supervised versions of DBN. As a new deep learning model, DCNN demonstrated great classification ability due to its ability to preserve visual locality and space structure.

5.5.1 Experiments on MNIST

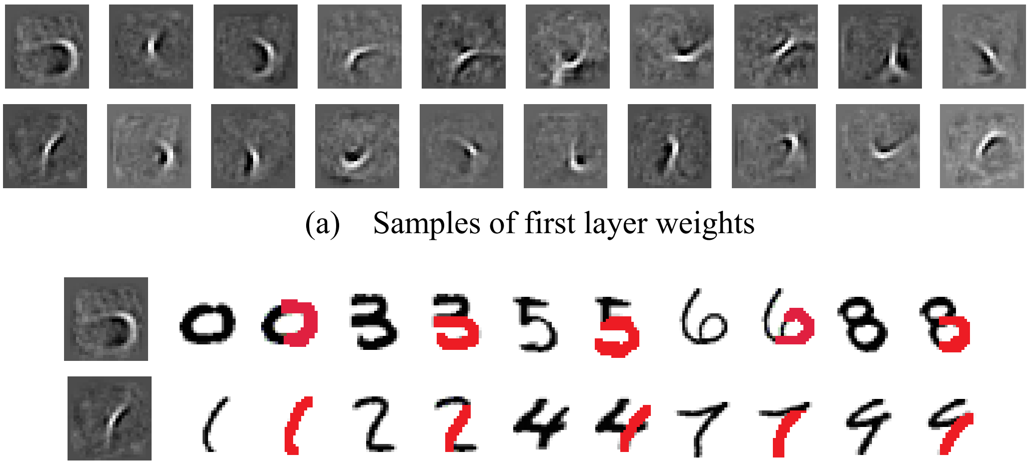
In this part, we demonstrate the performance of BDBN on image dataset of handwritten digits MNIST (LeCun et al., 1998). MNIST is a standard large database of hand written digits containing 60,000 training images and 10,000 test images with 10 classes. The resolution of images is 28×28. MNIST is often used to compare deep learning performance (Salakhutdinov & Hinton, 2007) (Weston et al., 2008).

The first experiment in this dataset is used to demonstrate the effectiveness of BDBN on MNIST. Different numbers of images of training data are randomly selected and labeled while the other training data remain unlabeled. The number of selected labeled data in each category is equal to 1, 2, 5 and 10, respectively. We perform 10 random splits and report the average results over the 10 trials. Table 5.1 shows the classification accuracy rate of the test dataset from different classifiers. From Table 5.1, it can easily be seen that, compared with supervised learning algorithms, the semi-supervised versions achieved a better performance. For example, EmbedNN is better than NN; Semi-DBN, DBN-rNCA and DDBN are better than DBN. It is obvious that deep learning models such as Semi-DBN, DBN-rNCA, DCNN, DDBN and BDBN achieve a competitive performance. This proves that the deep learning model has great abstraction ability. Owing to the discriminant information obtained from the labeled data between adjacent layers, the separability of abstraction in layer by layer has been promoted. Thus, our proposed BDBN has better performance than other classifiers.

In the second experiment, some samples of first layer weights learned from MNIST are given. The current consensus seems to be that the responses of V1 neurons consist of tiled sets of selective spatial information filters. The functioning of V1 can be thought of as similar to many spatially local, complex Fourier transforms, or more accurately, Gabor transforms. From Figure 5.4 (a), it is easy to see that many weights found by the algorithm roughly represent different "strokes" of which handwritten digits are comprised. And these weights of first layer are also oriented, Gabor-like and resemble the receptive fields of V1 simple cells. As shown in Figure 5.4 (b), with the help of these weights, we could classify hand written digitals easily.

Table 5.1. Classification accuracy rate (%) on the test data with different numbers of labeled data per category on MNIST.

Num./Cat.	1	2	5	10
KNN	51.61	54.10	67,65	77.32
SVM	52.55	55.65	70.07	73.00
NN	51.27	55.01	68.68	74.11
EmbedNN	51.50	55.40	64.00	76.80
Semi-DBN	55.23	60.64	79.10	83.56
DBN-rNCA	**60.45**	62.72	74.16	81.29
DCNN	55.70	62.99	78.37	85.51
DDBN	58.78	64.81	79.33	85.87
BDBN	60.07	**66.05**	**79.36**	**86.01**



(a) Samples of first layer weights

(b) Examples represent different “strokes” of handwritten digital

Figure 5.4. Samples of first layer weights learned from MNIST.

5.5.2 Experiments on Caltech101

In this experiment, we work on the frequently used subset of the Caltech101 (Zhou et al., 2010), which includes 2,935 images from the first five categories. The subset includes: 435 images of “Faces_easy,” 435 images of “Faces,” 798 images of “Motorbikes,” 800 images of “Airplanes,” and 467 images of “Back_google.” As shown in Figure 5.5, the images in the same category vary greatly.

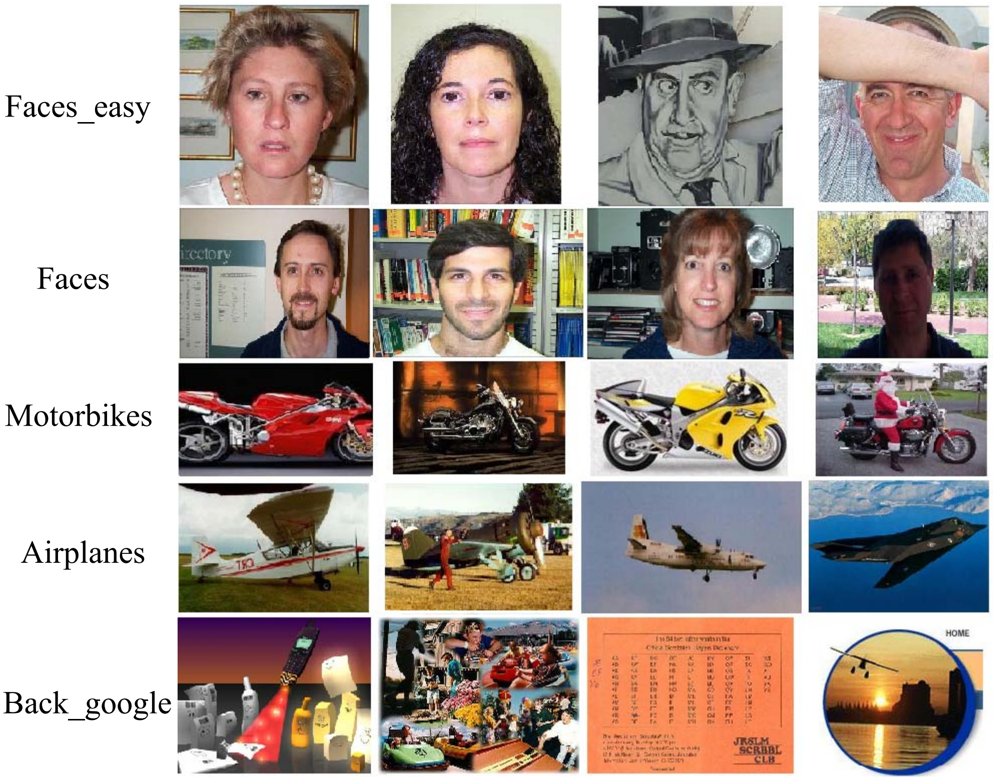


Figure 5.5. Sample images from the dataset Caltech101.

First, we compare the classification accuracy of different methods with a various number of labeled data. Because the number of images in each category in Caltech101 is different, 50 images are randomly selected for each category to form the test set and the rest to form the training set. As the previous setting in (Zhou et al., 2010), the number of labeled data is equal to 5, 25, 50, and 75 per category, respectively. We perform 10 random splits and report the average results over the 10 trials. As shown in Table 5.2, the performance of BDBN is stable and impressive.

Then, we compare the convergence of the proposed BDBN with two other deep learning models: Semi-DBN and DDBN, all of which have a fine-tuning stage. Figure 5.6 shows that BDBN converges much more quickly than Semi-DBN and DDBN. Although they are all deep learning models, BDBN requires an average of 106 iterations in comparison to 290 iterations for Semi-DBN and 161 iterations for

DDBN. The improvement comes from the strategy of Bilinear discriminant projection. This strategy helps BDBN to achieve a better “initial guess” when constructing the symmetrically weighted connections between layers.

Table 5.2. Classification accuracy rate (%) on the test data with different number of labeled data per category on Caltech101.

Num./Cat.	5	25	50	75
KNN	44.60	58.20	63.20	64.60
SVM	49.80	66.20	67.40	68.20
TSVM	50.00	70.20	70.50	72.80
NN	53.20	64.00	66.80	70.60
EmbedNN	51.20	55.50	58.60	64.00
Semi-DBN	55.40	65.80	67.60	69.60
DBN-rNCA	55.80	64.20	65.40	69.80
DDBN	58.30	71.40	72.00	74.20
DCNN	58.20	70.80	73.40	75.20
BDBN	**61.80**	**71.60**	**75.60**	**78.80**

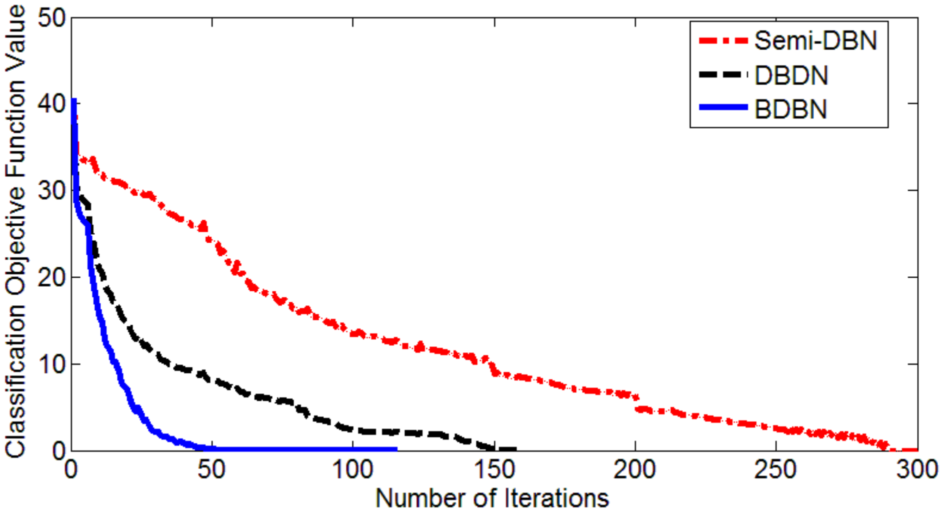


Figure 5.6. Convergence curve of Semi-DBN, DDBN and BDBN on Caltech 101.

Table 5.3. Comparisons of run-time (s) and classification accuracy (%) with different labeled numbers and different deep architectures on Urban and Natural Scene.

Num. / Cat.	5		25		50		75	
	Run-time (s)	Acc. (%)	Run-time (s)	Acc. (%)	Run-time (s)	Acc. (%)	Run-time (s)	Acc. (%)
NN_d	**378**	22.25	1340	30.50	2693	31.50	5796	32.75
NN_c	438	22.50	3602	27.25	6791	30.25	9948	32.50
EmbedNN_d	435	26.75	1373	32.50	2722	35.00	5913	37.50
EmbedNN_c	523	27.50	3702	32.75	6831	36.50	10219	38.25
Semi-DBN_d	769	29.50	1275	33.50	2402	37.25	5945	40.25
Semi-DBN_c	1394	30.50	3467	34.25	7792	37.70	22887	39.50
DBN-rNCA_d	712	29.25	1156	35.25	2209	36.50	5197	41.25
DBN-rNCA_c	1134	30.75	3223	35.25	6565	37.00	18452	42.50
DDBN_d	658	31.25	1051	37.00	2126	41.25	5142	49.20
DDBN_c	1045	32.00	2987	38.25	5292	42.50	16737	51.00
BDBN	392	**35.25**	**963**	**42.50**	**2056**	**50.75**	**5101**	**55.25**

5.5.3 Experiments on Urban and Natural Scene

In this section, we demonstrate the performance of BDBN on the Urban and Natural Scene dataset (Oliva & Torralba, 2001). This dataset is composed of 2,688 color images with eight categories, namely “coast & beach,” “highway,” “open country,” “tall building,” “forest,” “street,” “mountain,” and “city center.” In the preprocessing stage, images are downsampled to 32×32 as the input of BDBN. In our experiment, 50 images are randomly selected from each category to form the test set and the rest of the images are used for training. Sample images of each category are shown in Figure 5.7.

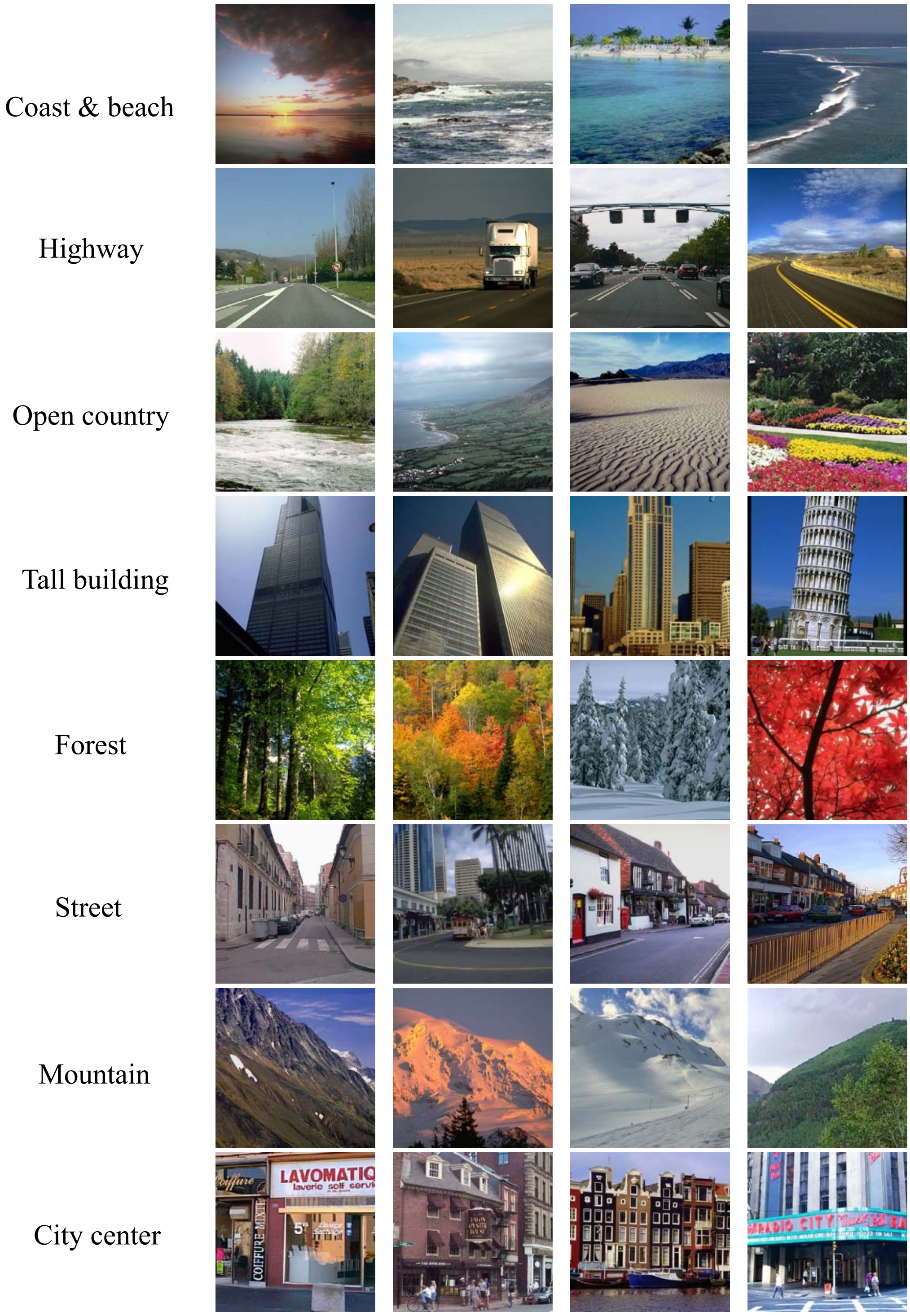


Figure 5.7. Sample images from the Urban & Natural Scene.

All existing deep learning models determine the structure, such as the sizes of the hidden layers, based on researchers' intuition. In our model, the number of the neurons in each layer can be determined automatically based on bilinear discriminant strategy. Table 5.3 demonstrates this advantage by comparing the real running time and classification accuracy of BDBN with other five neural networks. The number of labeled data is equal to 5, 25, 50 and 75 per category, respectively. We perform 10 random splits and report the average results over the 10 trials. For BDBN, the number of neurons in layer H^1 is the size of the input image, i.e. 32×32. The number of neurons in H^2, H^3, H^4 is 24×24, 21×21, and 19×20, respectively. The classical setting of neurons numbers in H^2, H^3, H^4 are 500, 500, and 2000, according to previous publications. The results with different sizes of the deep architecture are provided for the models under comparison. In the table, "_d" is used to represent the compared models with the same size of BDBN, and "_c" is utilized to represent the compared models with the classical sizes. Clearly, BDBN has lower time complexity and better classification accuracy.

In Figure 5.8, we discuss the limitation of image classification based on visual similarity. Figure 5.8 (a) is a representative image of "Street", and Figure 5.8 (b) is a representative image of "Highway". Figure 5.8 (c) is classified to be "Highway" by BDBN and all other classifiers in this experiment, although the ground-truth of this image is "Street". Only according to visual similarity, Figure 5.8 (b) and Figure 5.8 (c) should be grouped together. However, human can give the correct judgment of Figure 5.8 (c) by referencing the buildings and cars along the street, which is a kind of contextual cueing acquired from past experiences of regularities. We list it as the future work of integrating contextual cueing in the deep modeling.

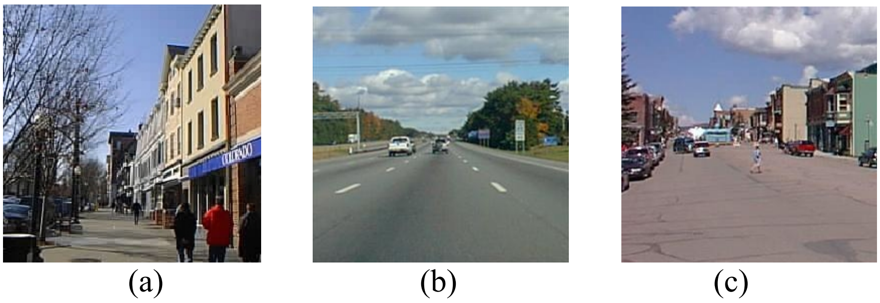


Figure 5.8. Limitation of image classification via visual similarity. (a) A representative image of “Street”. (b) A representative image of “Highway”. (c) The misclassified image. The ground-truth category of it is “Street” and the misclassified category is “Highway”.

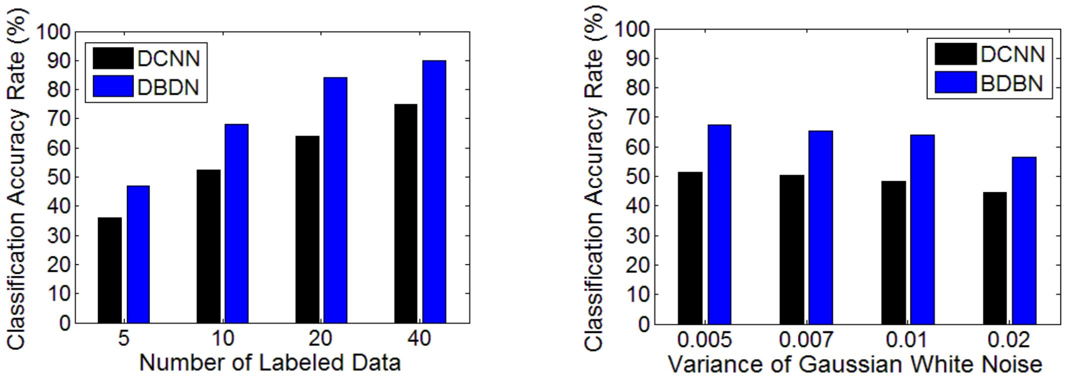
5.5.4 Experiments on CMU PIE

In this part, we demonstrate the performance of BDBN on image dataset of the CMU PIE dataset (Sim & Baker, 2003). The CMU PIE face dataset contains 68 subjects with a total of 41,368 face images. The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination and expression. As with the general setting of experiments to build the sub dataset (He et al., 2005), we use all the images under different illuminations and expressions with five near frontal poses (C05, C07, C09, C27, C29). In this way, about 170 images with the resolution of 32×32 are obtained for each individual. The preprocessing is applied following the general setting of experiment (He et al., 2005).

In the above experiments, the convolutional deep learning model demonstrates a better performance than other existing deep models. Therefore, in the experiment for dataset PIE, we compare the robustness of our deep model BDBN with that of the convolutional deep model DCNN. For the dataset, 120 images are randomly selected for each person to form the training set and the rest to form the test set. We perform 10 random splits and report the average results over the 10 trials.

First, we compare the influence from different number of labeled data with the same extent of noise. Of the 120 images for each person, different numbers of images are randomly selected and labeled while the others remain unlabeled. The number of labeled data per subject is equal to 5, 10, 20 and 40. The Gaussian white noise with a mean of 0 and a variance 0.003 is added to the intensity image. According to the average classification results shown in Figure 5.9 (a), it is obvious that the classification accuracy increases with the number of labeled data. In addition, BDBN exhibits better performance than DCNN under all conditions.

Second, we compare the influence from different extents of noise with same number of labeled data. Here, we fix the number of labeled data per subject to be 10. The variance of Gaussian white noise changes from 0.005 to 0.02. From Figure 5.9 (b), although the classification accuracy decreases along with the increase of noise in both BDBN and DCNN, our technique performs better. Thus, we are able to conclude that, although DCNN is famously invariant to variations or noises (Lee et al., 2009) (Taylor et al., 2010), our proposed BDBN is more robust than DCNN.



(a) (b)

Figure 5.9. (a) Classification Accuracy rate (%) with different number of labeled data (b) Classification Accuracy rate (%) with different extents of noise.

Why does BDBN always performs better than DCNN for noisy images? Figure 5.10 is intended to provide some interpretation from the data reconstruction. The images with Gaussian white noise with a mean of 0 and variance of 0.005 are inputted to BDBN, as shown in the first row. The results of the reconstruction in every layer are shown from the second to the fourth row. It is apparent that, after three layer-wise information reconstruction, the noises have been removed. In addition, the reconstructed images are more similar to the original images shown in the fifth row.

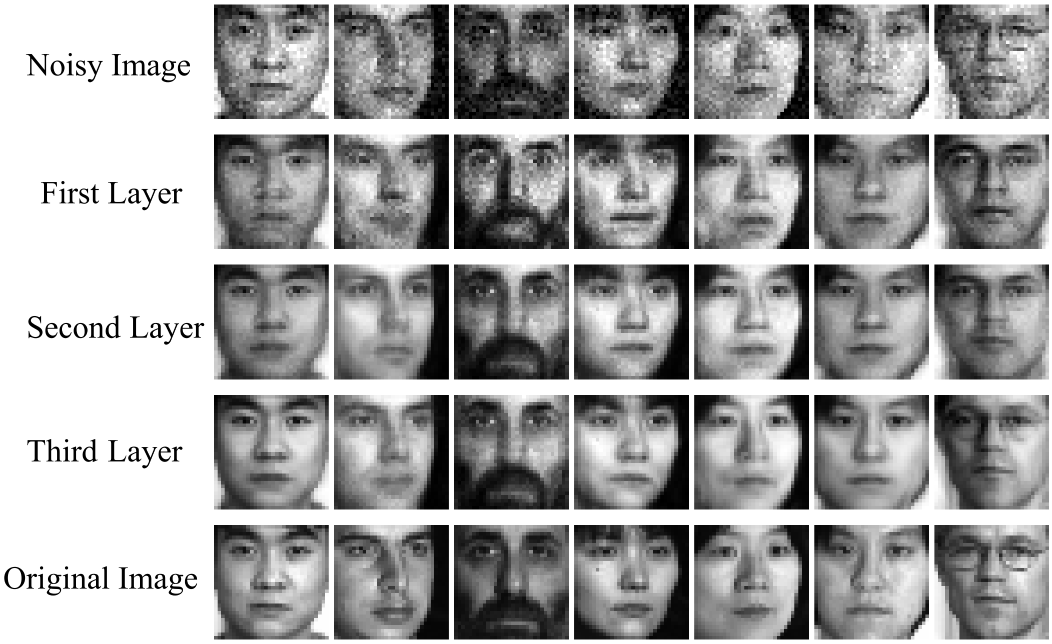


Figure 5.10. The reconstruction of BDBN in every layer. The first row shows the noisy images. The reconstruction results of every layer are shown from the second to the fourth row. The original images are shown in the fifth row.

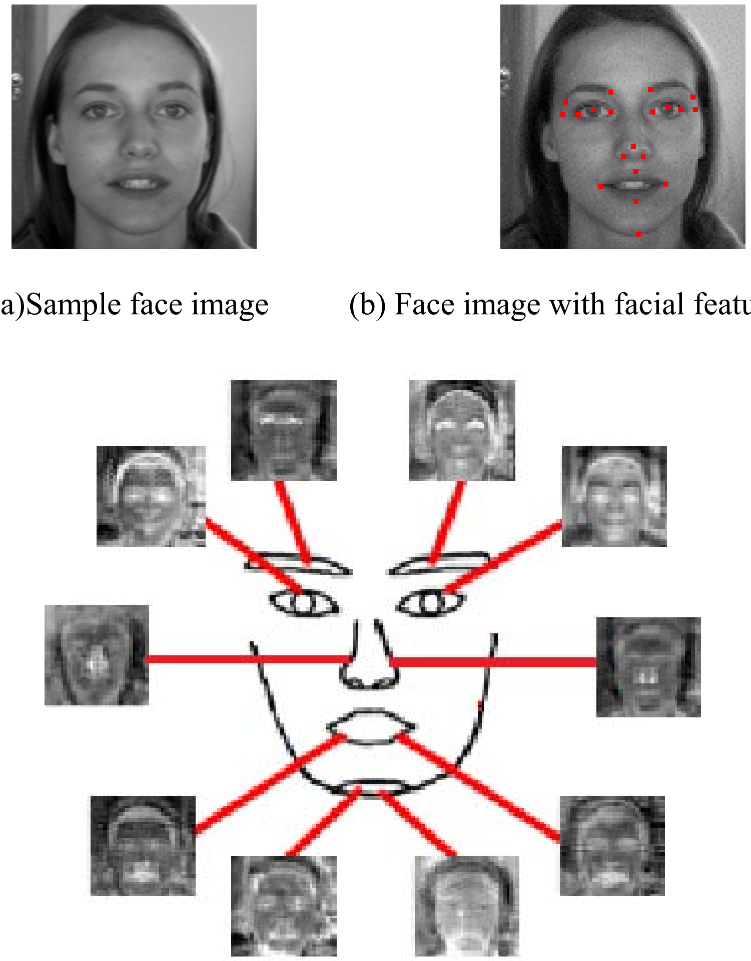
5.5.5 Experiments on BioID

In this experiment, we intend to investigate the consistency between the emphasized regions in BDBN and the attention map of human being. In BioID, for every image, 20 important facial feature points are manually selected out and placed. With marked facial feature points, the attention model based on deep learning model could be evaluated without eye tracking recordings (Jesorsky et al., 2001).

The number of images in every category of BioID is varied, from 35 to 118. Therefore, we choose the categories with more than 50 face images as the subset we work on. To demonstrate the effectiveness of our model, firstly, the visualization of the parameter space of proposed model is observed. Figure 5.11 (a) shows a sample image, and Figure 5.11 (b) shows the sample image with the facial feature points. Figure 5.11 (c) visualizes the parameter spaces between the input layers and the first hidden layer in BDBN. Each picture shown below represents one neuron in the hidden layer and each pixel quantizes the weight value between that neuron and the one in the input layer. Obviously, the proposed BDBN can automatically extract and emphasize the important areas of human's face, such as the eyes, eyebrows, noses, cheeks, mouths and chins.

Then, we construct the saliency regions based on the emphasized regions of BDBN. Just like the Figure 5.11 (c), the weight value between each neuron in the input layer to the one in the first hidden layer is calculated at first. Then, the weight value of every neuron is normalized and combined into a saliency map. According to the x and the y coordinates of the 20 important facial feature points of every face image in the dataset, we statistically analysis the percentage of all facial feature points located in the saliency regions of the saliency map.

There are 63.71% facial feature points are located inside 30% most saliency regions and only about 1% facial feature points are located outside 80% most saliency regions. In Figure 5.12, the comparison of different computational attention maps are provided, including Graph Gabor attention map (Harel et al., 2007), Itti classical attention map (Itti & Koch, 2000) and BDBN attention map. It is obviously that BDBN has better coverage than other models. It proves that BDBN provides a human-like judgment by referencing the human visual system.

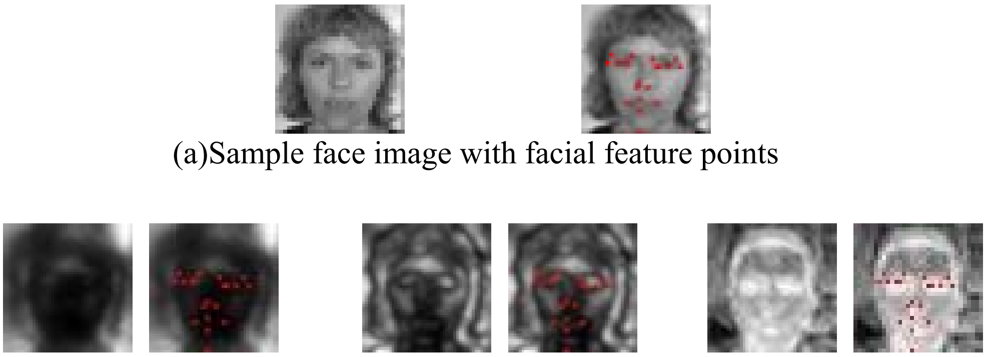


(a)Sample face image

(b) Face image with facial feature points

(c) Emphasized regions of BDBN corresponding to facial feature regions

Figure 5.11. Samples of first layer weights learned by BDBN, and the consistency of these weights with facial feature points.



(a)Sample face image with facial feature points

(b) Graph Gabor map

(c) Itti map

(d) BDBN map

Figure 5.12. The comparison of different attention maps with facial feature points.

5.6 Summary

In this chapter, we propose a novel learning model, BDBN for a classical multimedia content analysis task, image classification. This learning model is faithful to the physiology and the anatomy of the visual cortex, provides a realistic alternative to engineered artificial vision systems. BDBN has several attractive characters. First, the novel deep architecture of BDBN simulates the multi-layer physical structure of the visual cortex and enables the preservation of the natural tensor structure of the input image in the information propagation. Second, the three-stage learning of BDBN faithfully realizes the procedure of object recognition by human beings, especially for the “initial guess” part, which has never been modeled in deep learning. Third, the bilinear discriminant initialization of BDBN not only prevents the propagation of information from falling into a bad local optimum but also provides a more meaningful setting for deep architecture. Fourth, the semi-supervised learning ability of BDBN causes the proposed deep techniques to work well with an insufficient number of labeled data.

Experiments on four real-world image classification tasks and one attention map construction task not only show the distinguishing classification ability of BDBN but also clearly demonstrate our intention of providing a human-like image analysis by referencing the human visual system and perception procedure.

Chapter 6 Visual Cortex Simulation for Image Recognition with Incomplete Data

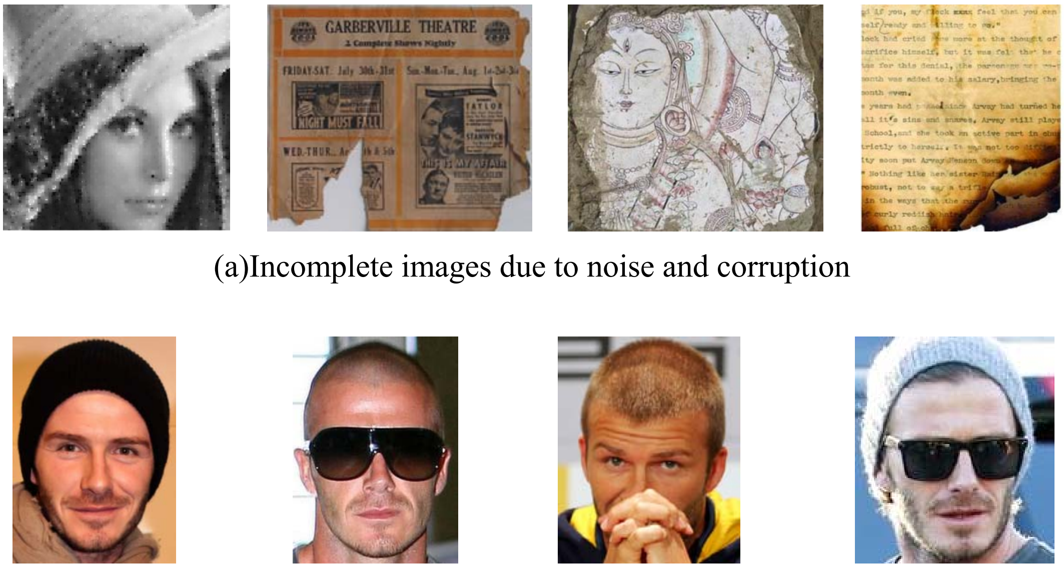
6.1 Overview

Image recognition with incomplete data is a well-known hard problem in computer vision and machine learning. This chapter proposes a novel deep learning technique called Field effect bilinear deep belief networks (FBDBN) to seek the recognition discriminant boundary and estimate the missing features jointly (Zhong et al., 2012c). Inheriting from deep belief networks, FBDBN simulates the laminar structure of human's cerebral cortex and the neural loop in human's visual areas, hence shows good performance in visual data analysis. To address the difficulties of incomplete data, we design a novel second-order deep architecture with the Field effect restricted boltzmann machines, which models the reliability of the delivered information according to the availability of the features. Based on the new architecture, two peaks activation with the bi-directional inference of human's perception is implemented by three learning stages of Field effect bilinear discriminant initialization, Field effect layer-wise abstraction and estimation, and global fine-tuning with missing features adjustment.

Owing to construct decision boundary from the reliable feature spaces and filling the missing value to the greatest extend, the proposed FBDBN has demonstrated impressive recognition performance on three datasets under supervised, semi-supervised and unsupervised framework. Furthermore, the missing features in the incomplete images are inferred and estimated effectively.

6.2 Background and Motivation

Incomplete data, data values are partially observed (Liao et al., 2007), exists in a wide range of fields, including social sciences, computer vision, and remote sensing (Williams et al., 2007). In general, features missing in real-world data are resulted from measurement noise, corruption or occlusion (Chechik et al., 2008). Figure 6.1 shows some real examples of incomplete data. Figure 6.1 (a) provides some images with missing features due to noise and corruption, including: noisy photo, old broken movie poster, ancient fresco, and a burned paper with some available handwriting. Obviously, it is more difficult for computer to recognize meaningful patterns with the incomplete data. If the image distortion is very serious, even human beings can’t recognize the images correctly. Figure 6.1 (b) provides more general cases of incomplete data in our daily life. David Beckham is one of the most iconic athletes and most fans have no difficulty to recognize him from these four images. But it is not an easy task for many face recognition models because some key facial features to identify persons, such as characters of eyes and mouth, are not observable.



(a)Incomplete images due to noise and corruption

(b) Incomplete face images due to the occlusion in important facial feature regions

Figure 6.1. The examples of incomplete images due to noise, corruption or occlusion.

Current works on incomplete data can be roughly categorized into three groups based on the modeling of the missing values (Dick et al., 2008). The first kind of techniques doesn't intend to estimate the missing values. They learn the decision function only based on the visible features. Chechik et al. recognized the incomplete data directly without any completion of the missing features using a max-margin learning framework (Chechik et al., 2007) (Chechik et al., 2008). For each sample, the margin is rescaled according to the visible attributes. Although their methods can avoid the additional workload introduced by estimating the unknown values, the recognition performance of them is limited by the numbers of available features. The second kind of techniques fills the missing values based on the modeling of the available information, and then learns the decision function in a general way. Williams et al. developed a logistic regression classification algorithm for incomplete data. Conditional density functions were estimated using a Gaussian mixture model, with parameter estimation performed using both expectation maximization (LRCEM) (Williams et al., 2005) and Variational Bayesian expectation maximization (LRCVBEM) (Williams et al., 2007). Shivaswamy et al. proposed a novel second order cone programming formulation (SOCP) for designing robust classifiers which can handle uncertainty in observations (Shivaswamy et al., 2006). The third kind of techniques seeks the final decision boundary by estimating the missing value and constructing predictive model jointly. Liao et al. proposed a statistical model names Quadratically gated mixture of experts (QGME) for multi-class nonlinear recognition (Liao et al., 2007). The model used linear classifiers as basic building blocks and mixed them through one-level quadratic gating. In their paper, they proved that the missing values entail joint estimation of the data and the classifier. Dick et al. derived a generic joint optimization Weighted infinite imputations (WII) method, which learned the decision function and the distribution of imputations dependently (Dick et al., 2008). The experiments demonstrated significant improvements over the methods that separate estimation from classifier learning.

Previous works on incomplete data show that the joint learning methods consistently outperform separated ones. Hence, this chapter intends to design a novel classifier to seek the decision boundary and estimate the missing values synchronously for incomplete image classification. We choose deep learning, which models the learning tasks using deep architectures composed of multiple layers of parameterized modules because of two considerations. Firstly, deep learning model has demonstrated distinguishing ability of information abstraction and robust performance of data classification in various visual data analysis tasks (Hinton & Salakhutdinov, 2006). Secondly, according to our previous works, deep learning has demonstrated impressive results in noisy data analysis, hence, it shows great potentials to address incomplete data problem.

To our knowledge, deep learning model has never been used for incomplete data classification mainly because of the difficulty of handling missing features. To address this problem, we propose a novel and flexible deep learning framework by jointly constructing the recognition discriminant boundary and estimating the missing values. The basic idea of our work is providing human-like judgment by referencing the architecture of human visual system and procedure of intelligent perception. Inhering from previous work in deep learning, the proposed framework simulates the laminar structure of human's cerebral cortex and the neural loop in human's visual areas. Moreover, the new framework utilizes second-order planes and bilinear disciriminant strategy to adapt the natural tensor character of the visual data.

The most important contribution of this chapter is that the proposed framework provides a novel and flexible framework to address image recognition with incomplete data by referencing human's perception. There are two characters of human visual system in incomplete images recognition. Firstly, humans can automatically adjust their attention to the available features and emphasize the contributions from them consciously and, actually sometimes unconsciously. Especially to the contrast gain attention, in comparison with the occluded part, the firing rates of the neurons will be increased preferring the available features (Ranzato et al., 2011). Secondly, additive attention could lead to the occluded parts of the object becoming active, as the feedback from higher levels travels down the visual stream based on the feedback connections in the visual cortex (Taylor et al., 2006). This process allows us to hallucinate occluded/undetected parts by filling-in the missing features based on top-down knowledge from the model, which plays an important role in identifying and complete objects when different portions are visible, or when parts are occluded or degraded (Kosslyn, 1994) (Enns & Rensink, 1998) (Aleman et al., 2003). It means that human can infer and estimate the incomplete parts by the reference data. In case the feature value of the training data is missing and no related information is useful to estimate this missing feature, human will automatically neglect the specific missing feature in the recognition.

To address these special characters of visual perception caused by incomplete data, we construct a reliability function to model the quality of the features, and propose a novel framework called Field effect bilinear deep belief networks (FBDBN). The reliability function has the similar output characteristic curve with field effect transistor, a common electronic device. To our model, we borrow this item to explain the physical meaning of the function. There are three different operating modes existing in the Filed-effect transistor (FET), including: the cutoff mode, ohmic mode, and the saturation mode. The operating modes are controlled by the gate-to-source voltage V_{GS}, the threshold voltage V_{th}, and the drain-to-source voltage V_{DS} as follows.

$$\begin{cases} \text{Cutoff mode,} & \text{if } V_{GS} \le V_{th}, \\ \text{Ohmic mode,} & \text{if } V_{GS} > V_{th} \text{ and } V_{DS} < V_{GS} - V_{th}, \\ \text{Saturation mode,} & \text{if } V_{GS} > V_{th} \text{ and } V_{DS} \ge V_{GS} - V_{th} \end{cases} \tag{6.1}$$

Similar with FET, the proposed reliability function also has three operating modes, which is consistent to the perception in human beings. In case the feature value of the training data is missing and no related information is useful, human will automatically neglect the specific missing feature in the recognition. To our model, the reliability function goes into the cutoff mode and the reliability related with the missing feature is set to be zero. If some related information is helpful to estimate the missing feature, human will attempt to estimate the missing feature and adjust the reliability derived from the missing feature. To our model, it is identical to the ohmic mode of FET and the reliability of the corresponding connection will be updated in the process of the recognition. Given that the estimation is stable and no further information is available and useful, human will stop adjusting the estimation and keep the reliability without change. To our model, the reliability function related with the missing feature will not change which is corresponding to the saturation mode of FET.

Compared with existing deep models, the proposed FBDBN has several attractive characters:

1) FBDBN is the first deep learning model developed specially for the incomplete image recognition task. Although by exploiting the generative ability of deep learning model, such as DBN, deep model has demonstrated the effectiveness in coping with occlusion (Ranzato et al., 2011), the information reconstruction relying on all features equally is not reliable when some features are missing. The proposed FBDBN integrates the reliability of the features in the learning procedure to make use of the availability information to the greatest extent.
2) FBDBN utilizes a FET-like characteristic curve to define the reliability of the features and bi-directional inference to deliver the information. By seeking the optimal decision boundary and estimate the missing values jointly, FBDBN shows impressive results in incomplete data recognition.
3) FBDBN is a unified framework of supervised learning, semi-supervised learning, and unsupervised learning schemes according to the availability of the labels, hence, it is facilitating for real-world applications with different kinds of data for different tasks.

The remainder of this chapter is organized as follows. A novel deep architecture and a new deep learning algorithm are introduced in Section 6.3. Section 6.4 shows the performance of the proposed techniques in image recognition and retrieval tasks and Section 6.5 concludes this chapter.

6.3 Field Effect Bilinear Deep Belief Networks

In this section, we propose a novel deep learning architecture based on Field effect bilinear deep belief networks (FBDBN). Our FBDBN, which is aimed at the task of incomplete image recognition, is described in Section 6.3.1. The Field effect bilinear discriminant initialization stage is discussed in Section 6.3.2. Section 6.3.3 contains details of the layer-wise abstraction and estimation by Field effect RBMs. The global fine-tuning process of the whole deep network is described in Section 6.3.4. Finally, the algorithm and some discussion about the generalization of FBDBN are provided in Section 6.3.5.

6.3.1 Framework of Field effect bilinear deep belief networks

Let *X* be a set of incomplete data samples as shown below:

$$X = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k, \ldots, \mathbf{X}_K] \tag{6.2}$$

where $\mathbf{X}_k$ is a sample datum with missing features in the image space $\mathbb{R}^{I \times J}$ and *K* is the number of sample data. Let F_k denote the set of missing features of the sample $\mathbf{X}_k$, $(X_k)_{ij}$ is missing if $(X_k)_{ij} \in F_k$. Let *Y* be a set of labels corresponding to *X*, which can be seen as:

$$Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k, \ldots, \mathbf{y}_K] \tag{6.3}$$

And $\mathbf{y}_k$ is the label vector of $\mathbf{X}_k$ in $\mathbb{R}^C$, where *C* is the number of classes.

$$y_k^c = \begin{cases} 1 & \text{if } \mathbf{X}_k \in c\text{th class} \\ 0 & \text{if } \mathbf{X}_k \notin c\text{th class} \end{cases} \tag{6.4}$$

Based on the given training set, the goal in image recognition is to learn a mapping function from the image set *X* to the label set *Y*, and then recognize the new coming data points according to the learned mapping function.

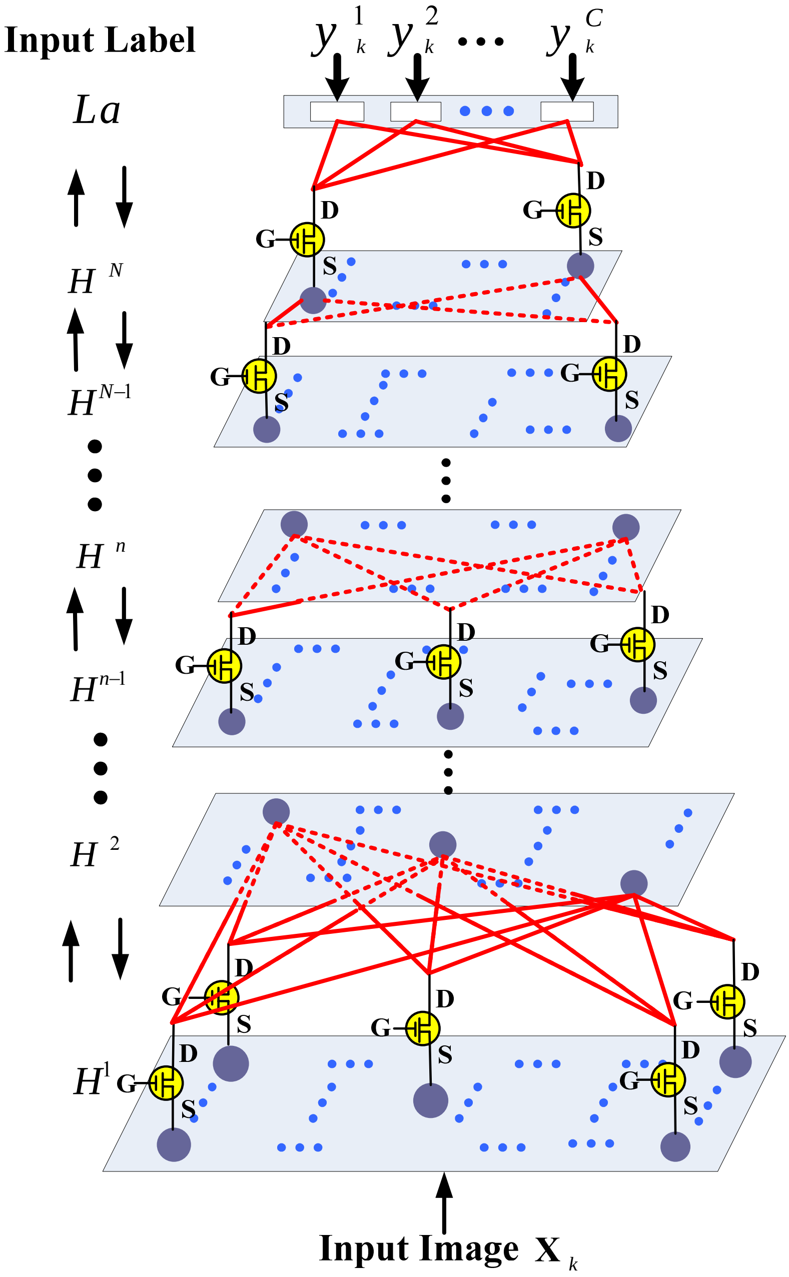


Figure 6.2. Architecture of FBDBN.

To address the problem of incomplete image recognition, we propose a novel deep learning technique called Field effect bilinear deep belief networks (FBDBN). Figure 6.2 shows the architecture of FBDBN. A fully interconnected directed belief network includes the input layer H^1, hidden layer $H^2, \ldots, H^N$, and one label layer La at the top. The input layer H^1 has $I \times J$ units, and this size is equal to the dimension of the input features. In our model, we use the pixel values of sample datum $\mathbf{X}_k$ as the original input features. In the top, the label layer has *C* units, which is equal to the number of classes. The search of the mapping function from *X* to *Y* is transformed to the problem of finding the optimum parameter space θ^*.

The most obvious character of our architecture is that the novel Field effect RBMs are proposed instead of the original RBMs. RBMs in Deep belief networks help us to abstract the embedding information by layer-wise reconstruction. Unfortunately, RBM cannot work when some features are missing, and the corresponding units of the networks are empty. Inspired from electronic circuits (Malik, 1995), FRBM construct the reliability weighted connection by FET analogy between the lower layer and the upper layer. To our model, the reliability parameter $\Re$ is defined by Eq. (6.5) via referring the operating mode and the output characteristic curve of FET.

$$\Re = \begin{cases} 0, & \text{if } V_{GS} \le V_{th}, \\ \sqrt{2(V_{GS} - V_{th})V_{DS} - V_{DS}{}^2}, & \text{if } V_{GS} > V_{th} \text{ and } V_{DS} < V_{GS} - V_{th}, \\ V_{GS} - V_{th}, & \text{if } V_{GS} > V_{th} \text{ and } V_{DS} \ge V_{GS} - V_{th} \end{cases} \tag{6.5}$$

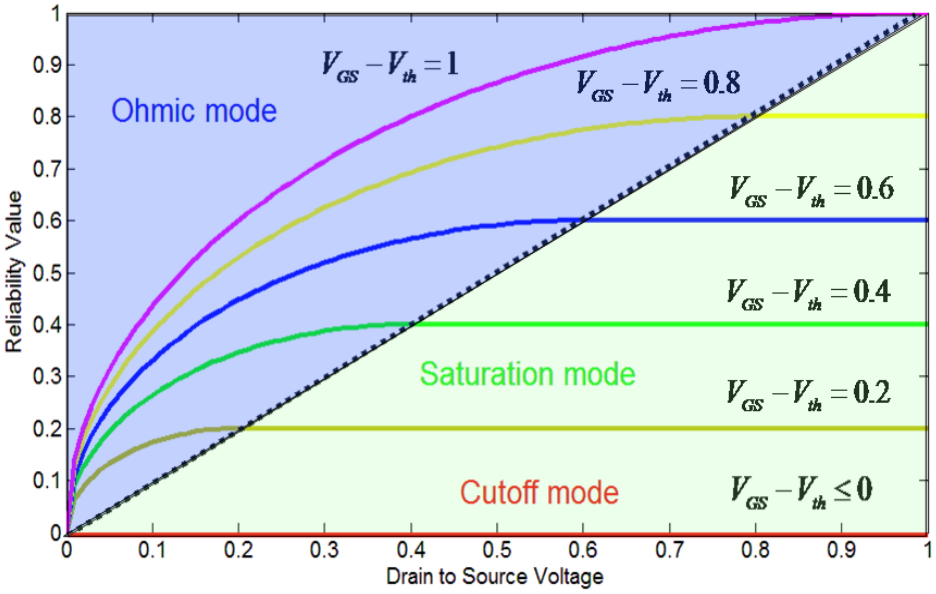


Figure 6.3. The output characteristic curve and the operating mode of Field effect RBMs depends on the voltage V_{GS}, V_{th} , and V_{DS}.

The output characteristic curve of FRBM is shown in Figure 6.3. Relying on three voltage parameters, the reliability could be automatically and adaptively adjusted. In our model, the gate-to-source voltage V_{GS} is defined based on the probability level of the estimated feature according to the distribution of all available features. The drain-to-source voltage V_{DS} is defined based on the similarity of the reference datum and incomplete datum.

The procedure under supervised or semi-supervised learning framework of our FBDBN is listed below:

1) The strategy of Field effect bilinear discriminant projection is utilized to construct a projection to map the original data into a discriminant bilinear subspace based on the features with high reliability. The initial symmetrically weighted connections are constructed between adjacent layers according to the "initial guess" based on the discriminant information of the features with high reliability. The size of the deep architecture is determined automatically based on the optimum dimension to retain the discriminant information.

2) In the “bi-directional inference” stage, the parameter space is refined using Field effect RBMs as building blocks. In the bottom-up inference, the whole deep learning model is constructed by the available features and the estimated features based on the reliability. In the top-down inference, the missing features are estimated by the higher layer activations of the reference datum.
3) In the “post activation” stage, the whole deep model is fine-tuned to minimize the recognition error and slightly adjust the estimation values of missing features data based on backpropagation.

6.3.2 Initial guess by Field effect bilinear discriminant initialization

In this subsection, we introduce the Field effect bilinear discriminant projection (FBDP), which is utilized to extract the discriminant information from the image datasets with incomplete features.

Given the training data points $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_K \in \mathbb{R}^{I \times J}$ with missing features set F_k, $(X_k)_{ij}$ is missing if $(X_k)_{ij} \in F_k$. FBDP aims to find two projection matrices $\mathbf{U} \in \mathbb{R}^{I \times P}$ and $\mathbf{V} \in \mathbb{R}^{J \times Q}$ such that the latent representation $\mathbf{TX_1, TX_2, ..., TX}_K \in \mathbb{R}^{I \times J}$ can be obtained by $\mathbf{TX}_k = \mathbf{U}^T \mathbf{X}_k \mathbf{V} \quad (k = 1, ..., K)$ from features with high reliability. Here, we define the reliability matrix $\mathbf{R}_k^F \in \mathbb{R}^{I \times J}$ of the features in $\mathbf{X}_k$. In the initial guess stage, the reliability matrix $\mathbf{R}_k^F$ is assigned as Eq. (6.6), just as the cutoff mode of FET.

$$(R_k^F)_{ij} = \begin{cases} 0, & \text{if } (X_k)_{ij} \in F_k \\ 1, & \text{else} \end{cases} \tag{6.6}$$

In order to preserve the discriminant information from features with high reliability in the learning procedure, the objective function of FBDP could be represented as follows:

$$\begin{aligned} &\arg\max_{\mathbf{U},\mathbf{V}} J(\mathbf{U},\mathbf{V}) = \sum_{s,t=1}^{K} (\alpha \mathbf{B}_{st} - (1-\alpha)\mathbf{W}_{st}) \| \mathbf{U}^T (\mathbf{X}_s .* \mathbf{R}^F_{st} - \mathbf{X}_t .* \mathbf{R}^F_{st}) \mathbf{V} \|^2 \\ &s.t.\ \mathbf{R}^F_{st} = \mathbf{R}^F_s .* \mathbf{R}^F_t, \mathbf{U}^T\mathbf{U} = \mathbf{I}_P,\ \ \mathbf{V}^T\mathbf{V} = \mathbf{I}_Q, \mathbf{X}_s \in X, \mathbf{X}_t \in X \end{aligned} \tag{6.7}$$

Different with the Bilinear discriminant projection (BDP) in (Zhong et al., 2011c), we extract the discriminant information based on the features with high reliability. In Eq. (6.7), $\alpha \in [0,1]$ is the parameter used to balance the between-class weights $\mathbf{B}_{st}$ and the within class weights $\mathbf{W}_{st}$, which are defined below:

$$\mathbf{B}_{st} = \begin{cases} \dfrac{1}{n_d} - \dfrac{1}{n_c}, & \text{if } \mathbf{y}^c_s = \mathbf{y}^c_t = 1, \\ \dfrac{1}{n_d}, & \text{else,} \end{cases}, \mathbf{W}_{st} = \begin{cases} \dfrac{1}{n_c}, & \text{if } \mathbf{y}^c_s = \mathbf{y}^c_t = 1, \\ 0, & \text{else,} \end{cases} \tag{6.8}$$

where $\mathbf{y}^c_s$ denotes the class label of datum point $\mathbf{X}_s$, n_d is the number of data points in all class and n_c is the number of data points in class c, where $c \in \{1, ..., C\}$.

By simultaneously maximizing the distances between data points from different classes and minimizing the distance between data points from the same class, the discriminant information is preserved at the greatest extent in the projected feature space. Optimizing $J(\mathbf{U},\mathbf{V})$ by solving **U** (or **V**) with fixed **V** (or **U**) is a convex optimization problem. Let $\mathbf{E}_{st} = \alpha \mathbf{B}_{st} - (1-\alpha)\mathbf{W}_{st}$, with the fixed **V**. The optimal **U** is composed of the first *P* eigenvectors of the following eigendecomposition problem:

$$\mathbf{D}_{\mathrm{V}} \mathbf{u} = \boldsymbol{\lambda} \mathbf{u} \tag{6.9}$$

where $\mathbf{D}_{\mathrm{V}} = \sum_{st} \mathbf{E}_{st} (\mathbf{X}_s .* \mathbf{R}^F_{st} - \mathbf{X}_t .* \mathbf{R}^F_{st}) \mathbf{V}\mathbf{V}^T (\mathbf{X}_s .* \mathbf{R}^F_{st} - \mathbf{X}_t .* \mathbf{R}^F_{st})^T$. Similarly, with the fixed **U**, the optimal **V** is composed of the first Q eigenvectors of the following eigendecomposition problem:

$$\mathbf{D}_{\mathrm{U}}\mathbf{v} = \boldsymbol{\lambda}\mathbf{v} \tag{6.10}$$

where $\mathbf{D}_{\mathrm{U}} = \sum_{st} \mathbf{E}_{st}(\mathbf{X}_s.*\mathbf{R}_{st}^F - \mathbf{X}_t.*\mathbf{R}_{st}^F)^T \mathbf{U}\mathbf{U}^T(\mathbf{X}_s.*\mathbf{R}_{st}^F - \mathbf{X}_t.*\mathbf{R}_{st}^F)$.

Therefore, we can alternately optimize **U** (with a fixed **V**) and **V** (with a fixed **U**). The above steps monotonically increase $J(\mathbf{U},\mathbf{V})$ and since the function is upper bounded, it will converge to a critical point with transformation matrices **U** and **V**.

In FBDP, the sizes of P and Q are determined by the number of positive eigenvalues in $\mathbf{D}_{\mathrm{V}}$ and $\mathbf{D}_{\mathrm{U}}$, respectively, since adding the eigenvectors corresponding to the nonpositive eigenvalues will not increase $J(\mathbf{U},\mathbf{V})$ in Eq. (6.6). As a result, the original dimension $I \times J$ is automatically reduced into $P \times Q$ after the FBDP procedure**.**

6.3.3 Greedy layer-wise reconstruction by semiconducting RBMs

In visual cortex, bi-directional inference includes bottom-up inference and top-down inference, and they are not separated processes. Therefore, in our model, bottom-up inference and top-down inference are integrated together to simulate the human visual perception. The whole deep learning model with the parameter space is constructed based on the bottom-up inference from available features and estimated features. Simultaneously, the estimated features with their reliability parameters are obtained by the top-down inference.

The incomplete data including the available features and estimated features are input to the deep architecture as the state of the input layer H^1 to construct an FRBM with the first hidden layer H^2. The energy function of the state ($\mathbf{h}^1,\mathbf{h}^2$) in the first Field effect RBM is shown in Eq. (6.11). Here, while the feature is available, the corresponding h^1_{ij} is the value of the available feature; if the feature is missing, the corresponding h^1_{ij} is the estimated value of the feature.

$$E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)=-\sum_{i=1,j=1}^{i\le I,j\le J}\sum_{p=1,q=1}^{p\le P^2,q\le Q^2} h^1_{ij}A^1_{ij,pq}R^{A,1}_{ij,pq}h^2_{pq}-\sum_{i=1,j=1}^{i\le I,j\le J} b^1_{ij}R^{b,1}_{ij}h^1_{ij}-\sum_{p=1,q=1}^{p\le P^2,q\le Q^2} c^1_{pq}h^2_{pq} \tag{6.11}$$

In Eq. (6.11), $I\times J$ is the number of units in H^1, while $P^2\times Q^2$ is the number of units in H^2. $\theta^1=\left(\mathbf{A}^1,\mathbf{b}^1,\mathbf{c}^1\right)$ are the model parameters between the input layer H^1 and first hidden layer H^2. $A^1_{ij,pq}$ is the symmetric interaction term between the input unit (i,j) in H^1 and the hidden unit (p,q) in H^2. b^1_{ij} is the $(i,j)^{th}$ bias of layer H^1 and c^1_{pq} is the $(p,q)^{th}$ bias of layer H^2. $R^{\theta,1}=\left(\mathbf{R}^{A,1},\mathbf{R}^{b,1}\right)$ are the reliability parameters between the input layer H^1 and the first layer H^2 to control the reliability of corresponding parameters θ^1. $R^{A,1}_{ij,pq}$ and $R^{b,1}_{ij}$ are the weights to control the reliability of corresponding parameters $(\boldsymbol{\theta})_{\mathbf{ij}}$ related to $(X_k)_{ij}$. To simplify the problem, reliability parameters $R^{\theta,1}_{ij}=\left(R^{A,1}_{ij,pq},R^{b,1}_{ij}\right)$ is depended on the reliability of the estimated value of missing feature $(X_k)_{ij}\in F_k$.

$$R^{A,1}_{ij,pq}=R^{A,1}_{ij,\bullet}=R^{b,1}_{ij}=\Re^{\theta,1}_{ij} \tag{6.12}$$

Therefore the first RBM has the following joint distribution:

$$P\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)=\frac{\exp^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)}}{\sum_{\mathbf{h}^1}\sum_{\mathbf{h}^2}\exp^{-E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)}} \tag{6.13}$$

The log probability of the model assigned to $\mathbf{h}^1$ in H^1 is:

$$\log P\left(\mathbf{h}^{1}\right)=\log\sum_{\mathbf{h}^{2}}\exp^{-E\left(\mathbf{h}^{1},\mathbf{h}^{2};\theta^{1}\right)}-\log\sum_{\mathbf{h}^{1}}\sum_{\mathbf{h}^{2}}\exp^{-E\left(\mathbf{h}^{1},\mathbf{h}^{2};\theta^{1}\right)} \qquad (6.14)$$

Similar with existing deep learning models, we utilize the stochastic steepest ascent in the log probability of the training data to update the parameter space $\theta^{1}=\left(\mathbf{A}^{1},\mathbf{b}^{1},\mathbf{c}^{1}\right)$.

$$A_{ij,pq}^{1}=\vartheta A_{ij,pq}^{1}+\triangle A_{ij,pq}^{1}R_{ij,pq}^{A,1} \qquad (6.15)$$

$$\triangle A_{ij,pq}^{1}=\varepsilon_{\mathbf{A}}(<h_{ij}^{1}(0)h_{pq}^{2}(0)>_{data}-<h_{ij}^{1}(1)h_{pq}^{2}(1)>_{recon}) \qquad (6.16)$$

where $\langle\cdot\rangle_{data}$ denotes an expectation with respect to the data distribution and $\langle\cdot\rangle_{recon}$ denotes the "reconstruction" distribution of data after one step. Other parameters in θ^{1} update function can be calculated in a similar manner.

$$b_{ij}^{1}=\vartheta b_{ij}^{1}+\triangle b_{ij}^{1}R_{ij}^{b,1}=\vartheta b_{ij}^{1}+\varepsilon_{\mathbf{b}}(h_{ij}^{1}(0)-h_{ij}^{1}(1))R_{ij}^{b,1} \qquad (6.17)$$

$$c_{pq}^{1}=\vartheta c_{pq}^{1}+\triangle c_{pq}^{1}=\vartheta c_{pq}^{1}+\varepsilon_{\mathbf{c}}(h_{pq}^{2}(0)-h_{pq}^{2}(1)) \qquad (6.18)$$

where ϑ is the momentum and $\varepsilon_{\mathbf{A}}$, $\varepsilon_{\mathbf{b}}$, $\varepsilon_{\mathbf{c}}$ are the learning rate of model parameters $\mathbf{A}$, $\mathbf{b}$ and $\mathbf{c}$.

As we described before, we find a Field effect bilinear projection based on the reliable features that can automatically reduce the original dimension $I\times J$ to $P\times Q$ through the transformation matrices $\mathbf{U}^{1}$ and $\mathbf{V}^{1}$. As a result, in our model, the number of neurons in layer H^{2} is determined by the row and column size of the transformation matrices $\mathbf{U}^{1}$ and $\mathbf{V}^{1}$.

$$P^{2}=row(\mathbf{U}^{1}),\ Q^{2}=column(\mathbf{V}^{1}) \qquad (6.19)$$

We set the discriminative transformation parameters obtained from the Field effect bilinear discriminant projection as the initial symmetrically connection weights by Eq. (6.20).

$$A^{1}_{ij,pq}(0) = (\mathbf{U}^{1}_{ip})^{T}\mathbf{V}^{1}_{jq} \tag{6.20}$$

The above discussion is the construction of the first Field effect RBM. Similar operations are performed on the higher layer pairs to construct the whole initial parameter space of the deep learning model.

The estimated feature of the missing feature is obtained by the top-down inference. To the incomplete sample datum $\mathbf{X}_s (1 \le s \le K)$, we define $(f_s)^{n}_{pq}$ to denote the corresponding activation code in the hidden unit (p,q) of the layer $n(1 \le n \le N)$. The activation code $(f_s)^{n}_{pq}$ is calculated by Eq. (6.21), where $\sigma(x)$ is the logistic function $\sigma(x) = 1/[1+\exp(-x)]$.

$$(f_s)^{n}_{pq} = h^{n}_{pq} = \sigma(h^{n-1}_{ij} A^{n-1}_{ij,pq} R^{A,n-1}_{ij,pq} + c^{n-1}_{pq}), \quad n \ge 2 \tag{6.21}$$

The Euclidean distances sequence $\{g^{n}_{s,t}\}$ between datum point $\mathbf{X}_s$ and $\mathbf{X}_t$ of layer *n* is denoted as below:

$$\{g^{n}_{s,t}\} = \left\| (f_s)^{n}_{pq} - (f_t)^{n}_{pq} \right\|, \quad 1 \le s,t \le K, s \ne t \tag{6.22}$$

To the current datum point $\mathbf{X}_s$, we sort the distances sequence $\{g^{n}_{s,t}\}$ in ascending order. The ranking position of the datum point $\mathbf{X}_t$ in the sorted list is denoted as $L^{n}_{s,t}$. To infer and estimate the missing features $(X_s)_{ij}$ in $\mathbf{X}_s$, the nearest datum point $\mathbf{X}_{t^*}$ is calculated and selected out as the reference datum of $\mathbf{X}_s$ by Eq. (6.23),

$$t^{*} = \arg\min_{t}[\sum_{n} \varepsilon_n L^{n}_{s,t}], \ s.t. \ (X_s)_{ij} \in F_s, (X_t)_{ij} \notin F_t, \mathbf{y}_s = \mathbf{y}_t \tag{6.23}$$

where ε_n is the weight of the activation codes in layer *n*. The higher layer activation of the reference datum is utilized to infer and estimate the missing features of incomplete datum just like Eq. (6.24). Let EF_s denote the set of estimated features of the sample $\mathbf{X}_s$.

$$(X_s)_{ij} = \sigma[A^1_{ij,pq}(f_{t^*})^1_{pq} R^{A,1}_{ij,pq} + b^1_{ij}],\ \ s.t.(X_s)_{ij} \in EF_s \tag{6.24}$$

In the bi-directional inference, the threshold voltage V_{th} in Eq. (6.5) is set to zero. Based on the estimated feature and the available features, the gate-to-source voltage V_{GS} in Eq. (6.5) is defined as $(z_s)_{ij}$ in Eq. (6.25),

$$\begin{aligned}
&(V_{GS})_{ij} = (z_s)_{ij} = \int_{(\mu)_{ij}+\|(X_s)_{ij}-(\mu)_{ij}\|}^{\infty} \frac{2}{\sqrt{2\pi}(\lambda)_{ij}} \exp^{-\frac{[y-(\mu)_{ij}]^2}{2(\lambda)_{ij}^2}} \mathrm{d}y, \\
&(\mu)_{ij} = \frac{1}{\sum_t t}\sum_t (X_t)_{ij},\ (\lambda)_{ij} = \sqrt{\frac{1}{\sum_t t}\sum_t [(X_t)_{ij} - (\mu)_{ij}]^2} \\
&s.t.\ (X_s)_{ij} \in EF_s, (X_t)_{ij} \notin F_t
\end{aligned} \tag{6.25}$$

where $(\mu)_{ij}$ and $(\lambda)_{ij}$ is the mean value and the standard deviation of all available features $(X_t)_{ij} \notin F(1 \le t \le K)$. $(z_s)_{ij}$ is the probability level of the estimated feature $(X_s)_{ij}$ according to the distribution of all available features. To estimated feature $(X_s)_{ij}$ in sample datum $\mathbf{X}_s(1 \le s \le K)$, the drain-to-source voltage V_{DS} in Eq. (6.5) is defined as $(m_s)_{ij}$.

$$(V_{DS})_{ij} = (m_s)_{ij} = \frac{2}{1+\exp(\sum_n \varepsilon_n L^n_{s,t^*}/8N)} \tag{6.26}$$

Therefore, the reliability parameter $\Re^{\theta,1}_{ij}$ of $(X_s)_{ij}$ could be written as (6.27).

$$\Re^{\theta,1}_{ij} = \begin{cases} 0, & \text{if } (z_s)_{ij} = 0, \\ \sqrt{2(z_s)_{ij}(m_s)_{ij} - (m_s)_{ij}^2}, & \text{if } (z_s)_{ij} > 0 \text{ and } (m_s)_{ij} < (z_s)_{ij}, \\ (z_s)_{ij}, & \text{if } (z_s)_{ij} > 0 \text{ and } (m_s)_{ij} \ge (z_s)_{ij} \end{cases} \tag{6.27}$$

6.3.4 Post activation by global fine-tuning

Above, we use the bi-directional inference algorithm by Field effect RBMs to construct a deep model. In this section, we use backpropagation through the whole deep model to fine-tune the parameters $\theta = [\mathbf{A}, \mathbf{b}, \mathbf{c}]$ for optimal recognition and estimation. Different with the usage of backpropagation in DBN, in our fine-tuning stage, the missing features in incomplete data are re-estimated.

In the layer-by-layer bi-directional inference stage, a global search has been performed for a sensible and good region in the whole parameter space. Therefore, before proceeding to the process of fine-tuning, we have already constructed a good data concept extraction model, and most of the missing features are roughly inferred and estimated. Now, backpropagation is utilized to tune the entire parameter space of FBDBN. Two tasks are involved in the stage of fine-tuning: finding good local optimum parameters $\theta^* = [\mathbf{A}^*, \mathbf{b}^*, \mathbf{c}^*]$ to recognize the data effectively, and adjusting the estimation values of missing features elaborately.

To the first task, the learning algorithm is used to minimize the recognition error $[-\sum_s \mathbf{y}_s \log \hat{\mathbf{y}}_s]$, where $\mathbf{y}_s$ and $\hat{\mathbf{y}}_s$ are the correct label and the output label value of sample datum $\mathbf{X}_s$. Simultaneously, to the second task, the roughly estimated missing features are slightly adjusted and updated by Eq. (6.28) and Eq. (6.29). Let OF_s denote the set of output estimated features of the missing features in sample datum $\mathbf{X}_s$.

$$t^* = \arg\min_t [\sum_n \varepsilon_n L^n_{s,t}],\ s.t. (X_s)_{ij} \in F_s, (X_t)_{ij} \notin F_t, \mathbf{y}_s = \mathbf{y}_t \tag{6.28}$$

$$(X_s)_{ij} = \sigma[A^1_{ij,pq} (f_{t^*})^1_{pq} R^{A,1}_{ij,pq} + b^1_{ij}],\ \ s.t. (X_s)_{ij} \in OF_s \tag{6.29}$$

For the test data, similar activation codes in the higher layer of the same predicted category are utilized to infer and estimate the value of the missing features. Eq. (6.28) is substituted by the equation below.

$$t^* = \arg\min_t [\sum_n \varepsilon_n L_{s,t}^n], \quad s.t. (X_s)_{ij} \in F_s, (X_t)_{ij} \notin F_t, \hat{\mathbf{y}}_s^c = \hat{\mathbf{y}}_t^c = 1 \qquad (6.30)$$

6.3.5 Algorithm and discussion

Firstly, the detailed procedure of the FBDBN is described in Algorithm 6.1. Then, in this section, we show the generalization of FBDBN by analyzing its relation with our previous work Bilinear deep belief networks (BDBN) (Zhong et al., 2011c) and the opposite extreme of BDBN, called SBDBN. Furthermore, we describe how to extend the supervised learning algorithm to semi-supervised and unsupervised learning scheme.

```
Input:  Training data set $X$, Corresponding labels set $Y$, Missing features set $F_k$
        Number of layers $N$, Number of epochs $M$
        Between-class weights $\mathbf{B}_{st}$,Within class weights $\mathbf{W}_{st}$
        Initial bias parameters b and c, Momentum $\vartheta$, Parameter $\alpha$
Output: Optimal parameter space $\theta^* = [\mathbf{A}^*, \mathbf{b}^*, \mathbf{c}^*]$,Estimated features $(X_s)_{ij} \in OF_s$
for $m = 1,\dots, M$ do
    for $n = 1,\dots, N$ do
        if $n = 1$   $T^n = X$   else
            for $k = 1,\dots, K$ do   $\mathbf{T}_k^n = \sigma(\mathbf{T}_k^{n-1} A^{n-1} + c^{n-1})$   end for
        end if
        while not convergent do
            $\mathbf{D}_{\mathrm{V}} = \sum_{st} \mathbf{E}_{st}(\mathbf{X}_s .* \mathbf{R}_{st}^F - \mathbf{X}_t .* \mathbf{R}_{st}^F)\mathbf{V}\mathbf{V}^T(\mathbf{X}_s .* \mathbf{R}_{st}^F - \mathbf{X}_t .* \mathbf{R}_{st}^F)^T$
            $\mathbf{D}_{\mathrm{U}} = \sum_{st} \mathbf{E}_{st}(\mathbf{X}_s .* \mathbf{R}_{st}^F - \mathbf{X}_t .* \mathbf{R}_{st}^F)^T \mathbf{U}\mathbf{U}^T(\mathbf{X}_s .* \mathbf{R}_{st}^F - \mathbf{X}_t .* \mathbf{R}_{st}^F)$
            Fix V/ U, compute U/ V by solving $\mathbf{D}_{\mathrm{V}}\mathbf{u} = \boldsymbol{\lambda}\mathbf{u} / \mathbf{D}_{\mathrm{U}}\mathbf{v} = \boldsymbol{\lambda}\mathbf{v}$
        end while
        Determine the size of next layer $P^{n+1} = row(\mathbf{U}^n)$, $Q^{n+1} = column(\mathbf{V}^n)$
        Compute initial connection weights $A_{ij,pq}^n(0) = (\mathbf{U}_{ip}^n)^T \mathbf{V}_{jq}^n$
        The energy in the current Field effect RBM
            $E(\mathbf{h}^1, \mathbf{h}^2; \theta^1) = -\sum_{i=1,j=1}^{i \le I, j \le J} \sum_{p=1,q=1}^{p \le P^2, q \le Q^2} h_{ij}^1 A_{ij,pq}^1 R_{ij,pq}^{A,1} h_{pq}^2 - \sum_{i=1,j=1}^{i \le I, j \le J} b_{ij}^1 R_{ij}^{b,1} h_{ij}^1 - \sum_{p=1,q=1}^{p \le P^2, q \le Q^2} c_{pq}^1 h_{pq}^2$
        Update the weights and biases
            $A_{ij,pq}^1 = \vartheta A_{ij,pq}^1 + \triangle A_{ij,pq}^1 R_{ij,pq}^{A,1}, b_{ij}^1 = \vartheta b_{ij}^1 + \triangle b_{ij}^1 R_{ij}^{b,1}, c_{pq}^n = \vartheta c_{pq}^n + \triangle c_{pq}^n$
    end for
    for $s = 1,\dots, K$ do
        Obtain the estimated feature of missing features
            $t^* = \arg\min_t [\sum_n \varepsilon_n L_{s,t}^n], s.t. (X_s)_{ij} \in F_s, (X_t)_{ij} \notin F_t, \mathbf{y}_s = \mathbf{y}_t$
            $(X_s)_{ij} = \sigma[A_{ij,pq}^1 (f_{t^*})_{pq}^1 R_{ij,pq}^{A,1} + b_{ij}^1], \ s.t.(X_s)_{ij} \in EF_s$
        Calculate gate-to-source, drain-to-source voltage
            $(V_{GS})_{ij} = (z_s)_{ij} = \int_{(\mu)_{ij} + \|(X_s)_{ij} - (\mu)_{ij}\|}^{\infty} \frac{2}{\sqrt{2\pi}(\lambda)_{ij}} \exp^{-\frac{[y-(\mu)_{ij}]^2}{2(\lambda)_{ij}^2}} \mathrm{d}y$,
            $(V_{DS})_{ij} = (m_s)_{ij} = \frac{2}{1 + \exp(\sum_n \varepsilon_n L_{s,t^*}^n / 8N)}$
        Update the reliability parameter of missing features
            $\Re_{ij}^{\theta,1} = \begin{cases} 0, & \text{if } (z_s)_{ij} = 0, \\ \sqrt{2(z_s)_{ij}(m_s)_{ij} - (m_s)_{ij}^2}, & \text{if } (z_s)_{ij} > 0 \text{ and } (m_s)_{ij} < (z_s)_{ij}, \\ (z_s)_{ij}, & \text{if } (z_s)_{ij} > 0 \text{ and } (m_s)_{ij} \ge (z_s)_{ij} \end{cases}$
    end for end for
Calculate optimal parameter space $\theta^* = \arg\min_\theta [-\sum_k \mathbf{y}_k \log \hat{\mathbf{y}}_k]$
Reesimate the missing features
    $t^* = \arg\min_t [\sum_n \varepsilon_n L_{s,t}^n], s.t.(X_s)_{ij} \in F_s, (X_t)_{ij} \notin F_t, \mathbf{y}_s = \mathbf{y}_t$
    $(X_s)_{ij} = \sigma[A_{ij,pq}^1 (f_{t^*})_{pq}^1 R_{ij,pq}^{A,1} + b_{ij}^1], \ s.t.(X_s)_{ij} \in OF_s$
```

Algorithm 6.1: Field effect bilinear deep belief networks

Relation with BDBN

Unlike FBDBN, in all three learning stages, BDBN (Zhong et al., 2011c) equally relies on all the features no matter whether the features are missing. It can be viewed as one special version of FBDBN that only includes the saturation mode. And the reliability is “frozen” to be one.

The objective function of bilinear discriminant initialization in BDBN can be expressed as:

$$\begin{aligned}&\arg\max_{\mathbf{U},\mathbf{V}} J(\mathbf{U},\mathbf{V})=\sum_{s,t=1}^{K}(\alpha\mathbf{B}_{st}-(1-\alpha)\mathbf{W}_{st})\|\mathbf{U}^{T}(\mathbf{X}_{s}.*\mathbf{R}_{st}^{F}-\mathbf{X}_{t}.*\mathbf{R}_{st}^{F})\mathbf{V}\|^{2}\\&s.t.\ \mathbf{R}_{st}^{F}=\mathbf{R}_{s}^{F}.*\mathbf{R}_{t}^{F},(R_{s}^{F})_{ij}=1,(R_{t}^{F})_{ij}=1,\mathbf{U}^{T}\mathbf{U}=\mathbf{I}_{P},\ \mathbf{V}^{T}\mathbf{V}=\mathbf{I}_{Q},\mathbf{X}_{s}\in X,\mathbf{X}_{t}\in X\end{aligned}\tag{6.31}$$

And the energy function of the state ($\mathbf{h}^{1},\mathbf{h}^{2}$) in the first RBM can be expressed as

$$\begin{aligned}&E\left(\mathbf{h}^{1},\mathbf{h}^{2};\theta^{1}\right)=-\sum_{i=1,j=1}^{i\le I,j\le J}\sum_{p=1,q=1}^{p\le P^{2},q\le Q^{2}}h_{ij}^{1}A_{ij,pq}^{1}R_{ij,pq}^{A,1}h_{pq}^{2}-\sum_{i=1,j=1}^{i\le I,j\le J}b_{ij}^{1}R_{ij}^{b,1}h_{ij}^{1}-\sum_{p=1,q=1}^{p\le P^{2},q\le Q^{2}}c_{pq}^{1}h_{pq}^{2}\\&s.t.\ R_{ij,pq}^{A,1}=R_{ij,\bullet}^{A,1}=R_{ij}^{b,1}=1\end{aligned}\tag{6.32}$$

Relation with SBDBN

There is another special version of FBDBN, which is the opposite extreme of BDBN. We denote it as Semiconducting deep belief networks (SBDBN). Different with FBDBN, SBDBN not adaptively adjust the reliability of the estimated values for missing features. SBDBN only relies on fully exploiting the embedding information according to the available features rather than any completion of missing features. That is to say, there are only two modes in the SBDBN: cutoff mode and saturation mode. To the available features, the reliability connections are set to be one, and the reliability connections of the missing features are set to be zero.

The energy function of the state ($\mathbf{h}^{1},\mathbf{h}^{2}$) in the first RBM can be expressed as:

$$
\begin{aligned}
&E\left(\mathbf{h}^1,\mathbf{h}^2;\theta^1\right)=-\sum_{i=1,j=1}^{i\le I,j\le J}\sum_{p=1,q=1}^{p\le P^2,q\le Q^2} h_{ij}^1 A_{ij,pq}^1 R_{ij,pq}^{A,1} h_{pq}^2-\sum_{i=1,j=1}^{i\le I,j\le J} b_{ij}^1 R_{ij}^{b,1} h_{ij}^1-\sum_{p=1,q=1}^{p\le P^2,q\le Q^2} c_{pq}^1 h_{pq}^2 \\
&s.t.\ R_{ij,pq}^{A,1}=R_{ij,\bullet}^{A,1}=R_{ij}^{b,1}=\begin{cases}0, & \text{if } (X_k)_{ij}\in F_k \\ 1, & \text{else}\end{cases}
\end{aligned}
\tag{6.33}
$$

Semi-supervised learning scheme

Unlike the supervised learning, in semi-supervised learning, only the labeled samples X^L in training dataset *X* is utilized in field effect bilinear discriminant initialization. The objective function of FBDP can be represented as follows:

$$
\begin{aligned}
&\arg\max_{\mathbf{U},\mathbf{V}} J(\mathbf{U},\mathbf{V})=\sum_{s,t=1}^{K}(\alpha\mathbf{B}_{st}-(1-\alpha)\mathbf{W}_{st})\|\mathbf{U}^T(\mathbf{X}_s.*\mathbf{R}_{st}^F-\mathbf{X}_t.*\mathbf{R}_{st}^F)\mathbf{V}\|^2 \\
&s.t.\mathbf{R}_{st}^F=\mathbf{R}_s^F.*\mathbf{R}_t^F,(R_s^F)_{ij}=1,(R_t^F)_{ij}=1,\mathbf{U}^T\mathbf{U}=\mathbf{I}_P,\mathbf{V}^T\mathbf{V}=\mathbf{I}_Q,\mathbf{X}_s\in X^L,\mathbf{X}_t\in X^L
\end{aligned}
\tag{6.34}
$$

To obtain the estimated feature of missing features, the nearest datum point $\mathbf{X}_{t^*}$ is calculated and selected out as the reference datum of $\mathbf{X}_s$ only depended on the similarity of the activations in higher layers as follows.

$$
t^*=\arg\min_t[\sum_n \varepsilon_n L_{s,t}^n],\ s.t.\ (X_s)_{ij}\in F_s,(X_t)_{ij}\notin F_t \tag{6.35}
$$

Unsupervised learning scheme

Different with the supervised and semi-supervised learning schemes, the unsupervised learning scheme does not include Field effect bilinear discriminant initialization. The number of neurons in each layer is fixed and pre-defined intuitively.

$$
\begin{aligned}
&P^n=\text{Row size of layer } n\ , \\
&Q^n=\text{Column size of layer } n.
\end{aligned}
\tag{6.36}
$$

The initial connection parameter is initialized to small random values chosen from a zero-mean Gaussian with a standard deviation of 0.1,

$$
A_{ij,pq}^n(0)=0.1\times\mathrm{RAND}(1) \tag{6.37}
$$

where RAND(1) returns one random value drawn from the standard normal distribution.

In the fine-tuning stage, the recognition error will be substituted by the reconstruction discrepancy.

$$\theta^* = \arg\min_{\theta} [-\sum_{k} X_k \log \hat{X}_k] \tag{6.38}$$

6.4 Performance Evaluation

In this section, we demonstrate the performance of the proposed FBDBN on three standard datasets under three learning schemes: supervised, semi-supervised and unsupervised learning. The first dataset is MNIST, a standard large database of hand written digits, which is always used to illustrate the performance of deep models, and its subset has been used for performance comparison of incomplete data classification algorithms (LeCun et al., 1998). The second standard dataset is BioID face dataset, which is always used to illustrate the performance of face recognition. BioID consists of 1521 face images collected contains 23 subjects under a large variety of background and illumination (Jesorsky et al., 2001). We remove some parts of the face, such as eyes and mouth on purpose to show the performance for incomplete data. As shown in Figure 6.1 (b), the more general case of incomplete data in our daily life is that some key features in the data are not observable. So our group collected some incomplete face images due to the occlusion in the important facial feature regions and constructed a new dataset called StarFace, including 120 face images from four superstars and other 5000 face images from unknown persons downloaded from Google.

For the parameters used in the experiments, we follow the general setting of the previous work in deep learning. For example, the balance weight α is set as 0.5.

The weight ε_n which is the activation codes in layer n is set as 1. In greedy layer wise learning, the number of epochs M is fixed at 50 and the learning rate η is equal to 0.1. The initial momentum ϑ is set as 0.5. In the fine-tuning stage, the method of conjugate gradients is utilized and three line searches are performed in each epoch until convergence.

We compare the performance of FBDBN with various state-of-the-art incomplete image recognition algorithms and representative deep learning models, including *k*-nearest neighbor estimation (*k*-NNE), Support vector machines (SVM) (Boser et al., 1992), LRCEM (Williams et al., 2005), Maximize geometric margin (GEOM) (Chechik et al., 2008), Quadratically gated mixture of experts (QGME) (Liao et al., 2007), Weighted infinite imputations (WII) (Dick et al., 2008), Deep belief networks (DBN) (Hinton & Salakhutdinov, 2006), Bilinear deep belief networks (BDBN) (Zhong et al., 2011c), and Semiconducting deep belief networks (SBDBN). In *k*-NNE, the missing features were set with the mean value obtained from the nearest neighbors' instances. Neighborhood was measured using a Euclidean distance in the subspace relevant to each pair of samples. The number of neighbors was varied across 1, 3, 5, 10, 15, 20, and the best result is provided to make comparison. In LRCEM, a Gaussian mixture model is learned by iterating between (1) learning a GMM model of the filled data and (2) re-filling missing values using cluster means, weighted by the posterior probability that a cluster generated the sample. The number of clusters was varied across 1, 3, 5, 10, 15, 20, and the best result is reported.

6.4.1 Experiment on MNIST

In this section, we explore the performance of FBDBN under supervised learning scheme when features are missing at random. We test on the image dataset of handwritten digits MNIST (LeCun et al., 1998). MNIST is a standard large database of hand written digits containing 70,000 images with 10 classes. MNIST is often used to compare deep learning performance (Salakhutdinov & Hinton, 2007) (Weston et al., 2008).

The first experiment in this dataset is used to demonstrate the effectiveness of FBDBN for recognition on incomplete images with fixed missing ratio. We follow the same experimental setting of (Chechik et al., 2008). 1200 images including 600 images of the digits 5 and 600 images of digit 6 are randomly selected from MNIST. These images are partitioned to 1000 training data and 200 test data. We removed a square patch of pixels from each image that covered 25% of the total number of pixels. The location of the patch was uniformly sampled for each image, and typical examples are given in Figure 6.4.

We perform 5 random splits and report the average results over the 5 trials. The recognition performance of FBDBN with other incomplete image recognition algorithms is shown in Table 6.1. “Zero” means that the missing values were set to zero. “Mean” means that the missing values were set to the average value of the feature over all data. From Table 6.1, it can easily be observed that, compared with state-of-the-art incomplete image recognition algorithms, the deep learning models achieve better performance. This proves that the deep learning models have better recognition ability on incomplete data. Compared with two special versions of FBDBN, by fully exploiting the embedding information according to the available features, the recognition ability of SBDBN is better than BDBN. Furthermore, relying on the Field effect bilinear discriminant initialization and Field effect RBMs, FBDBN obtains the best accuracy rate in all deep models. In Figure 6.5, some samples of the block incomplete images and the corresponding estimated images are demonstrated. Although some occluded blocks located in the important parts that make digits 5 and 6 are similar with each other, the estimated images obtained by FBDBN are correct.

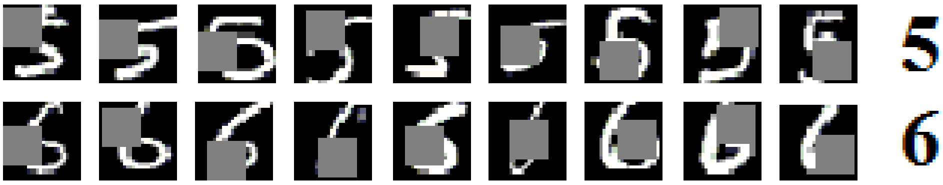


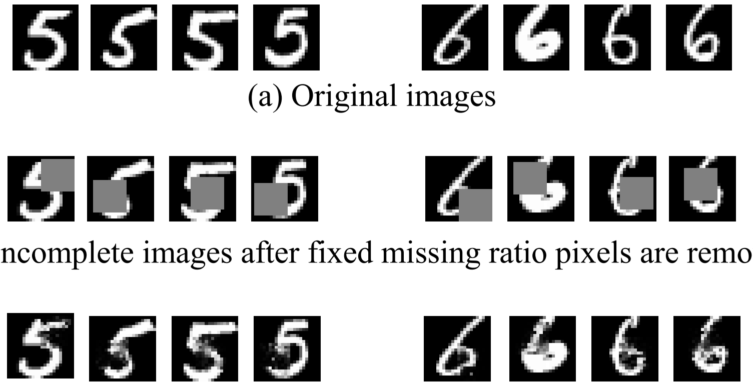
Figure 6.4. Examples of MNIST images of the digits ‘5’ and ‘6’ after fixed missing ratio pixels are removed with random centers.

Table 6.1. Incomplete image recognition accuracy rate (%) on test data.

Model			Acc.
Proposed model		**FBDBN**	**99**
		SBDBN	98.5
Other deep models	Bilinear deep model	BDBN(Zero)	97
		BDBN(Mean)	97.5
	Classical deep model	DBN(Zero)	96
		DBN(Mean)	96.5
Representative model for incomplete data	Without estimation	GEOM	95
	With estimation	SVM (Zero)	95
		SVM (Mean)	95
		k-NNE	94
		LRCEM	95
	Joint optimization	QGME	96.5
		WII	96

In the second experiment, we demonstrate the incomplete image recognition when features are missing at random under different missing ratio. 10,000 images of 10 classes from MNIST are utilized as the training data, and the remaining 60,000 images are utilized for test. 5 random missing trails are performed and the average results over the 5 trials are reported. Some sample images with different missing ratios are shown in Figure 6.6. Although the image samples selected in Figure 6.6 are not difficult to recognize, when the missing ratio becomes higher, these handwritten digits are not easily recognized even by human.

Table 6.2 shows the performance comparison under different missing ratios. Obviously, FBDBN shows higher incomplete image recognition accuracy rate. When 80% features are missing, although the recognition by human is adequately hard, our algorithms also demonstrate the acceptable performance.



(a) Original images

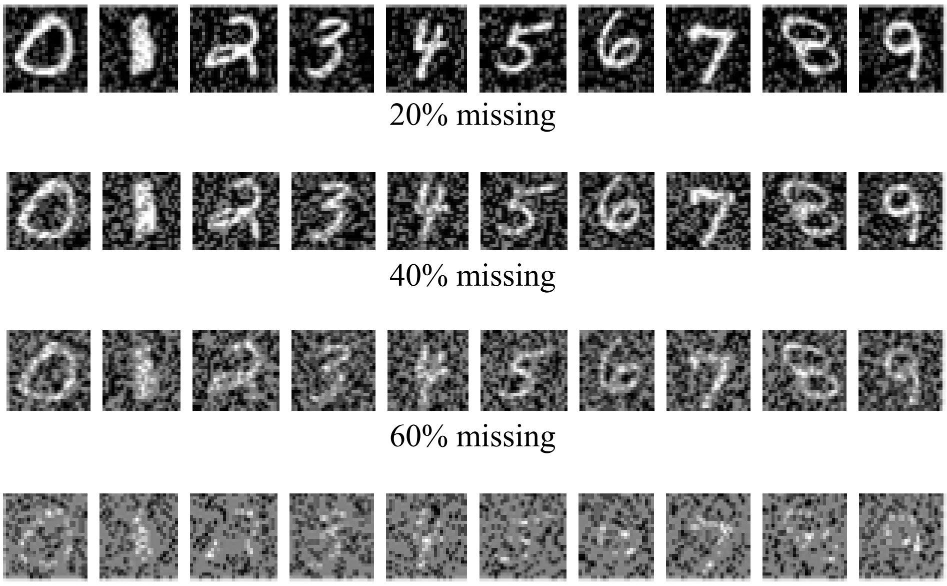
(b) Incomplete images after fixed missing ratio pixels are removed

(c) Estimated images via FBDBN.

Figure 6.5. Samples of estimated images by FBDBN of the block missing features with fixed missing ratio.

Table 6.2. Recognition accuracy rate (%) on test data with different missing ratios.

Missing Ratio	20%	40%	60%	80%
FBDBN	**97.12**	**95.88**	**92.96**	**79.23**
SBDBN	96.04	95.29	92.41	78.78
DBN(Zero)	95.99	93.62	89.70	77.68
DBN(Mean)	95.80	93.75	91.62	76.85
SVM(Zero)	84.84	81.10	77.96	60.47
SVM(Mean)	84.25	83.52	79.70	73.69



20% missing

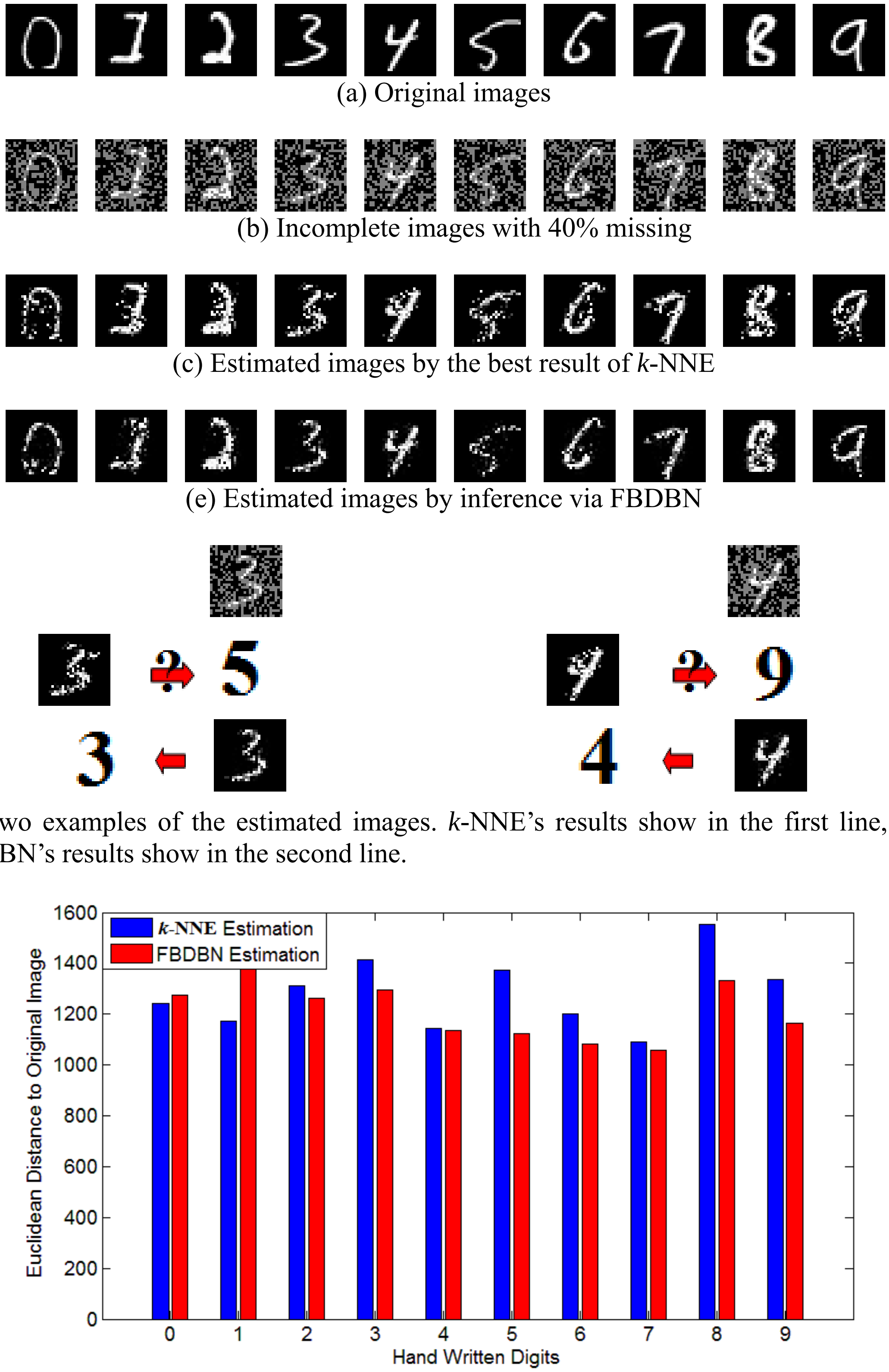
40% missing

60% missing

80% missing

Figure 6.6. Examples of images for different percent of pixels missing randomly.

Additionally, to evaluate whether FBDBN has the ability to estimate the missing features, we compare our algorithm with the baseline estimation algorithm *k*-NNE. In *k*-NNE, the number of neighbors was varied across 1, 3, 5, 10, 15, 20, and the result images with the shortest Euclidean distance from the original one are selected. Some samples of examples are demonstrated in Figure 6.7. From the Figure 6.7, we could find some hand written digits are incorrectly estimated by *k*-NNE, for example: digit 3 is restored just as digit 5, digit 4 is estimated just as digit 9, digit 6 is filled just like digit 0, and digit 7 is restored just as digit 9. And we list two comparisons in Figure 6.7 (f). Although the distance between the original one and estimated one by both of the algorithms is not too far in Figure 6.7 (g), the estimated images by *k*-NNE do not have enough discriminant information. In *k*-NNE, the estimations are relied on the similarity of pixel-level features. But in our FBDBN, the Field effect RBMs help us to infer the missing features with better discriminant ability. This strategy guarantees FBDBN to estimate the incomplete images successfully and effectively.



(a) Original images

(b) Incomplete images with 40% missing

(c) Estimated images by the best result of *k*-NNE

(e) Estimated images by inference via FBDBN

(f) Two examples of the estimated images. *k*-NNE's results show in the first line, FBDBN's results show in the second line.

(g)Euclidean distance comparison from the estimated images with the original images.

Figure 6.7. Samples of estimated images with 40% randomly missing ratio by *k*-NNE and FBDBN.

6.4.2 Experiment on BioID

In this section, we explore performance of FBDBN for face recognition on dataset of BioID (Jesorsky et al., 2001) when important facial features are missing under semi-supervised and unsupervised learning schemes.

BioID face dataset consists of 1521 face images collected contains 23 subjects. The number of images in every category of BioID is varied, from 35 to 118. In our experiments, firstly, we select the categories with more than 50 face images as the subset which we work on. This subset includes 1208 images in 14 categories. Then, just like the procedure on face datasets, the original images are normalized (in scale and orientation) so that the two eyes are aligned at the same position. Finally, the facial areas are cropped and downsampled into the final images. The size of each final image in all of the experiments is 32×32 pixels, with 256 gray levels per pixel. Some sample images after preprocessing are shown in Figure 6.8.

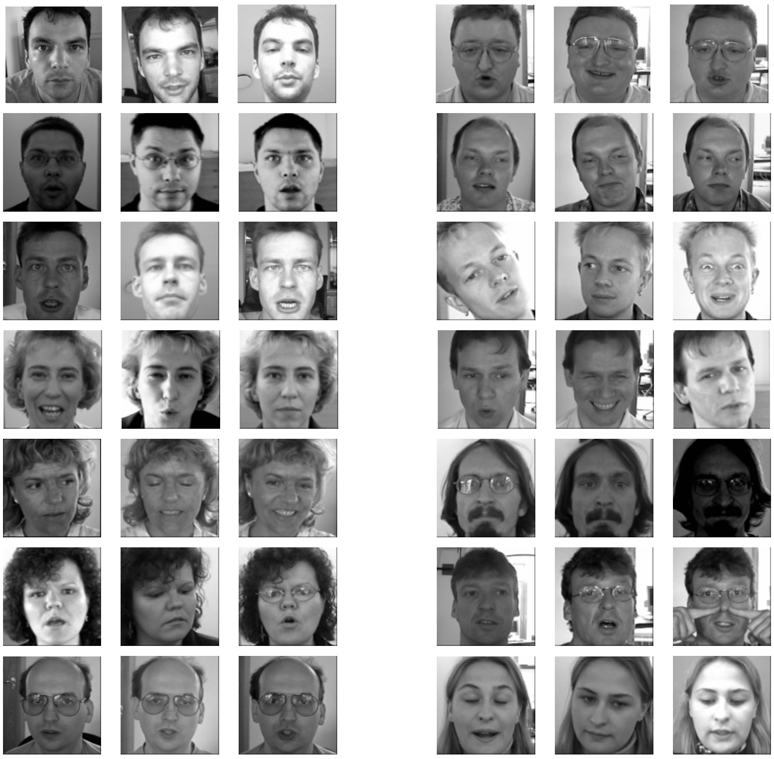


Figure 6.8. Some sample images in BioID.

The experiment in this dataset is used to demonstrate the face recognition effectiveness of FBDBN when important facial features are missing. To every image, we removed a rectangle region of pixels and generate five kinds of facial regions missing images. The locations of missing regions are related with important facial feature regions, including forehead, eye, nose, mouth, and chin. Sample images with important facial regions missing are given in Figure 6.9.

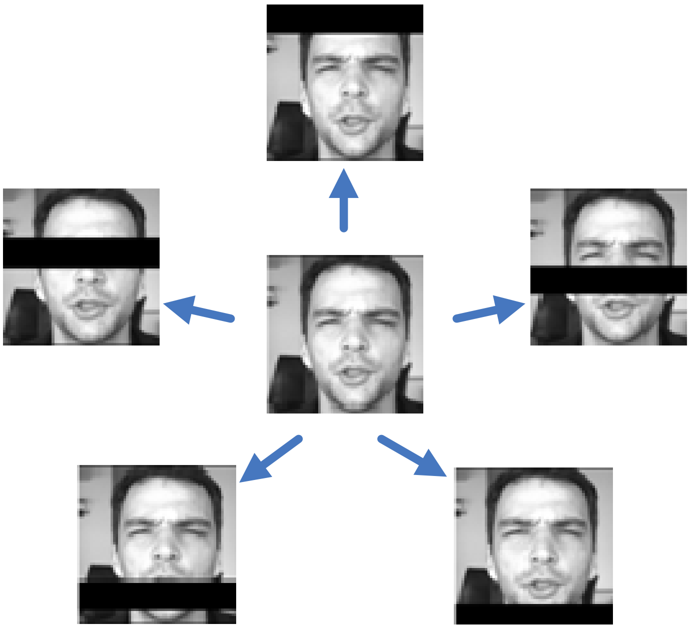


Figure 6.9. Sample images with important facial regions missing.

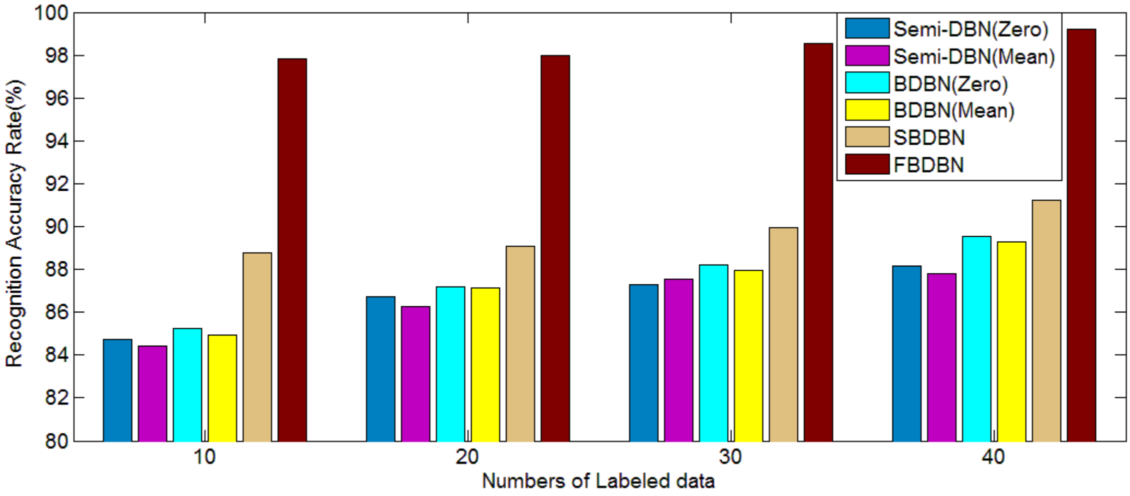
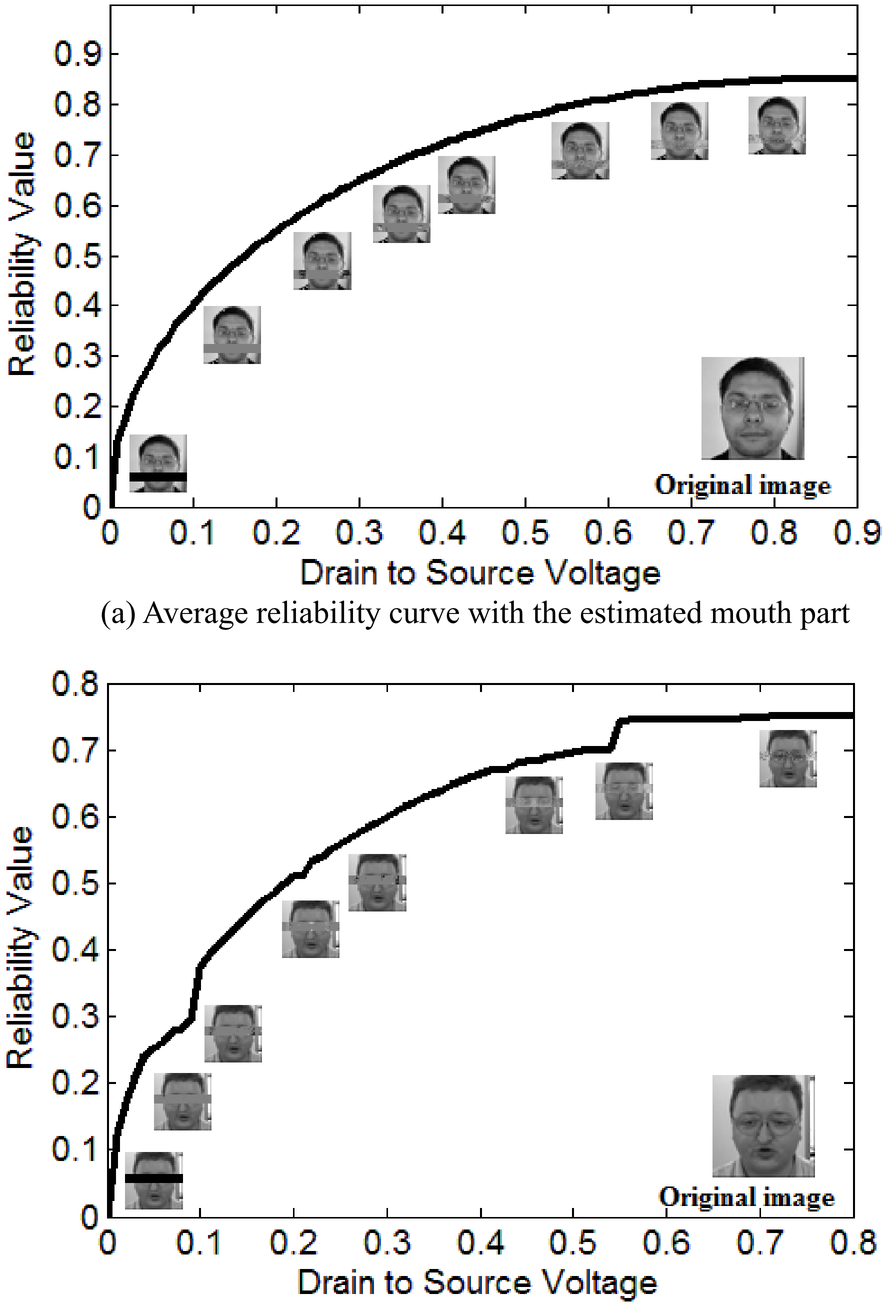


Figure 6.10. Recognition accuracy rate on test data in BioID with different numbers of labeled data.

In the above experiments, deep learning models demonstrated a better performance than other existing recognition models. Therefore, in this experiment, we compare proposed FBDBN with other deep learning models under semi-supervised learning scheme. For this dataset, 250 images for each person with different missing regions are randomly selected to form the training set and the rest 2540 images are utilized to form the test set. Different numbers of images of training data are randomly selected and labeled while the other training data remain unlabeled. The number of selected labeled data in each category is equal to 10, 20, 30, and 40, respectively. We perform 5 random splits and report the average results over the 5 trials. Figure 6.10 shows the face recognition accuracy rate of the test dataset. Although the recognition accuracy of Semi-DBN and BDBN are both higher than 80%, the recognition accuracy of FBDBN is the highest, which is nearly 100%. This phenomenon is due to Semi-DBN and BDBN equally trust on the available features and forecasting unreliable features. The missing features located in the important facial regions will influence the recognition accuracy. Two examples of the average reliability curve with the estimated image are shown in Figure 6.11. With the aid of the bi-directional inference by Field effect RBMs, the reliability of the estimated features is automatically and adaptively adjusted. With the increase of the reliability, the estimated images are more and more similar with the original image without missing features.

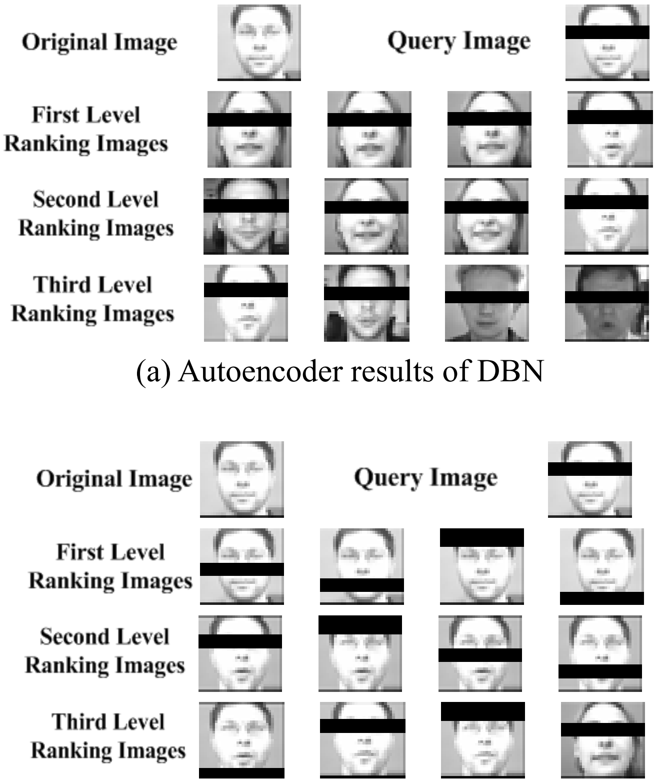


(a) Average reliability curve with the estimated mouth part

(b) Average reliability curve with the estimated eyes part

Figure 6.11. The reliability curve with estimated images.

In the second experiment of this dataset, we verify the auto-encode ability of proposed FBDBN via unsupervised learning. In this experiment, we follow the parameter setting of the DBN encoder in (Hinton & Salakhutdinov 2006). We use a 1024-1000-500-250-30 numbers of nodes in FBDBN. All units are logistic except for the 30 linear units in the code layer. To evaluate the encode ability of FBDBN, we compare the ranking of images based on the distance of the low-dimension output codes in FBDBN and DBN (Zero) in Figure 6.12. From Figure 6.12 (a), it is obvious that all the output codes with shorter distances are the faces missing eyes parts. And most of them are not the images of the identical person. In contrary to DBN, the output codes in the high ranking level are the images of the same person with same expression but missing different regions. With the help of FRBM and missing feature estimation in FBDBN, the missing features will not have an obvious influence to the low-dimension output codes. It proves that FBDBN is an effective tool to reduce the dimensionality of the incomplete data, which is very useful in the classical task in computer vision and multimedia: image retrieval.


Original Image
Query Image
First Level Ranking Images
Second Level Ranking Images
Third Level Ranking Images
Original Image
Query Image
First Level Ranking Images
Second Level Ranking Images
Third Level Ranking Images

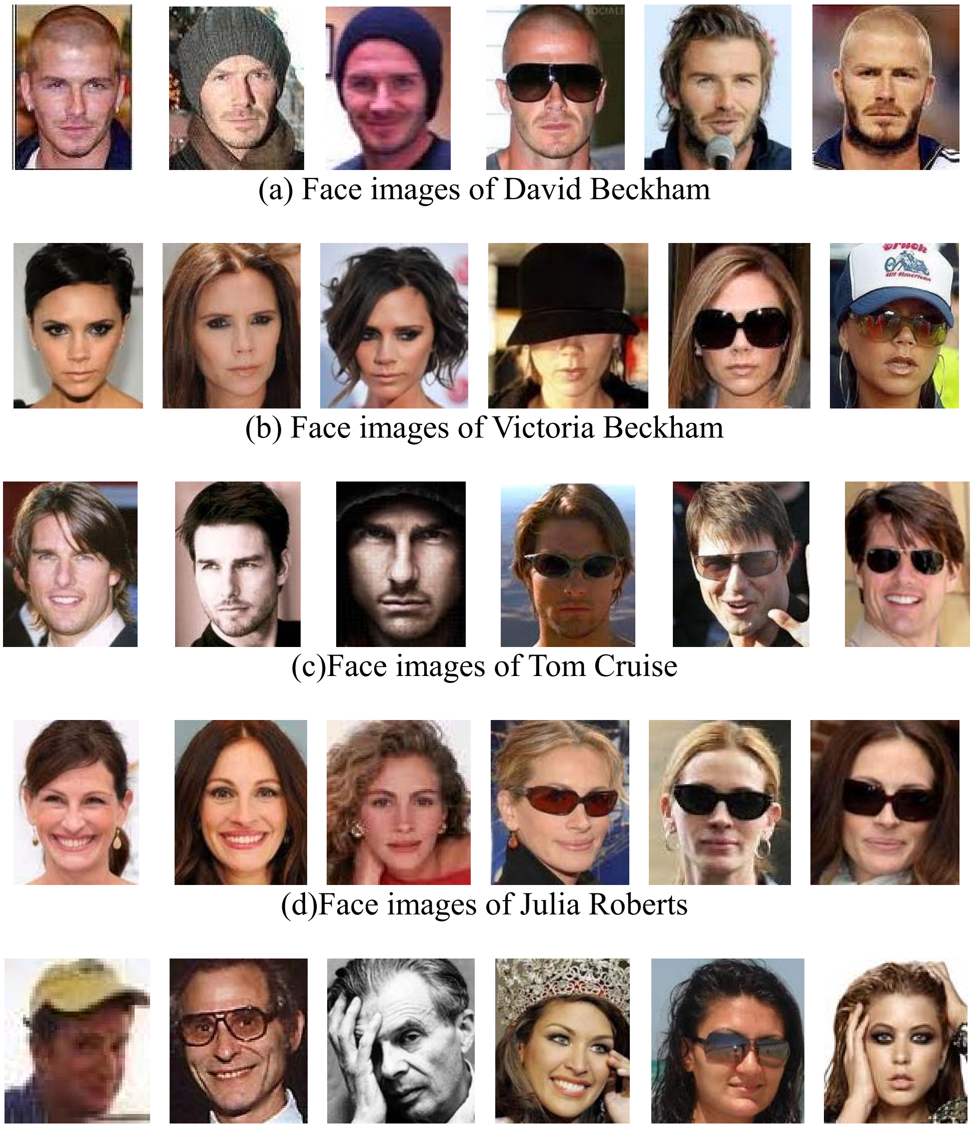
(a) Autoencoder results of DBN

(b) Autoencoder results of FBDBN

Figure 6.12. Autoencoder comparison of DBN and FBDBN.

6.4.3 Experiment on StarFace

To further prove the effectiveness of proposed FBDBN in real natural images, we collect and construct a new dataset StarFace[1] from Google, including 120 face images of David Beckham, Victoria Beckham, Tom Cruise, and Julia Roberts and 5000 face images from some unknown persons.



(a) Face images of David Beckham

(b) Face images of Victoria Beckham

(c)Face images of Tom Cruise

(d)Face images of Julia Roberts

(e)Face images of other persons

Figure 6.13. Sample images of StarFace.

[1] The dataset has been made publicly available and can be downloaded from http://www4.comp.polyu.edu.hk/~csshzhong/Star_Face.zip.

Figure 6.13 shows some samples images utilized in the retrieval experiment. From these sample images, it is obviously that some important facial feature regions have been occluded. To every occlusion region, we mark them as missing feature regions in preprocessing stage.

We evaluate the proposed FBDBN with proposed SBDBN and DBN for face retrieval via unsupervised learning. To every category, we randomly select one image with a kind of important facial region missing as the query image. We perform 5 random splits and report the average results over the 5 trials. The mean value of Normalized discounted cumulative gain (NDCG) is utilized to evaluate the retrieval results. From the NDCG scores in Table 6.3, the FBDBN has better retrieval performance. From Figure 6.14, it is obvious that our algorithm is effective in image retrieval although some important facial features are missing.

Table 6.3. Comparison of NDCG scores on StarFace.

	FBDBN	SBDBN	DBN(Zero)	DBN(Mean)
NDCG@10	**0.4904**	0.4208	0.3916	0.3787
NDCG@20	**0.4135**	0.3761	0.3289	0.3178

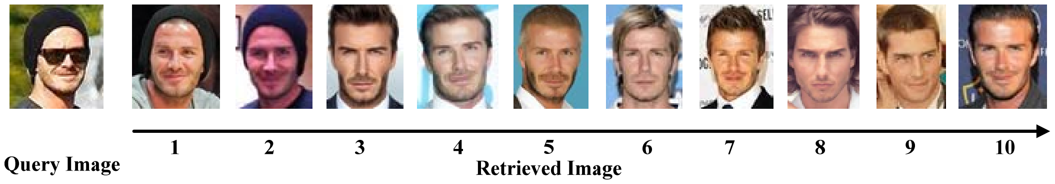


Figure 6.14. A query image with first ten images which are retrieved out.

6.5 Summary

In this chapter, we propose a novel deep learning model, FBDBN for image recognition with incomplete data. FBDBN has several attractive characters. First, incomplete image recognition is a classical challenge in computer vision and machine learning. Proposed FBDBN is the first deep learning model developed specially for this problem. Inherited from the merits of deep architectures, FBDBN simulates the laminar structure of human's cerebral cortex and the information delivery between multiple layers. Second, the three-stage learning of FBDBN faithfully realizes the human's object recognition procedure, including: "initial guess", "bi-directional inference" and "post activation". According to simulate the characters in human's recognition by Field effect bilinear discriminant projection and Field effect RBMs, the proposed algorithm with the adaptive reliability function is proved to be very competent in the task of incomplete image recognition. Third, in FBDBN, we propose a unified learning paradigm that integrates the merits of supervised learning, semi-supervised learning and unsupervised learning schemes into a uniform framework. Thus, it is facilitating for different kinds of data for different tasks.

In our experiments on real-world image recognition and retrieval tasks, FBDBN shows the distinguishing and robust recognition ability on the incomplete data. Furthermore, the missing features in the incomplete images are inferred and estimated effectively.

Chapter 7 Conclusion and Future Work

This dissertation presents a series of our studies for multimedia content analysis via computational human visual model. In Chapter 7, we first summarize and conclude the proposed work in this thesis. Then, the future work is discussed.

7.1 Conclusion

The most critical problem of multimedia content analysis is how to detect, classify and recognize the content of the multimedia data as if being understood by the computer. Despite more than twenty years of extensive research, multimedia content analysis for real-world applications remains a well-known challenge in the field of multimedia and computer vision. Due to the human-machine gap in multimedia content analysis, more and more researchers focused on constructing the computational human visual models to imitate human perception and intelligence. The primary focus of this thesis is seeking to understand the computational underpinnings of human visual processing through concerted efforts in both reverse- and forward-engineering based on the given tasks of multimedia content analysis. This thesis explores computational human visual modeling techniques from four aspects:

1) Retinal image formation for object detection and recognition. The first step of human visual system is to format the image on the retinal surface. Obviously, the characters of retinal image formation will influence the multimedia content analysis. Because human eyes expose pictures around 30 times per second, the relative motion

between the sensor and the scene in this exposure time will lead to motion blur. As a well known degradation factor, motion blur due to the water wave is too large to cause the distortion of the water part in water reflection image. Furthermore, motion blur leads to the existing reflection symmetry detection and recognition techniques invalid. To address the difficulty in water reflection recognition, a novel Invariant moment & Curvelet coefficient (IMCC) feature space is constructed. Moreover, based on the feature space, we propose two effective sub-algorithms to recognize the water reflection image, including: Reflection cost minimization (RCM) and High-frequency Curvelet coefficients discrimination (HCCD). By experimenting on real image datasets for different tasks, the proposed feature space and algorithms demonstrate impressive results in the water reflection image classification, the reflection axis detection, and the retrieval of the images with water reflection.

2) Attention allocation for image saliency detection and quality assessment. As a cognitive process of selectively concentrating on one aspect of the environment while ignoring other things, attention has been referred to as the allocation of processing resources. Construction of attention model in multimedia data is useful for applications in multimedia like object segmentation, object recognition and quality assessment. Our novel attention model integrates bottom-up saliency features and top-down targets information together. In bottom-up direction, the proposed Gabor & Curvelets based saliency map relying on 2D Gabor and Curvelet transforms with better representation ability of spatial and directional information. In top-down direction, the proposed Target based saliency map relying on the center priority and semantic target location obtained based on the tag information. Empirical validations on standard dataset demonstrate the effectiveness of the bottom-up saliency detection, and top-down saliency detection. Furthermore, in the image quality assessment task,

our technique based on Bi-directional saliency map (BSMP) outperforms the representative blurriness metrics.

3) Perceptual modeling for image annotation. Contextual cueing is the perceptual processing in which the human brain gathers information from visual elements and their surroundings, which acquires incidentally learn from past experiences of regularities. Five types of spatial invariants are thought to be important in contextual cueing, including: probability, co-occurrence, size, position and spatial topological. We propose a novel label to region assignment (LRA) technique called Fuzzy-based contextual-cueing label propagation (FCLP) to address the challenging problem in region level annotation to improve the semantic understanding of the images. Fuzzy representation and fuzzy reasoning are utilized to describe the spatial invariants in contextual cueing and imitate the contextual cueing process. The experiments on two public datasets demonstrate that the proposed technique achieves obvious performance improvement of label to region assignment for the images with multiple objects and complex background.

4) Visual cortex simulation for multimedia content analysis. The visual cortex of the brain is the most important part in the human visual system for processing visual information. Deep architecture composed of multiple layers of parameterized nonlinear modules is a representative paradigm that has achieved notable success in modeling the human visual system. By referencing the architecture of the visual cortex and the procedure of perception, we construct two novel deep networks model BDBN, and FBDBN for the two classical and intelligent tasks in multimedia content analysis: image classification, and incomplete data recognition. Extensive experiments on various standard datasets not only show the distinguishing ability of our model in various tasks but also clearly demonstrate our intention of providing a

human-like image analysis by referencing the human visual system and perception procedure.

7.2 Future work

Though we have made some progress in computational human visual modeling for multimedia content analysis tasks, it is far from the end of this road. A lot of work can be proposed to make further improvement in multimedia content analysis. In this section, we would like to provide some possible future work.

7.2.1 Deep learning model for video data analysis

The conventional deep learning model unfolds the input visual data to vectors. The proposed BDBN and FBDBN are constructed by a set of second-order planes, which are also consistent with the natural tensor structure of images. However, image is one kind of multimedia data. When we face to the real-world video data, the process would be much more complicated. The difficulty of video analysis is how to analysis the space-series information and the time-series information together. Different from image data, the order of natural structure in video data is generally not smaller than three. Furthermore, as we known, the text and audio information in video are also very helpful and important for the video analysis. From this point of view, the video data is not just fifth-order or sixth-order. Therefore, it is necessary to construct novel deep learning model to learn the video structure and analyze the content.

7.2.2 General deep learning model by simulating human visual system

The proposed computational models are constructed via simulating different parts of human visual system, such as retinal image formation, visual cortex simulation, attention allocation, and contextual cueing. In these models, the deep learning models BDBN and FBDBN demonstrate distinguished performance in visual data analysis by referring the visual cortex, which is the most important part of human visual system. In future work,we intend to propose a general deep learning model by simulating all four parts in human visual system. For example, we aim to work on how to integrate the contextual cueing in deep learning for intelligent understanding of images.

Reference

Abas, K.H. & Ono, O.. 2010. Implementation of multi-centroid moment invariants in thermal-based face identification system. In *American Journal of Applied Science,* 7 (3), ISSN 1546-9239, pages 283 – 289.

Achanta, R., Hemami,S., Estrada, F. and Süsstrunk, S.. 2009. Frequency-tuned salient region detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), pages 1597-1604.

Aleman, A., Bocker, K.B., Hijman,R, De Haan, E.H., Kahn, R.S.. 2003. Cognitive basis of hallucinations in schizophrenia: Role of top-down information processing. In *Schizophrenia Research*, 64(2-3), pages178-185.

Anderson, J. R.. 2004. Cognitive psychology and its implications, 6th Edition, Worth Publishers.

Attneave, F.. 1995. Symmetry information and memory for patterns. In *American Journal of Psychology*, 68, pages 209-222.

Avidan, S. & Shamir, A.. 2007. Seam carving for content-aware image resizing. In *ACM Transactions on Graphics*, pages 1-10.

Ballan, L., Bazzica, A., Bertini, M., Bimbo, A. D., Serra, G.. 2009. Deep networks for audio event classification in soccer videos. In Proceedings of the 2009 IEEE International Conference on Multimedia and Expo (ICME 2009), pages 474-477.

Bar, M.. 2004. Visual objects in context. In *Nature Reviews Neuroscience*, 5, pages 617-629.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.. 2006. Greedy layer-wise training of deep networks. In Proceedings of the 19th Annual Conference on Neural Information Processing Systems (NIPS 2006), pages 153-160.

Bengio, Y. & LeCun, Y.. 2007. Scaling learning algorithms towards AI. In *Large-Scale Kernel Machines*, MIT Press.

Biederman, I., Rabinowitz, J.C., Glass, A.L., Stacy E.J.. 1974. On the information extracted from a glance at a scene. In *Journal of Experimental Psychology*, 103,

pages 597-600.

Biederman, I., Mezzanotte, R.J., and Rabinowitz, J.C.. 1982. Scene perception: detecting and judging objects undergoing relational violations, In *Cognitive Psychology*, 14(2), pages 143-177.

Boiman,O., Shechtman, E., Irani, M.. 2008. In defense of nearest-neighbor based image classification. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008).

Bosch, A., Zisserman, A., Munoz, X.. 2007. Image classification using random forests and ferns. In Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV 2007).

Boser, B.E., Guyon, I. M., Vapnik, V.N.. 1992. A training algorithm for optimal margin classifiers. In Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT 1992), pages 144-152.

Cand`es, E.J. & Donoho, D.L.. 2004. New tight frames of curvelets and optimal representations of objects with piecewise-C^2 singularities. In *Communications on Pure and Applied Mathematics*, 57(2), pages 219-266.

Cao, L.L. & Li, F.F.. 2007. Spatially coherent latent topic model for concurrent object segmentation and classification. In Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV 2007), pages 1–8.

Cao, L.L., Yu, J., Luo, J.B., and Huang, T.S.. 2009. Enhancing semantic and geographic annotation of web Images via logistic canonical correlation regression. In Proceedings of the 17th ACM International Conference on Multimedia (ACMMM 2009), pages 125-134.

Cham, T.J. & Cipolla, R.. 1995. Symmetry detection through local skewed symmetries. In *Image and Vision Computing,* 13(5), pages 439-450.

Chechik, G., Heitz, G., Elidan, G., Abbeel, P., Koller, D.. 2007. Max-margin classification of incomplete data. In Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS 2007), 19, pages 233-240.

Chechik, G., Heitz, G., Elidan, G., Abbeel, P., Koller, D.. 2008. Max-margin classification of data with absent features. In *Journal of Machine Learning Research*, 9, pages 1-21.

Chertok, M., & Keller, Y.. 2010. Spectral symmetry analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7), pages 1227-1238.

Chu, W.T., Liu, W.L., Yu, J. Y.. 2010. Age classification for pose variant and occluded faces. In Proceedings of the 18th ACM International Conference on Multimedia (ACMMM 2010), pages 743-746.

Chun, M. M. & Jiang, Y.. 1998. Contextual cueing: implicit learning and memory of visual context guides spatial attention, In *Cognitive Psychology*, 36, pages 28-71.

Chun, M. M.. 2000. Contextual cueing of visual attention, In *Trends in Cognitive Sciences*, 4, pages170-178.

Collobert, R., Sinz, F., Weston, J., Bottou, L.. 2006. Large scale transductive SVMs, In *Journal of Machine Learning Research*, 7, pages 1687-1712.

Cornelius, H. & Loy, G.. 2006. Detecting bilateral symmetry in perspective. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop, pages 191-191.

Davenport, J.L. & Potter, M.C.. 2004. Scene consistency in object and background perception. In *Psychological Science*, 15, pages 559-564.

Derrode, S. & Ghorbel, F.. 2004. Shape analysis and symmetry detection in gray-level objects using the analytical Fourier-Mellin representation. In *Signal Processing*, 84(1), pages 25–39.

Dick, U., Haider, P., Scheffer, T.. 2008. Learning from incomplete data with infinite imputations. In Proceedings of the 25th International Conference on Machine learning (ICML 2008), 307, pages 232-239.

Donate, A., Dahme, G. and Ribeiro, E.. 2006. Classification of textures distorted by waterwaves. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006), pages 421-424.

Enns, J., & Rensink, R.. 1998. Early completion of occluded objects. In *Vision Research*, 38, pages 2489-2505.

Felzenszwalb, P. & Huttenlocher, D.. 2004. Efficient graph-based imagesegmentation, In *International Journal of Computer Vision*, 59(2), pages 167–181.

Feng, K., Jiang, Z.N., He, W., and Ma, B.. 2011. A recognition and novelty detection approach based on Curvelet transform, nonlinear PCA and SVM with application to indicator diagram diagnosis. In *Expert Systems with Applications*, 2011.

Fergus, R., Perona, P., and Zisserman, A.. 2003. Object class recognition by unsupervised scale-invariant learning. In Proceedings of the 2003 IEEE Computer Soecity Conference on Computer Vision and Pattern Recognition (CVPR 2003), pages 264-271.

Ferzli, R. & Karam, L. J.. 2009. A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). In *IEEE Transactions on Image Processing*, 18(4), pages 717-728.

Fukuchi, K., Miyazato, K., Kimura, A., Takagi,S. and Yamato, J.. 2009. Saliency-based video segmentation with graph cuts and sequentially updated priors. In Proceedings of the 2009 IEEE International Conference on Multimedia and Expo (ICME 2009), pages 638-641.

Flusser, J., Suk, T. and Saic, S.. 1996. Recognition of images degraded by linear motion blur without restoration. In *Computing Suppl.,* 11, pages 37-51.

Flusser, J., Suk, T. and Zitová, B.. 2009. Moments and moment invariants in pattern recognition. In *Wiley, Chichester*.

Friedberg, S.A.. 1986. Finding axes of skewed symmetry. In *Computer Vision Graphics Image Process*, 34, pages 138-155.

Fu, S.Y., Yang, G.-S., Kuai, X.-K.. 2012. A spiking neural network based cortex-like mechanism and application to facial expression recognition, In *Computational Intelligence and Neuroscience*, pages 1-13.

Fukushima, K.. 1980. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. In *Biological Cybernetics*, 36, pages 193-201.

Gabor, D.. 1946. Theory of communication. In *Journal of the Institution of Electrical Engineers*, 93, pages 429-459.

Galleguillos, C., Rabinovich, A. and Belongie, S.J.. 2008. Object categorization using co-occurrence, location and appearance, In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008).

Gong, Y. & Xu, W.. 2007. Machine learning for multimedia content analysis (Multimedia Systems and Applications), Springer-Verlag New York, Inc., Secaucus, NJ.

Gross, A.D. & Boult, T.E.. 1994. Analyzing skewed symmetries. In *International*

Journal of Computer Vision, 13(1), pages 91-111.

Gross, R., Sweeney, L., Torre, F.D., Baker, S.. 2008. Semi-supervised learning of multi-factor models for face de-identification. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008).

Guo, Q., Guo, F.L., and Shao, J.Q.. 2010. Irregular shape symmetry analysis: theory and application to quantitative galaxy classification. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1730-1743.

Harel, J., Koch, C., and Perona, P.. 2007. Graph-based visual saliency. In Proceedings of the 20th Annual Conference on Neural Information Processing Systems, pages 545-552.

He, X.F., Cai, D. and Niyogi, P.. 2005. Tensor subspace analysis. In Proceedings of the 18th Annual Conference on Neural Information Processing Systems (NIPS 2005).

Hinton, G.E.. 2002. Training products of experts by minimizing contrastive divergence. In *Neural Computation*, 14(8), pages 1711-1800.

Hinton, G.E. & Salakhutdinov,R.R.. 2006. Reducing the dimensionality of data with neural networks. In *Science*, 313(5786), pages 504-507.

Hinton, G.E., Osindero, S., Teh, Y.W.. 2006. A fast learning algorithm for deep belief nets. In *Neural Computation*, 18(7), pages 1527-1554.

Hinton, G.E.. 2007. Learning multiple layers of representation. In *Trends in Cognitive Sciences*, 11(10), pages. 428-434.

Hinton, G.E.. 2010. A practical guide to training restricted Boltzmann machine. Technical report, University of Toronto, pages 1–21.

Hörster, E. & Lienhart, R.. 2008. Deep networks for image retrieval on large-scale databases. In Proceedings of the 16th ACM International Conference on Multimedia (ACMMM 2008), pages 643-646.

Huang, P.P., Chen, C.F., Wang, J.H.. 2011. Experimental studies of several reflection detection methods. In *Energy Procedia*, 13.

Hu, M.K.. 1962. Visual pattern recognition by moment invariants. In *IRE Transactions on Information Theory*, 8(2).

Hubel, D.H. & Wiesel, T.N.. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. In *Journal of Physiology*, 160,

pages 106-154.

Intraub, H.. 1981. Rapid conceptual identification of sequentially presented pictures. In *Journal of Experimental Psychology:* Human Perception and Performance, 7(3), pages 604-610.

Itti, L., Koch, C. and Niebur, E.. 1998. A model of saliency-based visual attention for rapid scene analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), pages 1254-1259.

Itti, L. & Koch, C.. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. In *Vision Research*, pages 1489-1506.

Jarrett, K., Kavukcuoglu, K., Ranzato, M. and LeCun, Y.. 2009. What is the best multi-stage architecture for object recognition? In Proceedings of the 12th IEEE International Conference on Computer Vision (ICCV 2009), pages 2146-2153.

Jesorsky, O., Kirchberg, K., Frischholz, R.. 2001. Robust face detection using the hausdorff distance. In Proceedings of the 3rd International Conference on Audio- and Video-based Biometric Person Authentication, 2091, pages 90-95.

Ji, H., Liu & C.Q.. 2008. Motion blur identification from image gradients. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008).

Jiang, Y.G., Ngo, C.W., Chang, S.F.. 2009. Semantic context transfer across heterogeneous sources for domain adaptive video search, In Proceedings of the 17th ACM International Conference on Multimedia (ACMMM 2009), pages 155-164.

Jones, J.P. & Palmer, L.A.. 1987. An evaluation of the two-dimensional, gabor filter model of simple receptive fields in cat striate cortex. In *Journal of Neurophysiology*, 58, pages 1233-1258.

Judd, T., Ehinger, K., Durand, F. and Torralba, A.. 2009. Learning to predict where humans look. In Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV 2009), pages 2106-2113.

Kanade, T., & Kender, J.R.. 1983. Mapping image properties into shape constraints: skewed symmetry, affine-transformable patterns, and the shape-from-texture paradigm. In *Human and Machine Vision*, pages 237–257, Academic Press, 1983.

Koch, C. & Ullman, S.. 1985. Shifts in selective visual attention: Towards the

Underlying Neural Circuitry. In *Human Neurobiology*, pages 219-227.

Kosslyn, S. M.. 1994. Image and Brain. In The MIT Press, Cambridge, MA.

Kuehnle., A.. 1991. Symmetry-based recognition of vehicle rears. In *Pattern Recognition Letters*, 12(4), pages 249–258.

Kumar, A. & Sminchisescu, C.. 2007. Support kernel machines for object recognition. In Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV 2007).

Lambrecht, C. & Verscheure, O.. 1996. Perceptual quality measure using a spatio-temporal model of the human visual system. In Proceedings of the SPIE Digital Video Compression: Algorithms Technology, 2668, pages 450 -461.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y.. 2007. An empirical evaluation of deep architectures on problems with many factors of variation, In Proceedings of the 24th International Conference on Machine Learning, pages 473-480.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.. 1989. Backpropagation applied to handwritten zip code recognition. In *Neural Computation*, 1(4), pages 541-551 1989.

LeCun, Y., Bottou, L., Bengio,Y., Haffner, P.. 1998. Gradient-based learning applied to document recognition. In Proceedings of the IEEE, 86(11), pages 2278-2324.

Lee, S., Collins, R.T., Liu, Y.X.. 2008. Rotation symmetry group detection via frequency analysis of frieze-expansions. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), pages 1-8.

Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.. 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009), pages 609-616.

Lei, Y. & Wong, K.C.. 1999. Detection and localisation of reflectional and rotational symmetry under weak perspective projection. In *Pattern Recognition*, 32, pages 167-180.

Li, F.F., Fergus, R., Pernoa, P.. 2004. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In Proceedings of the 2004 IEEE Computer Society Conference on

Computer Vision and Pattern Recognition (CVPR 2004), 106(1), pages 59-70.

Li, J., Socher, R. and Li, F.F.. 2009. Towards total scene understanding: classification, annotation and segmentation in an automatic framework, In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20, pages 2036-2043.

Li, L.F., Zhang, N, Duan, L.Y., Huang, Q.M., Du, J., Guan, L.. 2009. Automatic sports genre categorization and view-type classification over large-scale dataset. In Proceedings of the 18th ACM International Conference on Multimedia (ACMMM 2009), pages 653-656.

Liao, X.J., Li, H., Carin, L.. 2007. Quadratically gated mixture of experts for incomplete data classification. In Proceedings of the 24th International Conference on Machine learning (ICML 2007), 227, pages 553-560.

Lindner, A., Shaji, A., Bonnier, N., Süsstrunk, S.. 2012. Joint statistical analysis of images and keywords with applications in semantic image enhancement. In Proceedings of the 20th ACM International Conference on Multimedia (ACMMM 2012), pages 489-498.

Liu, Y., Xu, D., Tsang, I.W.H., Luo, J.B.. 2009. Using large-scale web data to facilitate textual query based retrieval of consumer photos. In Proceedings of the 17th ACM International Conference on Multimedia (ACMMM 2009), pages 55-64.

Liu, X.B., Cheng, B., Yan, S.C., Tang, J.H., Chua, T.S., Jin, H.. 2009. Label to region by bi-layer sparsity priors. In Proceedings of the 18th ACM International Conference on Multimedia (ACMMM 2010), pages 115-124.

Liu, Y.. 2011. Manifold learning for visual data analysis. PhD. Dissertation. Dept. of Computing, Hong Kong Polytechnic University, Hong Kong.

Liu, Y., Zhong, S.H., Li, W.J.. 2012. Query-oriented multi-document summarization via unsupervised deep learning. In Proceedings of 26th AAAI Conference on Artificial Intelligence (AAAI 2012), pages 1699-1705.

Loy, G. & Eklundh, J.. 2006. Detecting symmetry and symmetric constellations of features. In Proceedings of the 9th European Conference on Computer Vision (ECCV 2006), Part II, LNCS 3952, pages 508-521.

Lucchese, L.. 2004. Frequency domain classification of cyclic and dihedral symmetries of finite 2-D Patterns. In *Pattern Recognition*, 37, pages

2263–2280.

Machajdik, J. & Hanbury, A.. 2010. Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th ACM International Conference on Multimedia (ACMMM 2010), pages 83-92.

Mahajan, D.K. & Slaney, M.. 2010. Image classification using the web graph. In Proceedings of the 18th ACM International Conference on Multimedia (ACMMM 2010), pages 991-994.

Malik, N. R.. 1995. Electronic circuits: analysis, simulation, and design, upper saddle river, N.J.: Prentice Hall.

Mancas, M., Gosselin, B., Macq, B.. 2005. Fast and automatic tumoral area localisation using symmetry. In Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), pages 725–728.

Mandal, M. K., Aboulnasr, T., and Panchanathan, S.. 1996. Image indexing using moments and wavelets. In *IEEE Transactions on Consumer Electronics*, 42, pages 557 -565.

Marr, D. & Poggio, T.. 1979. A computational theory of human stereo vision. In Proceedings of the Royal Society, London B, 204, pages 301-328.

Marr, D.. 1982. Vision: a computational investigation into the human representation and processing of visual information, MIT Press.

Memisevic, R. & Hinton, G.E.. 2010. Learning to represent spatial transformations with factored higher-order Boltzmann machines. In *Neural Computation*, 22(6), pages 1473-1492.

Miller, A.G.. 1995. WordNet: A lexical database for English. In *Communications of the ACM* , 38(11), pages 39–41.

Miyajima, K. & Ralescu, A.I.. 1994. Spatial organization in 2D images, In Proceedings of the 3rd IEEE International Conference on Fuzzy Systems, pages 100–105.

Newman, J. N.. 1977. Marine hydrodynamics. In *MIT Press*.

Mitchell, T.M. 1997. Machine learning. In *McGraw Hill*.

Mitra, H.S. & Liu, Y. 2004. Local facial asymmetry for expression classification. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 889 – 894.

Mokhtarian, F.. 1996. Silhouette-based object recognition with occlusion through curvature scale space. In Proceedings of the 4th European Conference on Computer Vision (ECCV 1996), pages 566-578.

Moosmann, F., Nowak, E. and Jurie, F.. 2008. Randomized clustering forests for image classification. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9), pages 1632-1646.

Narwaria, M., Lin W.S., Mcloughlin, I.V., Emmanuel, S., Chia, L.T.. 2012. Fourier transform-based scalable image quality measure, In *IEEE Transaction on image processing*, 21(8), pages 3364-3377.

Nasar, J.L. & Li, M.. 2004. Landscape mirror: the attractiveness of reflecting water. In *Landscape and Urban Planning*, 66, pages 233-238.

Oliva A. & Torralba, A.. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. In *Journal International Journal of Computer Vision*, 42(3), pages 145-175.

Oliva, A., & Torralba. A.. 2007. The role of context in object recognition In *Trends in Cognitive Sciences*, 11, pages 520–527.

Opelt, A., Fussenegger, M., Pinz, A., and Auer, P.. 2004. Weak hypotheses and boosting for generic object detection and recognition. In Proceedings of the 5th European Conference on Computer Vision (ECCV 2004), 3022, pages 71-84.

Parkhurst, D., Law, K., and Niebur, E.. 2002. Modeling the role of salience in the allocation of overt visual attention. In *Vision Research*, pages 107-113.

Podolak, J., Golovinskiy, A., and Rusinkiewicz, S.. 2007. Symmetry-enhanced remeshing of surfaces. In *Symposium on Geometry Processing*, 257, pages 235-242.

Potter, M.C.. 1976. Short-term conceptual memory for pictures. In *Journal of Experimental Psychology: Human Learning and Memory*, 2(5), pages 509-522.

Prasad, V.S.N. & Yegnanarayana, B.. 2004. Finding axes of symmetry from potential fields. In *IEEE Transactions on Image Processing*, 13(12), pages 1559-1556.

Qi, X.Y., Zhang, L. Tan, C.L.. 2005. Motion deblurring for optical character recognition. In Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR 2005), pages 389-393.

Ranzato, M., Mnih, V., Hinton, G. E.. 2010. Genearting more realistic images using gated MRF's. In Proceedings of the 2010 Conference on Advances in Neural

Information Processing Systems (NIPS 2010), pages 2002-2010.

Ranzato, M., Susskind,J., Mnih,V., Hinton,G.E.. 2011. On deep generative models with applications to recognition. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), pages 2587-2864.

Riesenhuber, M. & Poggio, T.. 1999. Hierarchical models of object recognition in cortex. In *Nature Neuroscience*, 2(11), pages 1019-1025.

Sadaka, N.G., Karam, L.J., Ferzli, R., and Abousleman, G.P.. 2008. A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling. In Proceedings of the 15th International Conference on Image Processing (ICIP 2008), pages 369-372.

Salakhutdinov, R.R., Hinton,G.E.. 2007. Learning a nonlinear embedding by preserving class neighbourhood structure, In *AI and Statistics*.

Serre, T.. 2006. Learning a dictionary of shape-components in visual cortex: comparison with neurons, humans and machines，Ph.D. Dissertation, In *MIT Press*.

Serre, T., Wolf, L., Bileschi, S.M., Riesenhuber, M., Poggio, T.. 2007. Robust object recognition with cortex-like mechanisms, In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3), pages 411-426.

Shen, D., Ip, H.H.S., and Teoh, E.K.. 2000. Robust detection of skewed symmetries. In Proceedings of the 10th International Conference on Pattern Recognition (ICPR 2000), pages 1010–1013.

Shivaswamy, P.K., Bhattacharyya,C., Smola, A.J.. 2006. Second order cone programming approaches for handling missing and uncertain data. In *Journal of Machine Learning Research*, 7, pages 1283-1314.

Siagian, C. & Itti, L.. 2007. Rapid biologically-inspired scene classification using features shared with visual attention. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2), pages 300 -312.

Sim, T. & Baker, S.. 2003. The CMU pose, illumination, and expression database. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12), pages 1615-1618, 2003.

Smolensky, P.. 1986. Information processing in dynamical systems: foundations of harmony theory. In *Parallel Distributed Processing: Explorations in The Microstructure of Cognition*, 2: *Psychological and Biological Models*, *MIT*

Press, 194-281.

Stern, A., Kruchakov, I. Yoavi, E. and Kopeika, N.S.. 2000. Recognition of motion-blurred images by use of the method of moments. In Proceedings of the 12th International Conference on Pattern Recognition (ICPR 2000), pages 3881-3884.

Styles, E.A.. 2005. Attention, perception, and memory: an integrated introduction. First edition, In *Psychology Press*.

Sugiyama, M.. 2007. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. In *Journal of Machine Learning Research*, 8, pages 1027-1061

Tamura, H., Mori, S. and Yamawaki, T.. 1978. Texture features corresponding to visual perception. In *IEEE Transactions on Systems, Man, and Cybernetics*, 6, pages 460 – 473.

Taylor, N.R., Panchev, C., Hartley, M., Kasderdis, S., Taylor, J.G.. 2006. Occlusion, attention and object representations. In Proceedings of the International Conference on Aritificial Neural Networks (ICANN), pages 592-601.

Taylor, G., Fergus, R., LeCun, Y., and Bregler, C.. 2010. Convolutional learning of spatio-temporal features. In Proceedings of the 11th European Conference on Computer Vision (ECCV 2010), pages 140-153.

Torralba, A., Oliva, A., Castelhano, M.S., and Henderson, J.M.. 2006. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search, In *Psychological Review*, pages 766-786.

Tsai, M.H., Tsai, S.F., Huang, T.S.. 2010. Hierarchical image feature extraction and classification. In Proceedings of the 18th ACM International Conference on Multimedia (ACMMM 2010), pages 1007-1010.

Valenti, R., Jaimes, A., Sebe, N.. 2010. Sonify your face: facial expressions for sound generation. In Proceedings of the 18th ACM International Conference on Multimedia (ACMMM 2010), pages 1363-1372.

Vangool, L., Moons, T., Ungureanu, D. Pauwels, E.. 1995a. Symmetry from shape and shape from symmetry. In *Interational Journal of Robotics Research*, 14 (5), pages 407-424.

Vangool, L., Moons, T., Ungureanu, D., Oosterlinck, A.. 1995b. The characterisation and detection of skewed symmetry. In *Computer Vision and Image*

Understanding, 61 (1), pages 138-150.

VanRullen, R. & Thorpe,S. J.. 2001. The time course of visual processing: from early perception to decision-making. In *Journal of Cognitive Neuroscience,*13(4), pages 454-461.

Varadarajan, S. & Karam, L.J.. 2008. An improved perception-based no-reference objective image sharpness metric using iterative edge refinement. In Proceedings of the 15th International Conference on Image Processing (ICIP 2008), pages 401-404.

Wallis, G. & Bülthoff, H.. 1999. Learning to recognize objects. In *Trends in Cognitive Sciences*, 3(1), pages 22-31.

Wang, Y., Liu, Z., and Huang, J.. 2000. Multimedia content analysis using both audio and visual clues. In *IEEE Signal Processing Magazine*, 17(6), pages 12-36.

Wang, Z. & Bovik, A.C.. 2001. Embedded foveation image coding. In *IEEE Transactions of Image Processing*, 10(10), pages 1397-1410.

Wang, Z., Xia, D., Chang, E.Y.. 2010. A deep-learning model-based and data-driven hybrid architecture for image annotation. In Proceedings of the 18th ACM International Workshop on Very Large Scale Multimedia Corpus, Mining and Retrieval (VLS-MCMR 2010), pages 13-18.

Wang, M., Li, J., Huang, T., Tian, Y. Duan, L., and Jia, G.. 2010. Saliency detection based on 2D log-gabor wavelets and center bias. In Proceedings of the 18th ACM International Conference on Multimedia (ACMMM 2010), pages 979-982.

Weston, J., Ratle, F., Collobert, R.. 2008. Deep learning via semi-supervised embedding. In Proceedings of the 25th International Conference on Machine Learning (ICML 2008), pages 1168-1175.

Weyl, H.. 1952. Symmetry. In *Princeton University Press*, ISBN 0-691-02374-3.

Williams, D., Liao, X.J., Xue, Y., Carin, L.. 2005. Incomplete-data classification using logistic regression. In Proceedings of the 22nd International Conference on Machine learning, (ICML 2005), 119, pages 972-979.

Williams, D., Liao, X.J., Xue, Y., Carin, L., Krishnapuram, B.. 2007. On classification with incomplete data. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3), pages 427-436.

Xian, X., Xu, C.S., Wang, J.Q., 2010. Landmark image classification using 3D point clouds. In Proceedings of the 18th ACM International Conference on Multimedia (ACMMM 2010), pages 717-722.

Xu, W., Yang, M., Yu, K.. 2013. 3D convolutional neural networks for human action recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1), pages 221-231.

Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., and Lin, S.. 2007. Graph embedding and extension: a general framework for dimensionality reduction. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1), pages 40-51.

Yang, J., Yu, K., Gong, Y., Huang, T.. 2009. Linear spatial pyramid matching using sparse coding for image classification. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), pages 1794-1801.

Yu, H., Li, J., Tian, Y., Huang, T.. 2010. Automatic interesting object extraction from images using complementary saliency maps, In Proceedings of the 18th ACM International Conference on Multimedia (ACMMM 2010), pages 891-894.

Yuan, J., Li, J., Zhang, B.. 2007. Exploiting spatial context constraints for automatic image region annotation, In Proceedings of the 15th ACM International Conference on Multimedia (ACMMM 2007), pages 595–604.

Yuen, S.Y. & Chan, W.W. 1994. Two methods for detecting symmetries. In *Pattern Recognition Letters*, 15 (3), pages 279-286.

Zahid, N., Abouelala, O., Limouri, M., Essaid, A.. 2001 Fuzzy clustering based on K-nearest-neighbors rule, In *Fuzzy sets and Systems*, 20(2), pages 239-247.

Zhang, H., Guo, X.J., Cao, X.C.. 2010. Water reflection detection using a flip invariant shape detector. In Proceedings of the 20th International Conference on Pattern Recognition (ICPR 2010), pages 633-636.

Zhong, S.H., Liu, Y., Liu, Y., Chung, F.L.. 2010a. A semantic no-reference image sharpness metric based on top-down and bottom-up saliency map modeling. In Proceedings of the 17th International Conference on Image Processing (ICIP 2010), pages 1553-1556.

Zhong, S.H., Liu, Y., Chung, F.L.. 2010b. Fuzzy based contextual cueing for region level annotation. In Proceedings of the 2nd International Conference on Internet Multimedia Computing and Service (ICIMCS 2010), pages 1-6.

Zhong, S.H., Liu, Y., Shao, L., Chung, F.L.. 2011a. Water reflection recognition via minimizing reflection cost based on motion blur invariant moments. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR 2011).

Zhong, S.H., Liu, Y., Shao, L., Wu, G.S.. 2011b. Unsupervised saliency detection based on 2D Gabor and Curvelets transforms. In Proceedings of the 3rd International Conference on Internet Multimedia Computing and Service (ICIMCS 2011), pages 146-149.

Zhong, S.H., Liu, Y., Liu, Y.. 2011c. Bilinear deep learning for image classification. In Proceedings of the 19th ACM International Conference on Multimedia (ACMMM 2011), pages 343-352.

Zhong, S.H., Liu, Y., Liu, Y., Chung, F.L.. 2012a. Region level annotation by fuzzy based contextual cueing label propagation. In *Multimedia Tools and Applications*.

Zhong, S.H., Liu, Y., Zhang, Y., Chung, F.L.. 2012b. Attention modeling for face recognition via deep learning. In Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci 2012), pages 2609-2614.

Zhong, S.H., Liu, Y., Chung, F.L., Wu, G.S.. 2012c. Semiconducting bilinear deep learning for incomplete image recognition. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (ICMR 2012).

Zhou, S.S., Chen, Q.C. and Wang, X.L.. 2010. Discriminate deep belief networks for image classification. In *Proceedings* of the 17th IEEE *International Conference on Image Processing (ICIP 2010),* pages 1561-1564.

Zhu, X.. 2006. Semi-supervised learning literature survey. Technical report 1530, University of Wisconsin-Madison.