The Dissertation Committee for Che-Chun Su
certifies that this is the approved version of the following dissertation:

Applied Statistical Modeling of Three-Dimensional Natural Scene Data

Committee:

Alan C. Bovik, Supervisor

Lawrence K. Cormack, Co-Supervisor

Constantine Caramanis

Donald S. Fussell

Kristen Grauman

Applied Statistical Modeling of Three-Dimensional Natural Scene Data

by

Che-Chun Su, B.S.; M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2014

To my family: Liu-Tiao, Chiu-Shia, Che-Wei, and Sih-Wun

Acknowledgments

A week after I successfully defended my Ph.D. dissertation, I still can feel the excitement and unreality when receiving congratulations from the committee members. With no doubt I will forever remember these emotions, which will also remind me of all the supports I was given during this long journey.

First and foremost, I would like to express my deepest gratitude to Dr. Bovik, who is absolutely the world's best advisor, for his guidance, encouragement, and trust. Dr. Bovik has not only taught me how to conduct world-class research, but also had immense influence on my life. I have learned from him it is discipline and passion that can lead you to the greatness. I have never lost confidence or thought of giving up when facing any challenge or difficulty in the pursuit of Ph.D., as I know he is always there to support me and has faith in me. Without Dr. Bovik's invaluable advice, I would have never finished. My appreciation for his fatherly support is beyond words.

Next up is Dr. Cormack, my co-advisor. I have been always inspired by his knowledgeable advice and comments, as well as his unparalleled sense of humor. I would like to thank Dr. Cormack for his support and being such a great role model.

My dad, Liu-Tiao Su, is the most important person that has influenced my life. It is his education and love that make me who I am today. I know he has always been looking after and shining on me from heaven, and I believe what I have accomplished has made him proud and smile. My mom, Chiu-Shia Lee, and my brother, Che-Wei Su, have given me their unconditional support and love whenever I needed them. I have been truly blessed and so grateful to have them in my life. I would like to thank my wife, Sih-Wun Yu, who is the loveliest woman around the world and my best life-time partner, for always being patient, understanding, and supportive. I would also like to thank my father-in-law, Jau-Song Yu, and mother-in-law, Su-Chen Liu, for always encouraging and supporting me like their son. There is no way I can repay the greatest support and love from my family.

I would like to thank all my fellow colleagues at the Laboratory for Image and Video Engineering (LIVE), both past and present, Yang, Joonsoo, Sina, Anush, Rajiv, Gautam, Ajay, Michele, Anish, Ming-Jun, Greg, Dinesh, Lark, Deepti, Janice, Todd, and Furkan. You have made LIVE the most enjoyable and memorable place during my academic career.

Finally, I would like to thank my distinguished committee members, Dr. Caramanis, Dr. Fussell, and Dr. Grauman, for their valuable and insightful comments that enhance the depth and breadth of this dissertation.

Applied Statistical Modeling of Three-Dimensional Natural Scene Data

Che-Chun Su, Ph.D.
The University of Texas at Austin, 2014

Supervisors: Alan C. Bovik
Lawrence K. Cormack

Natural scene statistics (NSS) have played an increasingly important role in both our understanding of the function and evolution of the human vision system, and in the development of modern image processing applications. Because depth/range, i.e., egocentric distance, is arguably the most important thing a visual system must compute (from an evolutionary perspective), the joint statistics between natural image and depth/range information are of particular interest. However, while there exist regular and reliable statistical models of two-dimensional (2D) natural images, there has been little work done on statistical modeling of natural luminance/chrominance and depth/disparity, and of their mutual relationships. One major reason is the dearth of high-quality three-dimensional (3D) image and depth/range database. To facilitate research progress on 3D natural scene statistics, this dissertation first presents a high-quality database of color images and accurately co-registered depth/range maps using an advanced laser range scanner mounted with a high-end digital single-lens reflex camera.

By utilizing this high-resolution, high-quality database, this dissertation performs reliable and robust statistical modeling of natural image and depth/disparity information, including new bivariate and spatial oriented correlation models. In particular, these new statistical models capture higher-order dependencies embedded in spatially adjacent bandpass responses projected from natural environments, which have not yet been well understood or explored in literature.

To demonstrate the efficacy and effectiveness of the advanced NSS models, this dissertation addresses two challenging, yet very important problems, depth estimation from monocular images and no-reference stereoscopic/3D (S3D) image quality assessment. A Bayesian depth estimation framework is proposed to consider the canonical depth/range patterns in natural scenes, and it forms priors and likelihoods using both univariate and bivariate NSS features. The no-reference S3D image quality index proposed in this dissertation exploits new bivariate and correlation NSS features to quantify different types of stereoscopic distortions. Experimental results show that the proposed framework and index achieve superior performance to state-of-the-art algorithms in both disciplines.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Given continuous, rapid advances in three-dimensional (3D) imaging and display technology, the quantity and quality of 3D and stereoscopic data, e.g., image, video, movies, geographic models, etc., has increased dramatically. A substantial amount of research has been conducted towards better understanding the perception of 3D content, with the aim of improving the quality of visual experience delivered by 3D technologies and products. For example, impairments in viewing and comfort when using 3D displays has been studied towards developing auto-stereoscopic 3D displays [1, 2]. There are numerous sources of distortion and visual discomfort that can be experienced when viewing 3D content. Understanding how the depth sensation is affected by improper geometry (stereography) and by signal distortion are crucial open problems [3–5].

Natural scene[1] statistics (NSS) have been proven to be important ingredients towards understanding both the evolution of the human visual system

[1]By 'natural scenes' we mean pictures of the real world, both arising in natural as well as man-made settings, obtained by a good-quality camera under good (photopic) conditions without distortion.

and the design of image processing algorithms [6]. Luminance/chrominance and depth/disparity information all play important roles in the perception of images of natural scenes and in stereoscopic vision. Likewise, models of the statistics of natural images and depth maps play an important role in modern image processing applications. Therefore, reliable statistical models of natural depth/disparity and luminance/chrominance image information can be used to not only improve the aforementioned 3D display and viewing experiences, but also benefit a variety of 3D image/video and vision algorithms.

However, while there exist regular and reliable statistical models of two-dimensional (2D) natural images, there has been little work done on statistical modeling of natural luminance/chrominance and depth/disparity, and of their mutual relationships. One major reason is the dearth of high-quality 3D image and depth/range database. To facilitate the variety of research relevant to natural scene statistics between 2D images and the depth/disparity information, we constructed a high-quality database of color images and co-registered depth/range maps using an advanced laser range scanner mounted with a high-end digital single-lens reflex camera. By utilizing this high-resolution, high-quality database, we develop reliable and robust statistical models of natural images and depth/disparity information, and show that these joint natural scene statistical models can be utilized to improve computational stereoscopic/3D (S3D) and vision problems, e.g., stereo correspondence, depth estimation, S3D image quality assessment, etc.

1.2 Contributions

The contributions of this dissertation are threefold. First, a high-quality data set of accurately co-registered color images and depth/range maps, the LIVE Color+3D Database [7, 8], has been presented and made publicly available. Second, we perform advanced statistical analyses on natural images and depth/range maps to develop reliable and robust NSS models. Finally, to demonstrate the efficacy and effectiveness of these new statistical models, we apply them to solve two practical, yet extremely challenging problems, depth estimation from monocular images and no-reference S3D image quality assessment. Experimental results show that with the aid of these advanced NSS models, we have achieved superior performance to the state-of-the-art algorithms in both disciplines.

1.2.1 LIVE Color+3D Database

We constructed the LIVE Color+3D Database [7, 8], a high-quality data set of accurately co-registered color images and depth/range maps, using an advanced range scanner, RIEGL VZ-400 [9], with a Nikon D700 digital camera mounted on top of it. This database serves as a solid basis on which a variety of statistical analysis and modeling of luminance/chrominance and depth/range data in natural images are performed. The LIVE Color+3D Database includes two releases. Release-1 contains 12 sets of color images with corresponding ground-truth range maps at a high-definition resolution of 1280×720, and Release-2 consists of 99 stereoscopic pairs, i.e., left- and

right-views, of color images and ground-truth range maps in high-definition resolution of 1920×1080. The natural environments where the image and depth/range data were collected are in areas around Austin, Texas, including the campus at The University of Texas, recreational parks, the Texas State Capitol, etc. These high-quality, high-resolution, and accurately co-registered color images and depth/range maps make the derived natural scene statistical models exceptionally robust and reliable.

1.2.2 Advanced Statistical Models of Natural Image and Depth Data

By utilizing the LIVE Color+3D Database, we first derive marginal and conditional priors relating natural luminance/chrominance and disparity, and demonstrate their efficacy with application to a chromatic Bayesian stereo algorithm. In particular, we use the univariate generalized Gaussian distribution (GGD) to fit the empirical histograms of sub-band coefficients after perceptual multi-scale, multi-orientation bandpass decomposition. Moreover, there exist higher-order dependencies between spatially neighboring bandpass responses that are not yet well understood or utilized in literature. Towards filling this gap, we further develop new bivariate and spatial oriented correlation models that capture statistical regularities between perceptually decomposed natural luminance and depth samples.

1.2.3 Depth Estimation from Monocular Natural Images

Inspired by psychophysical evidence of NSS-driven visual signal processing in HVS, we propose a new Bayesian model, which we call Natural3D, for estimating depths from single monocular images by employing reliable and robust NSS models of natural images and depth maps as priors. Specifically, we utilize the statistical relationships between local image features and depth variations inherent in natural images. By observing that similar depth structures may exist in different types of luminance textured regions in natural scenes, we build a dictionary of canonical depth patterns as the prior, and fit a multivariate Gaussian mixture (MGM) model to associate local image features to different depth patterns as the likelihood. Following the development of Natural3D, we describe how we trained and tested it on two publicly accessible databases of natural image and range data, the LIVE 3D+Color Database Release-2 [8], which consists of 99 pairs of natural images and accurately co-registered ground-truth depth maps of high-definition resolution (1920×1080), and the widely used Make3D Laser+Image Dataset-1 [10–12]. Compared with the state-of-the-art depth estimation method, we achieve superior performance in terms of pixel-wise estimated depth error, but better capability of recovering relative distant relationships between different objects and regions in natural images.

1.2.4 No-Reference Stereoscopic/3D Image Quality Assessment

We propose a no-reference S3D image quality assessment (IQA) framework utilizing both univariate and bivariate NSS models. A new convergent cyclopean image model is developed to gauge the perceptual quality of 3D percepts formed by HVS when viewing stereoscopic image pairs. The proposed framework, dubbed Stereoscopic/3D BLind Image Naturalness Quality (S3D-BLINQ) Index, deploys a novel set of NSS features including both spatial-domain and wavelet-domain univariate models, as well as recently explored bivariate and correlation statistics. We validate the robustness and effectiveness of the bivariate and correlation NSS features extracted from distorted stereopairs. Experimental results demonstrate that with the aid of convergent cyclopean images and the augmentation of bivariate and correlation NSS models, S3D-BLINQ Index outperforms state-of-the-art full- and no-reference 3D IQA algorithms on both symmetrically and asymmetrically distorted stereoscopic image pairs.

1.3 Organization

The rest of this dissertation is organized as follows. We first present the LIVE Color+3D Database in Chapter 3. Then, Chapter 2 reviews literature and provides background knowledge on natural scene statistical modeling, depth estimation, and S3D quality assessment. Next, Chapter 4 and 5 detail the univariate, bivariate, and spatial oriented correlation models we developed from natural images and depth/range maps. We describe the proposed

Bayesian framework of depth estimation from monocular natural images in Chapter 6. Chapter 7 deals with the proposed no-reference S3D image quality index. Finally, conclusion and future work are given in Chapter 8.

Chapter 2

Background

This chapter provides a brief review of previous work and background knowledge on the topics to be discussed in the rest of this dissertation. It is by no means exhaustive and only summarizes relevant literature, while pointing the interested readers to more comprehensive references.

2.1 Natural Scene Statistics

The evolution of the human vision apparatus has involved many different factors and driving forces, such as natural scene statistics, the computational resources available in the human brain, and the kinds of tasks that humans need to perform [13]. Natural scene statistics (NSS) have been proven to be important ingredients towards understanding both the evolution of the human vision system and the design of image processing algorithms [6].

2.1.1 2D Images

Extensive work has been conducted towards understanding the luminance statistics of natural scenes [14–17], and the link between natural scene statistics and neural processing of visual stimuli [18, 19]. It has been discovered

that the distributions of local quantities such as luminance contrast are scale invariant, and that the power spectra of natural images vary as $1/f^2$ with radial spatial frequency f. This has been successfully used to explain and predict early visual processing in both insects and higher vertebrates [14, 15, 20]. The statistics of natural images have been found to exhibit non-Gaussian behavior, but when projected onto appropriate multi-scale spaces, e.g., using wavelet bases [21], or 2D Gabor decompositions [14], the resulting coefficients are found to obey regular statistical models, such as Gaussian scale mixtures [22]. These statistical models have been successfully applied in a variety of image and video applications, such as image de-noising and restoration [23], and image quality assessment [6, 24–26]. Moreover, it has also been suggested that the spatial receptive fields of the simple cells in mammalian primary visual cortex (V1) can be characterized as being localized, oriented, and bandpass, which are comparable with the basis functions of wavelet or Gabor transforms [27]. It is also widely believed that the goal of the early stages of visual signal processing is to transform and encode the input stimuli from natural images into a sparse, efficient representation to utilize the available computation resources of neurons [18, 28]. This sparse and efficient coding strategy along with its over-complete basis leads to non-linear relationships between visual stimuli and neural responses, which can be used to help understand higher stages of cortical processing in human vision systems [29].

2.1.2 Stereoscopic/3D Images and Depth/Disparity Maps

Very little work has been done on analyzing the joint statistics of luminance and range in natural scenes, and we haven't found any relating the statistics of color and range. One major reason for the lack of studies on color and range statistics has been limited access to high quality databases of color images and associated ground-truth range maps. Potetz *et al.* [30] constructed a database of co-registered 2D color images and range maps, and discovered that there is a correlation between range and intensity of luminance in natural scenes. This negative range-luminance correlation merely reflects the fact that nearer objects tend to appear brighter than far objects, on average. The authors also deployed a few convex range filters, selected for specific structural properties relevant to computer vision, to filter both range and luminance images. Using a canonical correlation analysis, they found a relatively low degree of strictly linear correlation between the "structure-filtered" luminance and range patches. In a later study on the same dataset, Potetz *et al.* [31] examined the relationships between luminance and range over multiple scales and applied their results to shape-from-shading problems. In [32], Yang *et al.* explored the statistical relationships between luminance and disparity in the wavelet domain using a public co-registered database of range maps and luminance natural images from [30], and applied the derived models to a Bayesian stereo algorithm. The authors found that the correlations between bandpass luminance and bandpass disparity are stronger in coarser scales, and also showed that the statistical models of 3D natural scenes improve the qual-

ity of computed disparity maps. Recently, Su *et al.* [33] proposed reliable statistical models for both marginal and conditional distributions of luminance/chrominance and disparity in natural images, and used these models to significantly improve a chromatic Bayesian stereo algorithm.

2.2 Depth Estimation

Given the rapid growth and widespread popularity of 3D films and entertainment devices, understanding how statistical depth information relates to image luminance and color statistics in real-world images and videos has gained increased relevance to practical 3D image analysis problems over the past several years. One such problem that may benefit by a principled statistical approach, based on natural scene models, is recovering the three-dimensional structure of visual scenes from single monocular images. Success in this endeavor could give us a better understanding of the 3D relationships that exist between objects and their projected 2D images, with potential benefit to the solution of numerous 3D visual tasks, such as robotic navigation, visual surveillance, 3D cinema, predicting 3D video quality, and so on.

By seamlessly combining binocular and monocular cues, humans are able to perceive depth and reconstruct the geometry of the 3D visual space so quickly and effortlessly that an individual rarely feels how difficult and ill-posed this problem can be. Even given a single color image, or by gazing with one eye closed, a human viewer can still perceive meaningful depth structures and 3D relationships such as relative distances from the visible environment. Yet

automatically estimating range (egocentric distance) from a single monocular image remains a very difficult problem, which is generally attacked by using a combination of complementary depth cues, such as color, shading, texture, perspective, etc.

Of course, much of the work on 3D scene reconstruction has focused on depth from binocular vision, i.e., stereopsis. In [34], Scharstein and Szeliski provide a comprehensive review and summary of dense two-frame stereo algorithms. Other depth recovery algorithms require multiple images, e.g., structure from motion [35] and depth from defocus [36]. Such 'multi-view' algorithms consider geometric/triangulation differences between image samples, while largely ignoring the variety of strictly monocular cues that contain useful depth information.

Recently, many different methods and algorithms have been developed to tackle the problem of depth estimation from a single monocular image. These models typically deploy variants of shape from shading [37, 38] and shape from texture [39, 40]. However, the efficacy of these algorithms are limited by the information in the image luminance and texture variations, unless additional structural assumptions or specific constraints are placed on their solutions.

One of the first methods to utilize monocular image features to capture depths [41] reconstructs a simple 3D model of outdoor scenes by making the assumption that an image could be divided into a few planar surfaces, and that pixels could be classified using a small number of limited labels, e.g.,

ground, sky, and vertical walls. Along similar lines, Delage *et al.* [42] developed a dynamic Bayesian network to reconstruct the locations of walls, ceilings, and floors by finding the most likely floor-wall boundaries in indoor scenes. In [11, 12], a supervised learning strategy was devised to infer the absolute depth associated with each pixel of a monocular image. They assumed that most 3D scenes are made up of small planar surfaces, and used this assumption in conjunction with a Markov Random Field (MRF) model of textural and luminance gradient cues to infer depth. Nagai *et al.* [43] used Hidden Markov Models (HMM) to reconstruct surfaces of known classes of objects such as hands and faces from single images. In [44], an example-based approach was proposed by Hassner *et al.* to estimate the depths of objects given a set of known categories. In [45], Torralba and Oliva took a very different (but limited) approach by studying the relationship between the Fourier spectrum of an image and its mean depth. Specifically, they proposed a probability model to estimate the absolute mean depth of a 3D scene using information extracted from the global and local spectral signatures of a 2D image of it. In [46], Liu *et al.* incorporated semantic labels to guide a monocular 3D reconstruction process, thereby achieving better depth estimates. By conditioning on different semantic labels, they were able to better model absolute depth as a function of local pixel appearance. More recently, Karsch *et al.* [47] presented an optimization framework to generate a most likely depth map by first matching high-level image features to find candidates from a database, then warping the candidate depth maps under a set of spatial regularization constraints.

In addition, many static monocular Shape-from-X algorithms have been devised (too many to survey) that estimate relative local depths by assuming the presence of one or more specific attributes, e.g., texture or shading gradients. Our belief is that such cues are embedded in the local, scale-invariant but space-varying natural statistics of real-world images. Certain natural scene statistics (NSS) models have been shown to provide good descriptions of the statistical laws that govern the behavior of images of the world. NSS models are useful tools for both understanding the evolution of human vision systems (HVS) [18, 28] and solving diverse visual problems [6, 23, 48, 49]. In particular, there has been work conducted on exploring the 3D NSS of depth/disparity maps of the world, how they correlate with 2D luminance/color NSS, and how such models can be applied. For example, Potetz *et al.* [31] examined the relationships between luminance and range over multiple scales and applied their results to a shape-from-shading problem. Yang *et al.* [32] explored the statistical relationships between luminance and disparity in the wavelet domain, and applied the derived models to improve a canonical Bayesian stereo algorithm. In [33], Su *et al.* proposed new models of the marginal and conditional statistical distributions of the luminances/chrominances and the disparities/depths associated with natural images, and used these models to significantly improve a chromatic Bayesian stereo algorithm. Recently, Su *et al.* developed new bivariate and correlation NSS models that effectively capture the higher-order dependencies between spatially adjacent bandpass responses to both natural images and depth maps [50, 51]. In [52], the authors further utilized these

models to create a blind 3D perceptual image quality model that operates on distorted stereoscopic image pairs. An algorithm derived from this model was shown to deliver quality predictions that correlate very highly with recorded human subjective judgments.

2.3 Stereoscopic/3D Image Quality Assessment

As with other digital visual media [49], the quantity of S3D images and videos that are delivered by the cinema, television, and online entertainment industries on a daily basis for human consumption has been growing dramatically over the past few years. According to recent theatrical market statistics gathered by the Motion Picture Association of America (MPAA) [53], the proportion of cinema screens that are 3D has reached 35% worldwide, and approximate half of all moviegoers viewed at least one 3D movie in 2012. As Hollywood director James Cameron, who directed and produced Avatar, stated in an interview with BBC news in Aug. 2013 [54], "All forms of entertainment will eventually be 3D, because that's how we see the world." In fact, the wave of 3D has not been limited to the entertainment industry. Given the development of greatly improved acquisition and display technologies, S3D images and videos can provide natural and versatile visual presentations for numerous applications, including robot navigation [55], remote education [56], anatomical exploration [57], therapeutic treatment [58], and so forth. As these large volumes of S3D data are making their way to consumers and other users, a variety of issues have arisen regarding efficient compression and reliable transmission of

S3D content, especially when being transmitted over already-stressed wireless networks. At every stage of capture, compression, storage, and transmission, it is desirable to maximize the quality of the final visual experience, and in this regard, incorporating principles of the human perception of S3D quality is of importance [59, 60].

The ideal way to assess perceived visual quality is to run a subjective test to gauge human opinions [61]. However, subjective quality assessment has two obvious disadvantages, making it unsuitable for practical applications. First, the procedure of subjective quality assessment is expensive, tedious, and time-consuming as it has to be performed with great care in order to obtain meaningful results. Second, it is impossible to integrate subjective image quality assessment (IQA) tasks of any value into nearly any system for communicating real-time visual data to human users. Therefore, it is desirable to develop automated algorithms that can predict the perceptual quality of visual data streams, including S3D image pairs.

As an interesting and important application of the new parametric correlation model [51], we develop an automatic no-reference (NR) S3D image quality model that is able to automatically predict the perceptual quality of distorted S3D images, without benefit of any reference signal, making it useful for practical applications. Models that attempt to solve the S3D IQA problem may be distinguished by whether they utilize computed or measured depth/disparity information from the stereoscopic pairs. Thus, the simplest S3D IQA models apply off-the-shelf 2D IQA algorithms to both left and right

stereo images, then aggregate the two quality scores to form a final prediction of the quality of the fused stereopair. Both full-reference 2D models, e.g., PSNR [62], SSIM [63], and MS-SSIM [64], and 2D NR models, e.g., DIIVINE [25], BLIINDS [65], and BRISQUE [66], can be used in this way. Yasakethu *et al.* [67] applied a variety of 2D IQA algorithms to the left and right views independently, then averaged them to obtain S3D quality scores, achieving fairly good correlation with both perceived image and depth quality. Gorley *et al.* [68] reported a full-reference S3D IQA model that they found preferable to the PSNR for controlling practical S3D image compression rates. Recently, there has been increased emphasis on developing S3D IQA models that utilize the encoding of depth/disparity stimuli from the natural environment by modeling cortical neurons with disparity-tuned receptive fields [69–71]. Benoit *et al.* [72] predict the quality of S3D image pairs using the disparity information computed from off-the-shelf stereo algorithms [73, 74]. Recent studies have demonstrated the importance of depth/disparity for understanding perceptual S3D image quality. For example, Chen *et al.* [75] showed that when viewing S3D image pairs, subjects tend to agree on perceived image quality, but have more diverse opinions on their sensations of depth.

Although the depth/disparity information extracted from S3D image pairs does affect the perceptual quality of viewed stereoscopic images, the question of how best to exploit this information remains incompletely answered. You *et al.* [76] attempted to quantify the degradation of disparity information by applying 2D IQA algorithms on the disparity maps computed from both

reference and distorted left-right image pairs [74]. Disparity information can also be used indirectly to bolster an S3D IQA algorithm. For example, Sazzad *et al.* [77] utilized disparity information to design an NR IQA algorithm to predict the quality of both symmetrically and asymmetrically JPEG-coded stereo image pairs.

However, the ultimate goal of an S3D IQA algorithm is to form predictions of the quality of the ultimate cyclopean image [78] formed within an observer's mind when a left-right image pair is stereoscopically presented. Towards this end, several recent researchers have attempted to evaluate perceptual quality by synthesizing an intermediate image that more-or-less agrees with cyclopean perception. Maalouf *et al.* [79] proposed a reduced-reference quality metric that compares the sensitivity coefficients [80] extracted from the two cyclopean images synthesized from the reference and distorted stereopairs. Chen *et al.* [81] proposed a full-reference S3D IQA algorithm exploiting a perceptually synthesized cyclopean image to account for binocular rivalry. In [82], the authors extended this framework to create a no-reference model using 2D and 3D natural scene statistical features extracted from S3D image pairs.

Chapter 3

LIVE Color+3D Database

3.1 Data Acquisition

We constructed a high-quality database of color images and co-registered depth/range maps, the LIVE Color+3D Database [7, 8], using an advanced range scanner, RIEGL VZ-400 [9], with a Nikon D700 digital camera mounted on top of it. This dissertation utilize the natural image and depth/range data contained in the LIVE Color+3D Database. Figure 3.1 shows the flow of constructing the database. Since there are inevitable translational and rotational shifts when mounting the camera onto the range scanner, calibration needs to be performed before data acquisition. The mounting calibration is done manually by using the RIEGL RiSCAN Pro software, which is designed for scanner operation and data processing [83]. Next, to acquire the image and range data in natural scenes, the range scanner rotates and fires laser beams to measure distances, and then the digital camera takes the picture with the same field of view. The acquired range data are exported from the range scanner as point clouds with the three-dimensional coordinate and the range value, while the image data are stored in the digital camera as JPEG files. Finally, to obtain the aligned 2D range map with the 2D image, the 3D point clouds are projected and transformed into the 2D range map by applying the pinhole

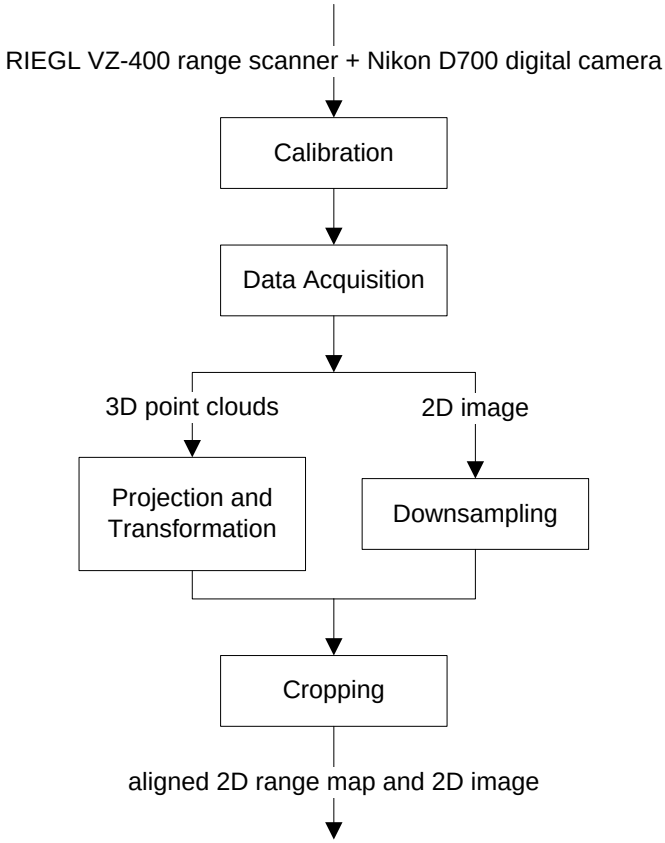


Figure 3.1: The flow of data acquisition.

camera model with lens distortion [84].

The angular step-width of the RIEGL VZ-400 range scanner can be as small as 0.0024° in both vertical and horizontal directions. However, higher scanning resolution means longer scanning time as well as higher probability of inconsistency between the 3D point clouds and the corresponding 2D image in natural scenes. Therefore, with careful consideration of the trade-off between the scanning resolution and the quality of range map, the angular step-width

of 0.04° is adopted while acquiring the database. In addition, since the digital camera is mounted in portrait mode onto the range scanner, the field of view for the 3D point clouds needs to be adjusted to match the aspect ratio of the portrait image, resulting in 60° and 100° field of views in the horizontal and vertical direction, respectively. As a result, the resolution of the 3D point clouds from the range scanner equals to $\frac{60}{0.04} \times \frac{100}{0.04} = 1500 \times 2500$ (points), which is smaller than the image resolution captured by the digital camera, which is 2823×4256 (pixels). To provide accurate aligned 2D range map and 2D image while keeping their resolution as high as possible, the 3D point clouds are projected and transformed into the 2D range map with the same resolution, and the original 2D image is down-sampled to the same size. Finally, the inaccurate range values at boundary pixels in the natural scene are removed by cropping the aligned 2D range map and 2D image into the target resolution, which is appropriate for display and viewing on digital TV and monitors.

The following equations summarize and explain how the aligned range and image data are acquired. First, the three-dimensional coordinates of the point clouds are converted into the undistorted two-dimensional pixel coordinates.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{A} \cdot \mathbf{RT} \cdot \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{3.1}$$

$$\mathbf{A} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

$$\mathbf{RT} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \tag{3.3}$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} x/z \\ y/z \end{bmatrix} \tag{3.4}$$

where $[X\ Y\ Z]^T$ is the three-dimensional coordinate of the point cloud, $\mathbf{A}$ is the camera's intrinsic matrix, $\mathbf{RT}$ is the joint rotation-translation matrix, and $[u\ v]^T$ is the undistorted two-dimensional pixel coordinate. In the intrinsic matrix $\mathbf{A}$, $[c_x\ c_y]^T$ is the coordinate of the principal point, which is usually at the image center, and (f_x, f_y) are the focal lengths along the x- and y- axes, all expressed in the unit of pixels. The parameters in the joint rotation-translation matrix $\mathbf{RT}$ are computed from the manual calibration after mounting the digital camera onto the range scanner.

Since real lens usually have distortions, ex. radial and tangential, the distorted two-dimensional pixel coordinates are computed by transforming the undistorted two-dimensional pixel coordinates as follows:

$$\begin{aligned} u_d = u + u' f_x (k_1 r^2 + k_2 r^4 + k_3 r^6 + k_4 r^8) + \\ 2 f_x u' v' p_1 + p_2 f_x (r^2 + 2u'^2) \end{aligned} \tag{3.5}$$

$$\begin{aligned} v_d = v + v' f_y (k_1 r^2 + k_2 r^4 + k_3 r^6 + k_4 r^8) + \\ 2 f_y u' v' p_2 + p_1 f_y (r^2 + 2v'^2) \end{aligned} \tag{3.6}$$

$$u' = (u - c_x)/f_x \tag{3.7}$$

$$v' = (v - c_y)/f_y \tag{3.8}$$

$$r = u'^2 + v'^2 \tag{3.9}$$

where $[u_d\ v_d]^T$ is the distorted two-dimensional pixel coordinate, (k_1, k_2, k_3, k_4) are the radial distortion coefficients, and (p_1, p_2) are the tangential distortion coefficients.

After the distorted two-dimensional pixel coordinate of each point cloud is computed, the aligned 2D range map is obtained by filling the range value at each pixel location with the one at the closest distorted two-dimensional pixel coordinate.

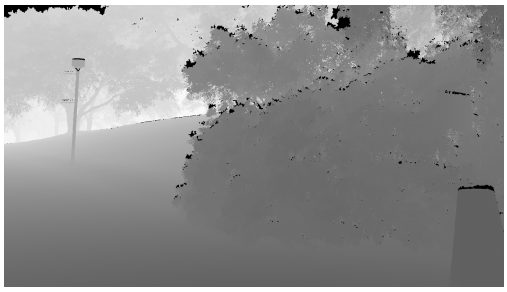
3.2 Example Natural Scenes

The LIVE Color+3D Database includes two releases. Release-1 contains 12 sets of color images with corresponding ground-truth range maps at a high-definition resolution of 1280×720, and Release-2 consists of 99 stereoscopic pairs, i.e., left- and right-views, of color images and ground-truth range maps in high-definition resolution of 1920×1080. The natural environments where the image and depth/range data were collected are in areas around Austin, Texas, including the campus at The University of Texas, recreational parks, the Texas State Capitol, etc. Figure 3.2 and 3.3 show example natural scenes, including both color images and corresponding ground-truth depth maps, in the LIVE Color+3D Database Release-1 and -2, respectively.

(a) Scene-1: image

(b) Scene-1: depth map

(c) Scene-2: image



(d) Scene-2: depth map

(e) Scene-3: image

(f) Scene-3: depth map

Figure 3.2: Example scenes in the LIVE Color+3D Database Release-1.

(a) Scene-1: image

(b) Scene-1: depth map

(c) Scene-2: image

(d) Scene-2: depth map

(e) Scene-3: image

(f) Scene-3: depth map

Figure 3.3: Example scenes in the LIVE Color+3D Database Release-2.

Chapter 4

Univariate Color and Depth Priors[1]

4.1 Introduction

Color is an important and dense natural visual cue that is used by the brain to reconstruct both low-level and high-level visual percepts. The cone photoreceptors, which are densely distributed in the fovea centralis of the retina, capture and convey rich information both in space and time. While the cones themselves do not encode color, they do come in three types that have different spectral sensitivities. Hence, comparisons of the outputs of the different cone types by the retinal ganglion cells allow dense spatiotemporal chromatic information to be transmitted from the retina to the primary visual cortex (V1). Likewise, color can be used at later processing stages to help infer large-scale shape information to better solve visual tasks by both humans and machine algorithms [85].

Moreover, it has been demonstrated that the perception of color and depth are related [86], and that chromatic information can be used to improve the solution of stereo correspondence problems [87, 88]. Therefore, interaction

[1]This chapter has been published with co-authors Alan C. Bovik and Lawrence K. Cormack in [33]. Dr. Bovik and Dr. Cormack helped build the database used in this research, and provided insightful comments on technical details.

and correlation between color and depth need further examination. Towards obtaining a better understanding of the statistical relationships between color and range, we studied both the marginal and joint statistics of color and range using the co-registered color images and ground-truth range maps in the LIVE Color+3D Database Release-1 [7]. To better approximate color perception in human vision systems, all color images in RGB were transformed into the more perceptually relevant CIELAB color space. We use CIELAB since it is optimized for quantifying perceptual color difference and it better corresponds to human color perception than does the perceptually nonuniform RGB space [89].

An important stereoscopic cue, disparity, comes from the angular difference between the two different retinal images received by the two frontally placed, horizontally separated eyes. It has been verified that there exist simple and complex neurons tuned to binocular disparity in V1 [70, 90], and the human vision system has very fine stereo acuity, which falls between 2 arcsec to 6 arcsec under the best conditions [91]. The visual system also has a large upper disparity limit, which reaches 7° for crossed disparities and 12° for uncrossed disparities [92]. The excellent acuity and broad operating range of stereopsis indicate that disparity is extensively used for depth perception. Liu *et al.* [93] studied the disparity distributions of natural scenes by converting forest range maps to disparity maps. They found that the disparity distributions at eye level are centered at zero, non-Gaussian, but well modeled as generalized Gaussian. A similar study on indoor range maps showed similar

results. Moreover, the authors correlated disparity sensitivity with naturally available disparities by showing that the proportion of near- and far-tuned disparity distributions qualitatively agrees with the distribution of disparity-tuning neurons in V1 [94]. This suggests that the human vision system may use the rich disparity cues both in near- and far-viewing distances to recover the depth information in natural scenes.

4.2 Data Pre-processing

Human vision systems extract abundant information from natural environments by processing visual stimuli through different levels of decomposition and interpretation. Since we want to learn and explore the statistical relationships between luminance/chrominance and range/disparity and how these statistics might be implicated in visual processing, and subsequently used in image processing algorithms, some pre-processing was performed on both the 2D color images and the co-registered 2D ground-truth range maps.

4.2.1 Color Space Conversion and Gabor Filter Bank

All color images were transformed into the perceptually relevant CIELAB color space having one luminance (L*) and two chrominance (a* and b*) components. CIELAB color space is optimized to quantify perceptual color differences and better corresponds to human color perception than does the perceptually nonuniform RGB space [89]. The coordinate L* of the CIELAB space represents the lightness of the color, a* represents its position between

red/magenta and green, and b* represents its position between yellow and blue. Moreover, the nonlinear relations for L*, a*, and b* mimic the nonlinear responses of human eyes, starting from the cone cells (L, M, and S) in the retina. Each image was then decomposed by a 2D Gabor filter bank over multiple scales and orientations, which serves to mimic the receptive fields of simple cells in V1 [14, 95–97]. Both the luminance and chrominance components of the transformed color images and the converted disparity maps were filtered by the same 2D Gabor filter bank.

Before discussing the statistical analysis and modeling, we wish to briefly motivate them by discussing the formation of receptive fields in V1 neurons and their relevance to understanding natural scene statistics. From physiological evidence [98], it is known that the simple cells in V1 process visual signals received from LGN (Lateral Geniculate Nucleus) neurons. The simple cells can be linearly modeled as having elongated, center-surround receptive fields that are highly selective in spatial frequency and orientation, remarkably like a Gabor filter. Thus, the physical statistics of the natural environment are manifested in the spectral responses of neurons in the visual cortex [18, 19], and likely in the responses of disparity-tuned neurons as well [93].

A complex 2-D Gabor filter can be written

$$\begin{aligned} &G(x, y, \sigma_1, \sigma_2, \zeta_x, \zeta_y, \theta) \\ &\qquad = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2}\left[\left(\frac{R_1}{\sigma_1}\right)^2 + \left(\frac{R_2}{\sigma_2}\right)^2\right]} e^{i(x\zeta_x + y\zeta_y)} \end{aligned} \tag{4.1}$$

where $R_1 = x\cos\theta + y\sin\theta$ and $R_2 = -x\sin\theta + y\cos\theta$, σ_1 and σ_2 are the

standard deviations of an elliptical Gaussian envelope along the rotated axes, ζ_x and ζ_y are the spatial center frequencies of the complex sinusoidal carrier, and θ is the orientation.

Since physiological evidence shows that visual neurons in primary visual cortex usually have an elliptical Gaussian envelope with an aspect ratio of 0.25-1.0, with propagating direction along the short axis of the elliptical Gaussian envelope [99], we use complex 2-D Gabor filters of the form

$$G(x, y, \gamma, \sigma, \omega, \theta) = \frac{1}{2\pi\gamma\sigma^2} e^{-\frac{1}{2}\left[\left(\frac{R_1}{\sigma}\right)^2 + \left(\frac{R_2}{\gamma\sigma}\right)^2\right]} e^{i\omega R_1} \tag{4.2}$$

where $\gamma = \sigma_y/\sigma_x$ is the aspect ratio of the elliptical Gaussian envelope, $\sigma = \sigma_x$, and $\omega = \sqrt{\zeta_x^2 + \zeta_y^2}$ is the radial center frequency. To create a suitable set of Gabor filter banks which can cover most of the frequency domain, the two parameters of the elliptical Gaussian envelope need to be chosen properly, including the aspect ratio, γ, and the standard deviation, σ [100]. Here, six spatial center frequencies, 0.84, 1.37, 2.22, 3.61, 5.87, and 9.53 (cycles/degree) are used, with four different sinusoidal grating orientations for each spatial frequency: horizontal (0-deg), diagonal-45 (45-deg), vertical (90-deg), and diagonal-135 (135-deg) [101, 102]. The aspect ratio, γ, is chosen to be 1.0 [99]. The spatial frequency bandwidth of each sub-band is 0.7 (octave), and neighboring filters intersect at half-power point, i.e. 3-dB point [101, 103].

4.2.2 Range and Disparity

In the following sections, we study the statistics of decomposed range data and explore the statistical relationship of range with decomposed lumi-

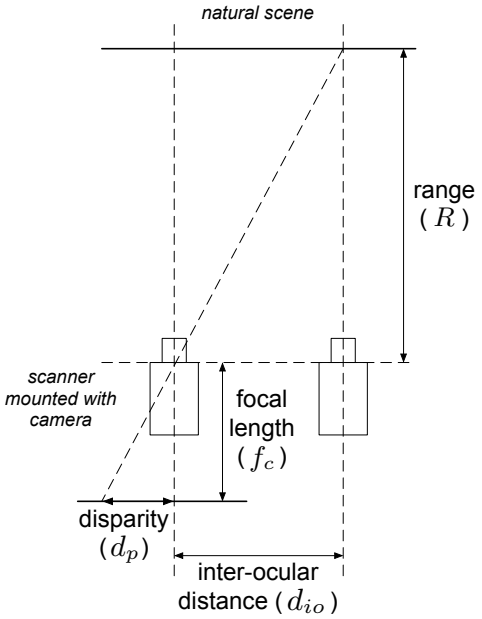


Figure 4.1: Geometry of the parallel-viewing model.

nance and chrominance in natural images. First, we examine the conditional distribution of range gradient data given luminance/chrominance. Since the depth information acquired by the human vision system is more relative than absolute, which means that we know which objects are further and which are closer, but we are not sure about the exact distance of each object from us, the disparity serves to be an important stereoscopic cue and its statistics is of most interest towards understanding depth perception. Therefore, in order to be able to examine statistical correlations between range/depth and luminance/chrominance information in the presence of stereoscopic fixation, we also convert ground-truth range maps into disparity maps under a parallel-viewing model. Figure 4.1 depicts the geometry of the parallel-viewing model.

The disparity values are computed as:

$$\frac{d_p}{f_c} = \frac{d_{io}}{R} \Rightarrow d_p = f_c \frac{d_{io}}{R} \tag{4.3}$$

where d_p is disparity, f_c is the focal length of the camera, d_{io} is the inter-ocular distance, and R is ground-truth range. Finally, the converted disparity maps are decomposed by the same multi-scale, multi-orientation 2D Gabor filter bank.

4.3 Statistical Analysis

The statistics of 2D and 3D natural scenes have previously been learned by the human visual apparatus over the eons. These powerful, physically and perceptually relevant constraints form priors which can be applied to solve visual tasks. Since acquiring geometric knowledge about the surrounding 3D environment is a basic element of human visual activity, accurate perception and consistent interpretation of natural range/depth information is an essential processing role of the early visual processing pathway.

Towards understanding the statistical basis of such computations, we first examine the marginal statistics of luminance and chrominance Gabor responses, and the conditional statistics of range gradients given measurements of these responses. Our analysis is performed on the (demodulated) magnitude responses of the Gabor quadrature functions, expressed as rms values of the sine and cosine responses [14]. Based on these measurements, we form models of the prior marginal and conditional distributions towards leveraging them in

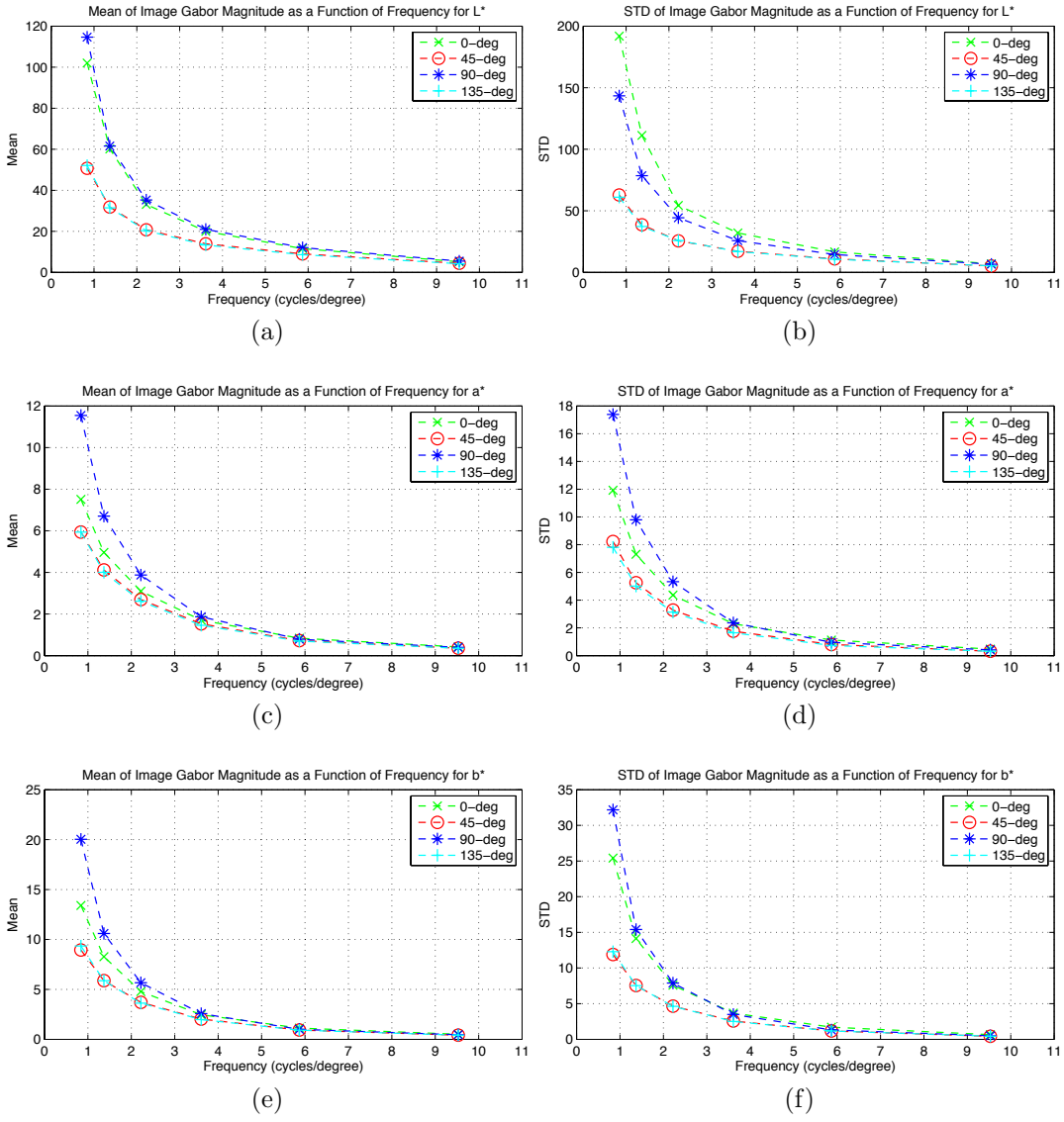


Figure 4.2: The mean and standard deviation (STD) of Gabor magnitude responses against radial spatial frequency: (a) mean for L*, (b) STD for L*, (c) mean for a*, (d) STD for b*, (e) mean for b*, and (f) STD for b*.

solving model-based visual processing problems.

4.3.1 Marginal Statistics

All color images were first transformed into the perceptually relevant CIELAB color space, then decomposed by a multi-scale, multi-orientation 2D Gabor filter bank. This serves the dual role of supplying optimally conjoint spatio-spectral decompositions of the data, while also providing a reasonable approximation of area V1 responses. As a first point of study, we computed the mean and standard deviations of the luminance and chrominance Gabor magnitude responses against spatial frequency and orientation. Specifically, we found the Gabor magnitude responses on the L* channel for all 12 test images in the LIVE Color+3D Database Release-1. We plot mean and standard deviation of these responses as a function of spatial frequency and orientation in Figs. 4.2 (a) and (b). The same computation was performed on the a* and b* channels as well, and is plotted in Figs. 4.2 (c) to (f).

As generally expected, mean magnitude responses fall off approximately as $1/f^2$ with spatial frequency f. This agrees with findings that the power spectra of natural images varies as $1/f^2$ with spatial frequency [14, 15, 104]. This fundamental model has been successfully used to explain and predict certain early stages of visual processing in insects and higher vertebrates [20, 105–107]. An interesting observation is that the curves for the diagonal-45 and diagonal-135 orientations nearly overlap in all three channels. For the L* channel, the curves for the horizontal and vertical orientations also overlap, while for the a* and b* channels, the curves are distinct.

The standard deviations of the magnitude responses also follows a $1/f^2$

shape for both luminance and chrominance channels. As observed for the mean magnitude responses, the curves for diagonal orientations overlap across all three channels. However, the standard deviation curves for horizontal and vertical orientations are distinct for both luminance and chrominance channels.

The $1/f^2$ distribution of mean and standard deviation of the Gabor responses implies equal energy within equal (octave) bandwidths, and also equal variation of energy within equal bands over different orientations in natural images. Moreover, the distribution of spectral energy contained in the luminance channel is different from that carried by the chrominance channels in natural environments. These findings can potentially be utilized to better understand and explain various stages of visual processing.

4.3.2 Conditional Statistics

It may be observed from natural scenes that there is substantial co-occurrence of luminance/chrominance edges and range/depth discontinuities. For example, if there is a discontinuity in the range/depth map, it is highly likely that an edge of the same orientation is co-located in the corresponding color image. However, the intuition of the other direction is rather weaker; many image edges must exist without corresponding discontinuities in range due to the plethora of shadows and textures in the natural environment. In order to examine the relationship between chromatic image and range singularities, we studied the conditional statistics of range gradients given Gabor image responses.

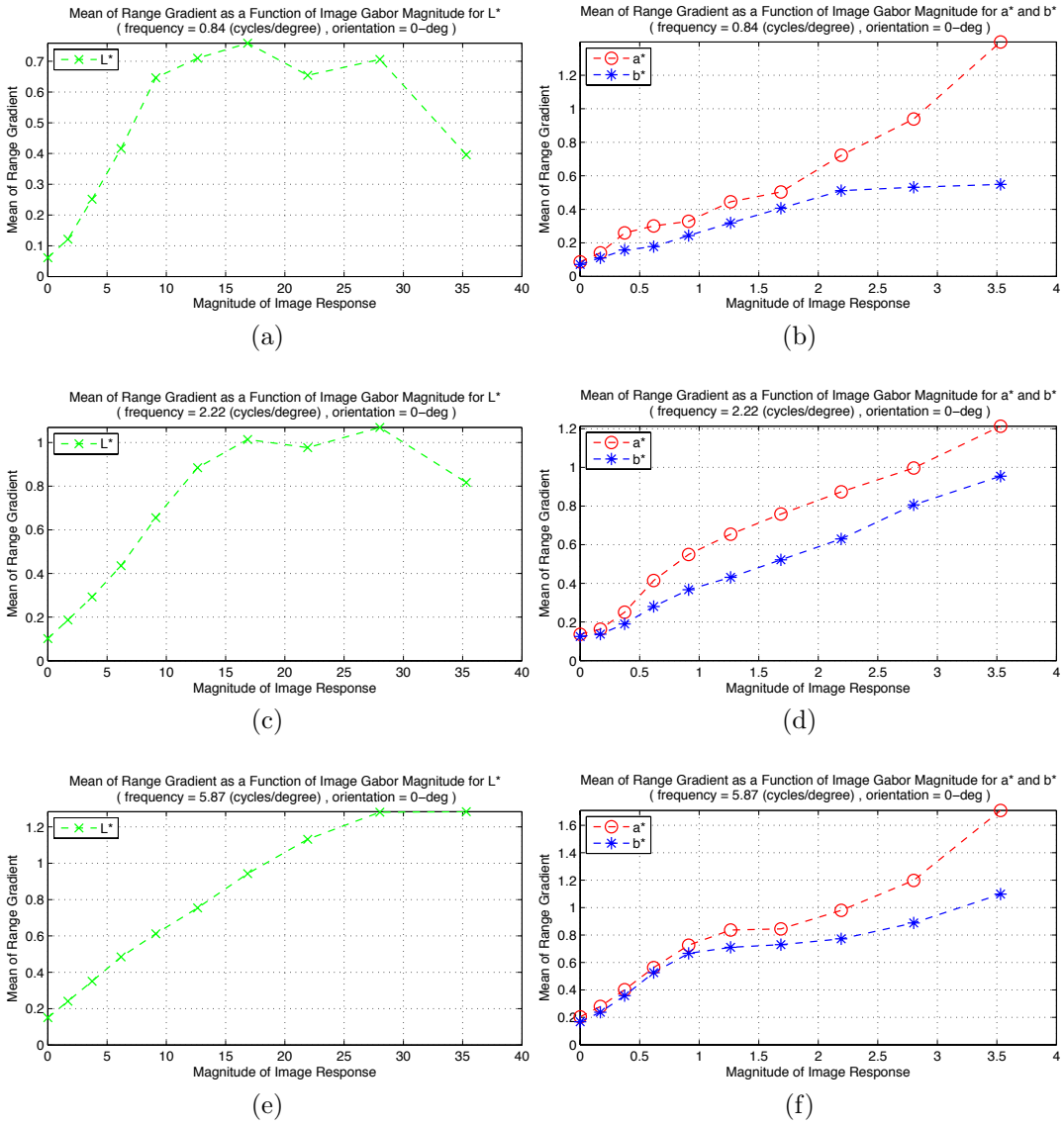


Figure 4.3: The means of range gradient magnitudes against Gabor magnitude responses over different spatial frequencies with the same horizontal (0-deg) orientation: (a), (c), and (e) for L*; (b), (d), and (f) for a* and b*.

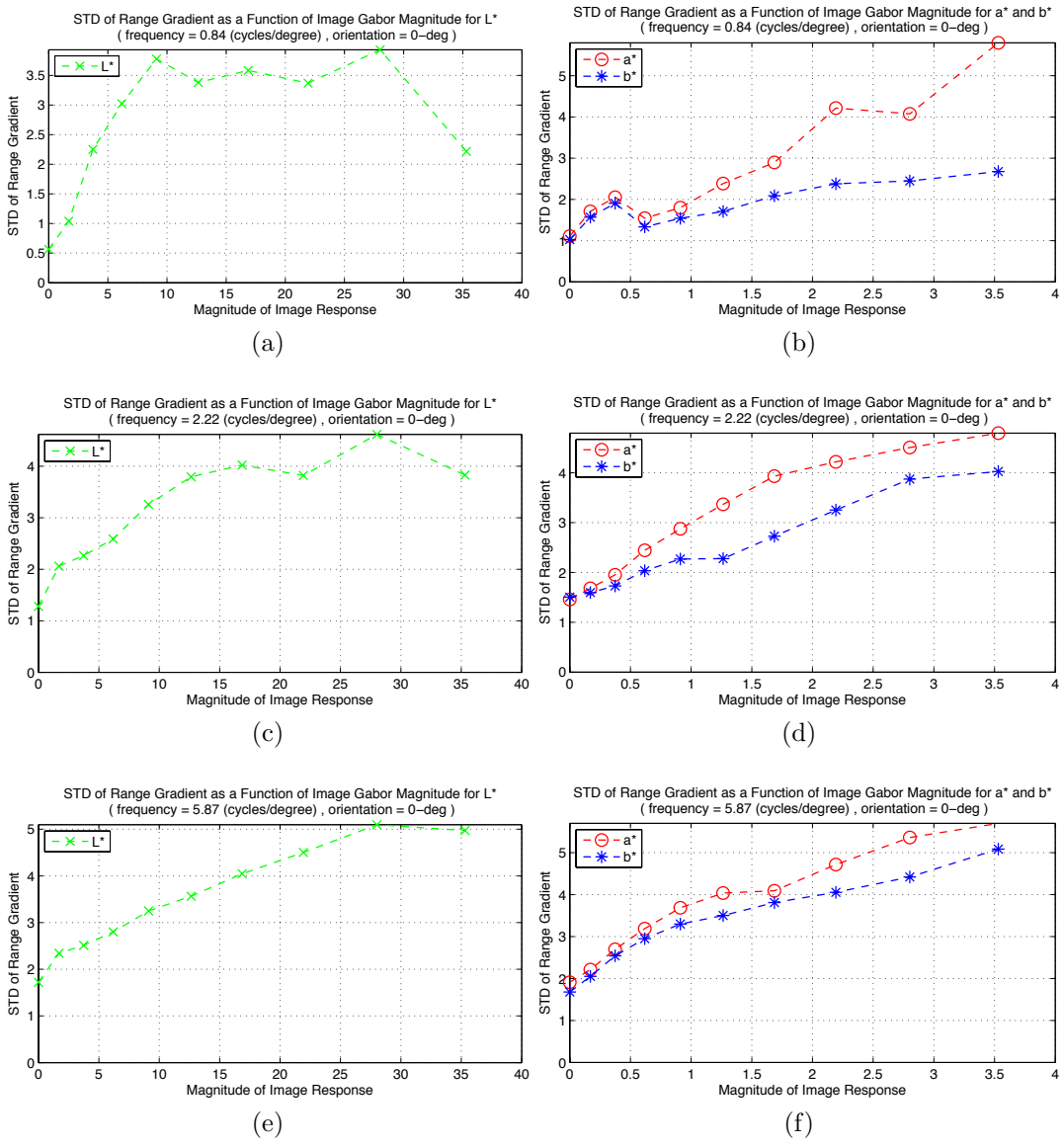


Figure 4.4: The standard deviations of range gradient magnitudes against Gabor magnitude responses over different spatial frequencies with the same horizontal (0-deg) orientation: (a), (c), and (e) for L*; (b), (d), and (f) for a* and b*.

Again, all color images were first transformed into the perceptually relevant CIELAB color space, then passed through the multi-scale, multi-orientation 2D Gabor filter bank, and the magnitude responses computed. We also computed the gradient magnitude matrix, R_g, of each range map, R, for each corresponding color image in the database, which is given by

$$R_g(i,j) = \|\nabla R(i,j)\| = \sqrt{\left(\frac{\partial R(i,j)}{\partial x}\right)^2 + \left(\frac{\partial R(i,j)}{\partial y}\right)^2} \tag{4.4}$$

where

$$\begin{aligned} \nabla R(i,j) &= \left[\frac{\partial R(i,j)}{\partial x}, \frac{\partial R(i,j)}{\partial y}\right]^T \\ \frac{\partial R(i,j)}{\partial x} &= \frac{R(i+1,j) - R(i-1,j)}{2} \\ \frac{\partial R(i,j)}{\partial y} &= \frac{R(i,j+1) - R(i,j-1)}{2} \end{aligned} \tag{4.5}$$

To obtain conditional statistics, we binned the magnitude responses across all images for each luminance and chrominance channel. Within each bin, we also collected the corresponding gradient magnitudes for all range maps. Finally, we computed the conditional mean and standard deviations of magnitude of the range gradients given the Gabor magnitude responses for the L*, a*, and b* channels. Figures 4.3 and 4.4 plot the mean and standard deviation, respectively, of the range gradient magnitudes against Gabor magnitude responses for luminance and chrominance channels for an exemplar sub-band. Very similar curves and results are observed at different sub-bands.

The six panels in Fig. 4.3 plot the conditional statistics of mean range gradient magnitude given magnitude responses of horizontal Gabors at three

different spatial frequencies. For the luminance channel, the range gradient magnitude increases monotonically with small Gabor responses, but saturates with larger Gabor responses over all frequencies. On the other hand, the magnitude of range gradients increases monotonically with Gabor magnitude responses at all frequencies for both chromatic channels, and there is no saturation for large Gabor responses. Similar trends are observed for the standard deviation of range gradients conditioned on the magnitude of horizontal Gabor responses at three different spatial frequencies, as shown in Fig. 4.4. These monotonic relationships between range gradients and luminance and chrominance Gabor responses demonstrate a high correlation between these quantities, while also strengthening the intuition that, if there are strong variations in a natural image, i.e. large Gabor responses, there is a high likelihood of co-located large variations, i.e. large range gradients, in the corresponding range map. Moreover, the luminance channel carries information that is different from that carried by the chrominance channels in the sense that the means and standard deviations of range gradient magnitudes both saturate given large luminance Gabor magnitude responses, which implies that the chromatic components in natural images can also possibly be utilized in depth perception, a concept that is supported by our prior human study [86].

To further validate the existence of strong correlations between range gradients and image Gabor responses, we performed a simple hypothesis test on the sample correlation coefficients between range gradients and luminance/ chrominance Gabor responses using the same sub-band as in Fig. 4.3(c) and

Table 4.1: Hypothesis Test of Correlation on Conditional Statistics (frequency = 5.87 (cycles/degree) and orientation = 0-deg)

Channel	$\lvert t$-score$\rvert$	p-value (10^{-3})	Decision
L*	3.342	0.205	Reject H_0
a*	2.933	7.681	Reject H_0
b*	3.343	0.357	Reject H_0

4.4(c). The t-score is given by

$$t = \frac{r\sqrt{n-2}}{1-r^2} \tag{4.6}$$

where r is the sample correlation coefficient and n is the number of samples. Since there are millions of points within each sub-band, we iterated 100 times, taking 1000 random samples per iteration to compute the correlation coefficient. The final t-score was obtained by finding the average correlation coefficient over 100 iterations with the sample size n equal to 1000. Table 4.1 lists the t-scores of the color channels and the corresponding decisions using a two-sided level of significance $\alpha = 0.05$. It can be seen from Table 4.1 that the null hypothesis, H_0, that there is no correlation between range gradients and image Gabor responses was rejected for all three channels. Very similar results were obtained for all other sub-bands. Note that this corresponds to a very conservative test. Testing the full data set would yield minuscule p-values because of the large n. Rather, we did the test using approximately the number of samples that would be available instantaneously over about 0.005 (mm^2) on the primate retina (or about 0.056 (deg^2) of visual angle).

4.4 Statistical Models

The statistical analysis described thus far discovers a link between range/depth variations and co-located luminance and chrominance variations in natural images, and by extension, in the neural responses of luminance/ chrominance in primary visual cortex (V1). In this section, we seek to quantitatively model the statistical relationships between co-located luminance/ chrominance and range/depth information in natural images. The Gabor magnitude responses were computed as in the previous analysis on both luminance and chrominance channels. To acquire statistics of the important stereoscopic cue, disparity, used in the perception of depth by the human vision system, the ground-truth range maps were converted into disparity maps using the parallel viewing model described in Section 4.2.2. These converted disparity maps were also subjected to a multi-scale, multi-orientation Gabor decomposition, from which the disparity Gabor magnitude responses were computed. Since we want to derive statistical models that relate disparity and color image data, and demonstrate their usefulness, the marginal distributions of the Gabor magnitude responses to luminance/chrominance and disparity, and the conditional distributions of luminance/chrominance given disparity Gabor magnitude responses are of great interest.

4.4.1 Marginal Distributions

In order to examine the marginal distributions of luminance, chrominance, and disparity processed by different sub-bands, the Gabor responses

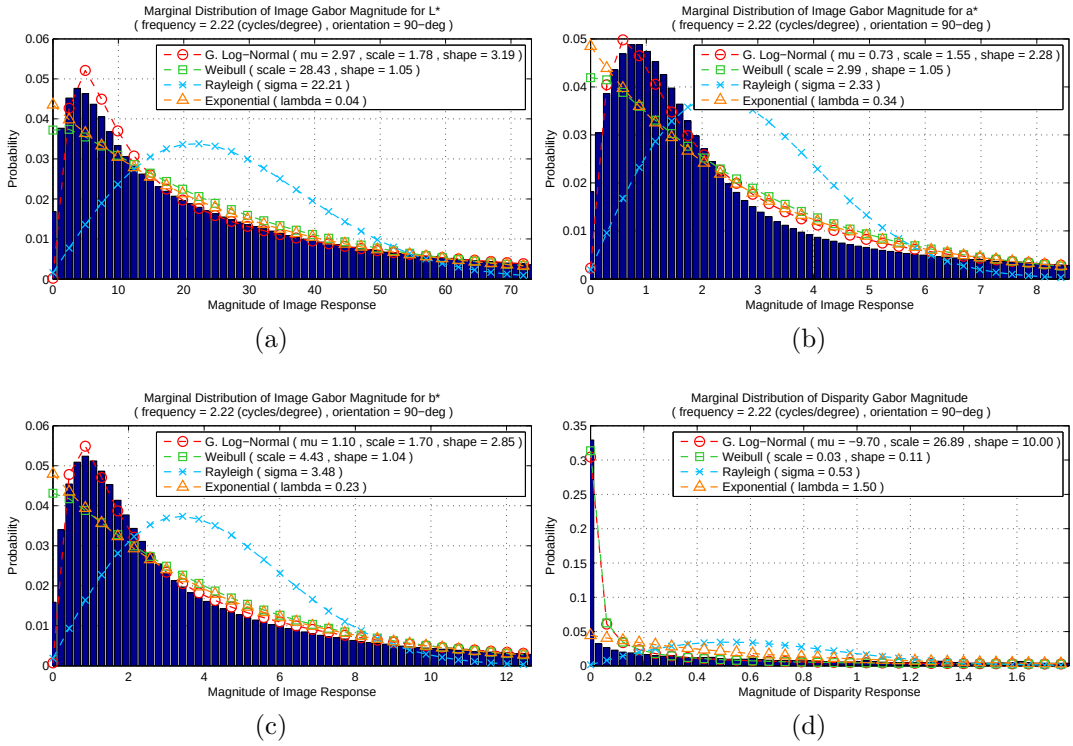


Figure 4.5: Marginal distributions of Gabor magnitude responses at one sub-band for: (a) L* , (b) a*, (c) b*, and (d) disparity.

Table 4.2: Comparison of Marginal Distribution Fits using Sum of Squared Error (10^{-2})

	Marginal Distribution			
	L*	a*	b*	Disparity
Exponential Fit	0.116	0.245	0.222	8.569
Rayleigh Fit	1.143	1.141	1.292	12.008
Weibull Fit	0.103	0.202	0.197	0.823
G. Log-Normal Fit	**0.078**	**0.057**	**0.041**	**0.785**

were first collected across all scenes in the database. The empirical marginal distributions for all quantities of interest within each sub-band were obtained

as histograms computed by binning all of the Gabor magnitude responses within each channel at that sub-band.

Figure 4.5 shows the marginal distributions of the luminance, chrominance, and disparity Gabor magnitude responses for one sub-band. The circle-dotted, square-dotted, cross-dotted, and triangle-dotted lines depict the best (least-squares) generalized log-normal, Weibull, Rayleigh, and exponential distribution fits, respectively, to each marginal distribution. In particular, the generalized log-normal distribution is given by

$$p_g(x) = \begin{cases} \frac{\beta_g}{2x\alpha\Gamma(\frac{1}{\beta_g})} \exp\left[-(\frac{|\ln(x)-\mu_g|}{\alpha_g})^{\beta_g}\right] & , x \geq 0 \\ 0 & , x < 0 \end{cases} \tag{4.7}$$

where $\Gamma(\cdot)$ is the gamma function, μ_g, α_g, and β_g are the location, scale, and shape parameters, respectively. The general Weibull distribution is given by

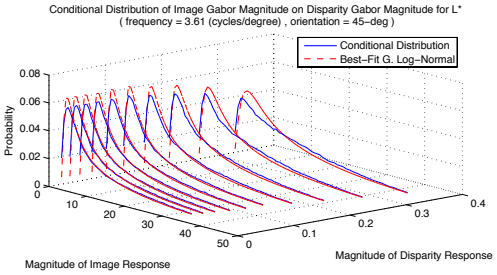
$$p_w(x) = \begin{cases} \frac{\beta_w}{\alpha_w}(\frac{x}{\alpha_w})^{(\beta_w-1)} e^{-(\frac{x}{\alpha_w})^{\beta_w}} & , x \geq 0 \\ 0 & , x < 0 \end{cases} \tag{4.8}$$

where α_w and β_w represent the scale and shape parameters, respectively, which allows the model to include exponential ($\beta_w = 1$) and Rayleigh ($\beta_w = 2$) distributions as special cases depending on the shape parameter. The characteristic shape of the marginal distributions is quite different from the symmetric Gaussian-like distributions used in other studies. These previous models have captured the statistics of the luminance channel of natural images using band-pass wavelet filter-banks, e.g., steerable pyramid decompositions, but without finding the magnitude (envelope) responses. Here we have used the Gabor filter-bank to match the receptive fields of simple neurons in primary

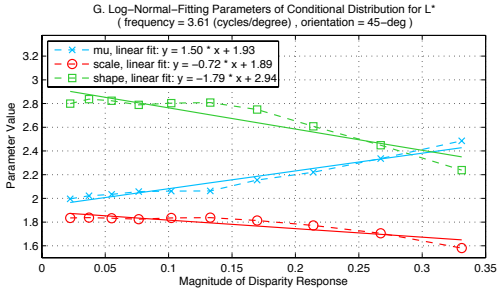
visual cortex (V1), and computed the Gabor magnitude responses to mimic the energy exchange in neural signal communication. It can be seen that the generalized log-normal fits better capture the shapes of all four marginal distributions, while the Weibull and exponential fits are not able to match the positive-skewed bell shapes, and the Rayleigh fits fail to model the heavy tails. In other words, with the three model parameters, location, scale and shape, the generalized log-normal function can flexibly adjust both its peak location and its variance, i.e., the width of distribution, to better fit the characteristic shape of the marginal distribution of image and disparity magnitude responses. We also performed a numerical comparison of different distribution fits at the same sub-band using the sum of squared error, as shown in Table 4.2. In accordance with the visual comparison in Fig. 4.5, the generalized log-normal functions yield the best fits among all four marginal distributions, i.e. fixing the parameters gives much worse fits. Note that the marginal distributions of the filtered luminance, chrominance, and disparity over different sub-bands all share similar shapes. For reference, we list the best-fit generalized log-normal parameters for the marginal distributions of luminance, chrominance, and disparity Gabor magnitudes at all sub-bands in [7].

4.4.2 Conditional Distributions

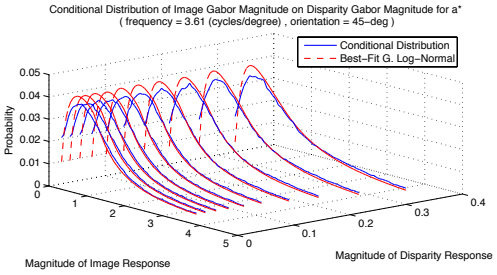
Similarly, the conditional distributions of luminance and chrominance given disparity at different sub-bands were obtained by first computing and collecting the filtered luminance, chrominance, and disparity magnitude responses



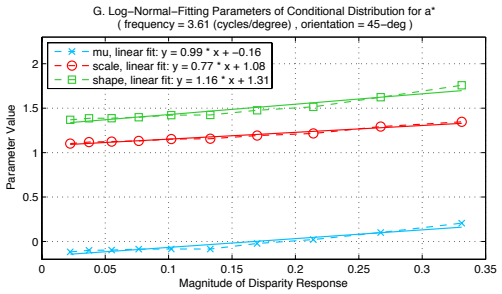
(a) Distribution for L* channel



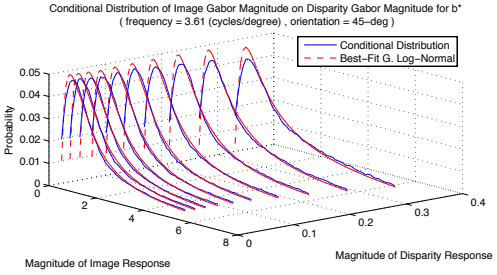
(b) Best-fit generalized log-normal parameters for L* channel



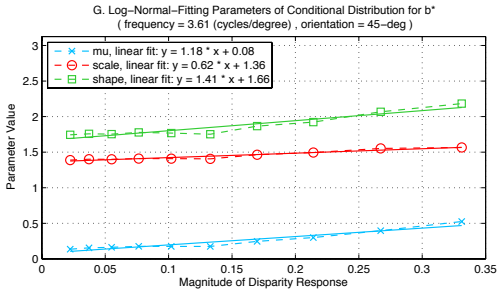
(c) Distribution for a* channel



(d) Best-fit generalized log-normal parameters for a* channel



(e) Distribution for b* channel



(f) Best-fit generalized log-normal parameters for b* channel

Figure 4.6: Conditional distributions of filtered luminance (L*) and chrominance (a* and b*) magnitudes given filtered disparity magnitudes from one sub-band: (a), (c), and (e) the conditional distributions (solid lines) along with the best-fit generalized log-normal models (dotted lines); (b), (d), and (f) their corresponding best-fit generalized log-normal parameters.

across all scenes, then computing the histograms for each sub-band. For each sub-band, the conditional histograms of luminance and chrominance given disparity were computed by first binning the disparity magnitude responses, then binning the filtered luminance/chrominance magnitude responses within each disparity bin. Figure 4.6 shows the conditional distributions of all three luminance and chrominance components (solid lines), as well as their corresponding best-fit generalized log-normal distributions (dotted lines) and their model parameters. It can be seen that the conditional distributions of luminance and chrominance given disparity are well-fitted by the generalized log-normal distribution. As discussed in Section 4.4.1, the flexible generalized log-normal function is a better model than other specific fits, such as Weibull, exponential, and Rayleigh functions. For the conditional distributions of luminance given disparity, the location parameter (μ_g) of the fitted generalized log-normal model increases monotonically and linearly as the disparity magnitude response increases, while both the scale (α_g) and shape (β_g) parameters decrease monotonically and linearly across the disparity magnitude responses. On the other hand, for the conditional distributions of chrominance given disparity, all three parameters exhibit a nearly linear relationship with the disparity magnitude responses. In general, as the disparity magnitude response increases, the conditional distributions of both luminance and chrominance become more heavy-tailed, which implies that if there is a large disparity variation, i.e. a large range/depth discontinuity, large luminance and chrominance variations are highly likely to be co-located in the corresponding

color images. These monotonic correlations between the Gabor responses of disparity and luminance/chrominance in natural images confirm the observations as well as the computed conditional statistics between range and luminance/chrominance variations discussed in Section 4.3.2. Moreover, the linear relationships between the parameters of the generalized log-normal model and the magnitude of disparity Gabor responses nicely captures the heavy-tailed conditional distributions of both filtered luminance and chrominance channels. Next, we will leverage these new joint NSS models to solve an exemplar 3D image processing problem: binocular correspondence.

4.5 Application to A Chromatic Bayesian Stereo Algorithm

Given a pair of left and right images, a binocular stereo algorithm computes a disparity map from one image to the other. The basic idea is to minimize an energy functional which captures differential binocular cues between left and right images within an optimization framework [34]. A Bayesian stereo algorithm is able to adapt a likelihood (conditional distribution) and a prior (marginal distribution) of natural scene statistics (NSS) within the energy functional to be minimized, thus forcing the solution to be consistent with the observed statistical relationships that occur between luminance, chrominance, and disparity data in natural scenes, as derived in Section 4.4. Given a pair of left and right images, I_l and I_r, then to estimate the disparity map, D, from the right (matching) to the left (reference) image, the canonical Bayesian

stereo formulation takes the form [108]

$$\begin{aligned} D &= \underset{D'}{\operatorname{argmax}}\, P(D'|(I_l, I_r)) \\ &= \underset{D'}{\operatorname{argmax}}\, P((I_l, I_r)|D')P(D') \end{aligned} \tag{4.9}$$

where $P(D'|(I_l, I_r))$ is the posterior probability to be maximized, and $P((I_l, I_r)|D')$ and $P(D')$ are the likelihood and prior probabilities, respectively. Taking the logarithm of the product of the likelihood and prior, the Bayesian formulation corresponds to minimization of the energy function:

$$D = \underset{D'}{\operatorname{argmin}}\, E_p + \lambda E_s \tag{4.10}$$

where E_p is the photometric energy expressed by the likelihood $P((I_l, I_r)|D')$, E_s is a smoothness term derived from the prior $P(D')$, and λ is a weight. Note that E_p can encapsulate all three luminance and chrominance components, L^*, a^*, and b^*, and be written

$$E_p = \sum_{i,j} \sum_{k \in \{L^*, a^*, b^*\}} |I_{l_k}(i, (j - D'(i,j))) - I_{r_k}(i,j)| \tag{4.11}$$

To incorporate the marginal and conditional NSS distributions that we have measured and modeled, the Bayesian stereo formulation can be re-written as

$$\begin{aligned} D &= \underset{D'}{\operatorname{argmax}}\, P(\tilde{D}'|(I_l, I_r), \tilde{I}_l) \\ &= \underset{D'}{\operatorname{argmax}}\, P((I_l, I_r)|\tilde{D}', \tilde{I}_l)P(\tilde{I}_l|\tilde{D}')P(\tilde{D}') \qquad (4.12) \\ &= \underset{D'}{\operatorname{argmin}}\, E_p + \lambda(E_{NSS_c} + E_{NSS_m}) \qquad (4.13) \end{aligned}$$

(taking logarithm of (4.12)), where $\tilde{I}_l$ and $\tilde{D}'$ are the magnitudes of the Gabor filtered responses of I_l and D', respectively, E_p is the photometric energy derived from $P((I_l, I_r)|\tilde{D}', \tilde{I}_l)$, E_{NSS_c} and E_{NSS_m} are energy terms related to the conditional and marginal NSS distributions, respectively, and λ is the constant weight.

Finally, since both the marginal distribution of disparity and the conditional distributions of luminance and chrominance given disparity can be modeled as generalized log-normal, the complete formulation of the proposed Bayesian stereo algorithm incorporating the NSS models can be written

$$D = \underset{D'}{\operatorname{argmin}} \sum_{i,j} \left[\sum_{k \in \{L^*, a^*, b^*\}} \left(E_{p,k} + \lambda_k E_{NSS_{c,k}} \right) + \lambda_m E_{NSS_m} \right] \tag{4.14}$$

where by introducing Eq. (4.7) and (4.11)

$$E_{p,k} = |I_{l_k}(i, (j - D'(i,j))) - I_{r_k}(i,j)| \tag{4.15}$$

$$E_{NSS_{c,k}} = \ln(\tilde{I}_{l_k}(i,j)) + \ln(\frac{2\alpha_k \Gamma(\frac{1}{\beta_k})}{\beta_k}) + (\frac{|\ln(\tilde{I}_{l_k}(i,j)) - \mu_k|}{\alpha_k})^{\beta_k} \tag{4.16}$$

$$E_{NSS_m} = \ln(\tilde{D}'(i,j)) + \ln(\frac{2\alpha_{\tilde{D}'} \Gamma(\frac{1}{\beta_{\tilde{D}'}})}{\beta_{\tilde{D}'}}) + (\frac{|\ln(\tilde{D}'(i,j)) - \mu_{\tilde{D}'}|}{\alpha_{\tilde{D}'}})^{\beta_{\tilde{D}'}} \tag{4.17}$$

where μ_k, α_k, and β_k are the location, scale, and shape parameters, respectively, of the best-fit generalized log-normal distributions of filtered luminance and chrominance conditioned on filtered disparity, $\mu_{\tilde{D}'}$, $\alpha_{\tilde{D}'}$, and $\beta_{\tilde{D}'}$ are the

location, scale, and shape parameters of the best-fit generalized log-normal distribution of filtered disparity, respectively, and λ_k and λ_m are their corresponding constant weights. Note that the three parameters, μ_k, α_k and β_k, can be further linearly modeled with the disparity Gabor magnitudes, as illustrated in Fig. 4.6:

$$\mu_k = m_{\mu,k}\tilde{D}' + b_{\mu,k} \tag{4.18}$$

$$\alpha_k = m_{\alpha,k}\tilde{D}' + b_{\alpha,k} \tag{4.19}$$

$$\beta_k = m_{\beta,k}\tilde{D}' + b_{\beta,k} \tag{4.20}$$

where $m_{\mu,k}$, $m_{\alpha,k}$, and $m_{\beta,k}$ are the slope parameters for μ_k, α_k, and β_k, respectively, and $b_{\mu,k}$, $b_{\alpha,k}$, and $b_{\beta,k}$ are the corresponding offset parameters. To solve the optimization of the proposed Bayesian stereo algorithm, we apply simulated annealing on the derived energy function (4.14) [109].

4.6 Experimental Results and Discussion

We simulate and evaluate the proposed Bayesian stereo algorithm utilizing the derived NSS models on stereo image pairs from the widely used Middlebury database [34]. To demonstrate the effectiveness of the derived statistical models relating luminance/chrominance and disparity in natural scenes, we compared the computed disparity maps using the Bayesian stereo algorithm with related formulations and models, including the canonical formulation using (4.10), the NSS model proposed in [32], and the proposed luminance-chrominance-range NSS model (4.14). In [32], the authors derived

an NSS model using only luminance information in the wavelet domain, and incorporated only the conditional distribution of luminance given disparity into the Bayesian stereo algorithm. Using the proposed luminance-chrominance-range NSS model, we implement the Bayesian stereo algorithm using two formulations: one includes only the luminance component (L*), while the other includes both luminance and chrominance (a* and b*) components.

4.6.1 Comparison with Previous Models

Figures 4.7 to 4.10 show simulation results of the four stereo image pairs, *Tsukuba*, *Venus*, *Cones*, and *Teddy*, from the Middlebury database, including the original left and right images, the ground-truth disparity map, and the computed disparity maps obtained by the three different Bayesian formulations. Generally, computed disparity maps delivered by the stereo model embodying both luminance (L*) and chrominance (a* and b*) NSS priors are very close to the corresponding ground-truth disparity maps, retaining more details than the canonical formulation, and better adherence to smooth regions than the one computed by the previous NSS model. On *Tsukuba*, the canonical formulation scrubs regions, e.g., around the camera, while the algorithm using the previous luminance-only NSS model tends to "over-segment". The proposed algorithm using luminance-chrominance-range NSS priors is better able to find a balance between disparity smoothness and 3D detail with the aid of the additional regularity supplied by modeling the disparity and luminance/chrominance channels. For *Venus* and *Teddy*, it can be seen that the

canonical formulation fails to find binocular correspondences in some smooth regions without edges, while the previous luminance-only NSS model is able to solve those disparity ambiguities using luminance and disparity priors. By introducing conditional priors of chrominance given depth, the luminance-chrominance-range NSS model further improves the accuracy of the computed disparity map by cleaning up the smooth 3D surfaces. On *Cones*, both the canonical formulation and the previous luminance-only NSS model do a good job matching image details while maintaining disparity smoothness; yet, they are not able to find binocular correspondences around some of the occluded regions on the cones. However, the luminance-chrominance-range NSS model allows most of the binocular correspondences around those occluded regions to be successfully resolved.

In addition to visual comparison, we also conducted a quantitative evaluation to compare the performance of the stereo algorithm embodying the derived luminance-chrominance-range NSS model with the canonical formulation and the previous luminance-only NSS model. Tables 4.3 through 4.5 give numerical comparisons between the proposed model and the other two formulations using three different metrics: bad-pixel percentage, including overall, non-occluded, and textured, for all four test image pairs.

Bad-pixel percentage, P_{bp}, is a commonly used error metric to measure pixel-wise differences between computed and ground-truth depth maps [34,

Table 4.3: Comparison of Bayesian Stereo Algorithm under Different Natural Scene Models using Overall Bad-Pixel Percentage (%)

	Tsukuba	*Venus*	*Cones*	*Teddy*
Canonical Formulation	9.71	11.21	24.31	32.18
Previous NSS Model in [32]	6.45	5.34	20.78	23.35
Proposed NSS Model using only L*	5.19	2.55	18.86	20.58
Proposed NSS Model using L*, a*, and b*	**4.91**	**2.21**	**18.57**	**20.37**

Table 4.4: Comparison of Bayesian Stereo Algorithm under Different Natural Scene Models using Non-occluded Bad-Pixel Percentage (%)

	Tsukuba	*Venus*	*Cones*	*Teddy*
Canonical Formulation	7.78	9.79	12.72	23.34
Previous NSS Model in [32]	4.26	3.69	8.54	13.20
Proposed NSS Model using only L*	2.92	1.44	7.58	12.37
Proposed NSS Model using L*, a*, and b*	**2.64**	**1.18**	**7.35**	**12.15**

Table 4.5: Comparison of Bayesian Stereo Algorithm under Different Natural Scene Models using Textured Bad-Pixel Percentage (%)

	Tsukuba	*Venus*	*Cones*	*Teddy*
Canonical Formulation	4.77	3.85	12.49	19.82
Previous NSS Model in [32]	4.70	2.96	8.39	12.20
Proposed NSS Model using only L*	3.34	1.62	7.61	11.10
Proposed NSS Model using L*, a*, and b*	**3.30**	**1.41**	**7.39**	**10.93**

110]. It takes the form:

$$P_{bp} = \frac{1}{N_{\mathcal{S}}} \sum_{i,j \in \mathcal{S}} \left(|D_C(i,j) - D_G(i,j)| > \delta_D \right) \tag{4.21}$$

where D_C and D_G are the computed and ground-truth disparity maps, respectively, $\mathcal{S}$ is the image region over which P_{bp} is calculated, $N_{\mathcal{S}}$ is the number of pixels in $\mathcal{S}$, and δ_D is a threshold expressing disparity error tolerance. Here we use $\delta_D = 1.0$, which coincides with previously published work comparing stereo algorithms [111, 112]. The three metrics of bad-pixel per-

centage used here are distinguished by the different regions $\mathcal{S}$. The overall bad-pixel percentage in Table 4.3 is calculated over the entire disparity map, i.e. $\mathcal{S} = \{(i,j) \mid 1 \leq i \leq h, 1 \leq j \leq w\}$ where h and w are the height and width of the disparity map, respectively. For the non-occluded bad-pixel percentage, $\mathcal{S}$ is defined as the region that is not occluded in the matching image, i.e. pixels appearing in the reference image have correspondences in the matching image. Finally, the textured bad-pixel percentage is calculated only over regions where the intensity of image horizontal gradients is beyond some threshold, i.e. pixels belonging to prominent image details, edges, and texture in the reference image.

From Tables 4.3 through 4.5, it is apparent that the numerical results support the visual comparisons: the Bayesian stereo algorithm using the luminance-chrominance-range NSS model outperforms the other two methods in terms of all three different metrics of bad-pixel percentage. Taking *Venus* for example, with respect to all three bad-pixel percentage metrics, the proposed luminance-chrominance-range NSS model achieves more than 100% improvement over the previous NSS model, which, in turn, significantly improves on the canonical formulation. By observing the textured bad-pixel percentage on *Tsukuba* in Table 4.5, it is apparent that the Bayesian stereo algorithms using the canonical formulation and the previous NSS model deliver similar performance, while the proposed luminance-chrominance-range NSS model yields a bolstered Bayesian stereo algorithm that delivers a significantly more accurate disparity map. Moreover, on the complicated image pairs *Cones* and

Teddy, while the previous NSS model generates fairly good disparity maps on non-occluded and textured regions, the new NSS models further improve the results with fewer pixel errors, demonstrating the utility of the derived marginal and conditional models that serve to regularize the range/depth and luminance/chrominance statistics of the stereo solution on natural images.

4.6.2 Augmentation by Chrominance

In Tables 4.3 to 4.5, we also list numerical results from the proposed luminance-chrominance-range NSS model using only the luminance channel (L*). It can be seen that for all four image pairs, the results using both the luminance (L*) and chrominance (a* and b*) channels yields better performance than using only the luminance channel with respect to all three different performance metrics. For example, the vivid and diverse colored objects in *Teddy* increase the difficulty of finding binocular correspondences; however, the proposed luminance-chrominance-range NSS model is better able to solve the problem by exploiting the derived conditional model between the natural depth and chrominance channels, resulting in more accurate disparity maps with lower non-occluded and textured bad-pixel percentages. Based on this quantitative comparison, we may conclude that chromatic information not only augments the performance of Bayesian stereo algorithms, but could also play a useful role in human binocular visual perception. For example, stereo processing in human vision systems could possibly leverage the statistical relationship between chrominance and range/depth cues in natural images to

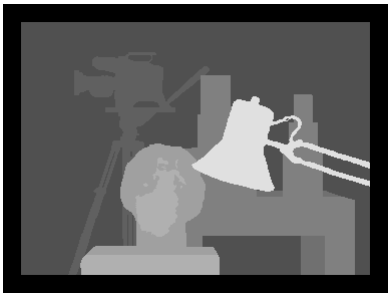
augment a variety of 3D visual tasks [86–88].

4.7 Summary

By utilizing high-resolution, high-quality color images and co-registered range maps in the LIVE Color+3D Database Release-1, we examined the statistical relationships between multi-scale, multi-orientation Gabor decompositions of luminance/chrominance and range/depth data in natural scenes. We showed that the marginal statistics of both image and range magnitude responses follow the well-known $1/f^2$ power law, and the conditional statistics of range gradients given image magnitude responses provide evidences supporting the co-occurrence of natural image and range variations. We further derived marginal and conditional priors relating natural luminance/chrominance and disparity, and demonstrated their efficacy with application to the Bayesian stereo algorithm. We also demonstrated that including the chrominance-range models augments the performance of the Bayesian stereo algorithm over using only the luminance information. More importantly, the superior performance incorporating color and range priors to previous luminance-only models bolsters the psychophysical evidence that not only image intensity, but also chromatic information is useful in 3D visual processing.

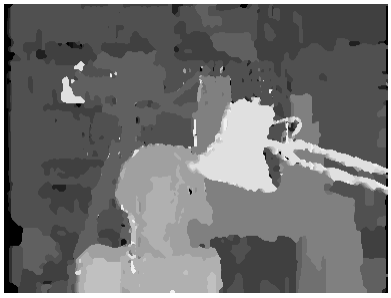
(a) Left image

(b) Right image

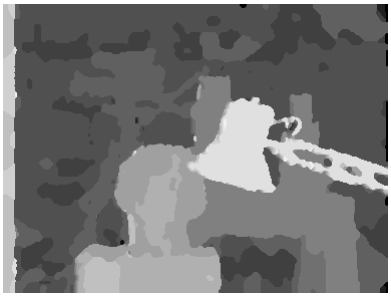


(c) Ground-truth disparity map

(d) Canonical formulation

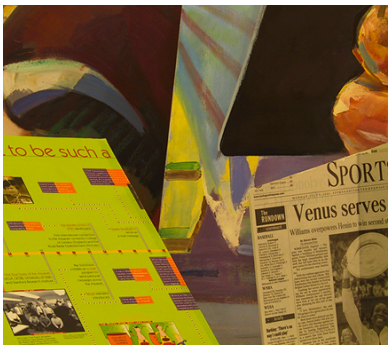


(e) Previous NSS model



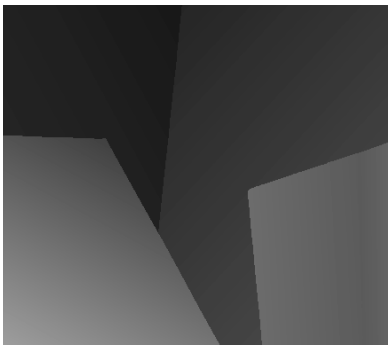
(f) Proposed luminance-chrominance-range NSS model with L*, a*, and b*

Figure 4.7: Simulation results on *Tsukuba* from the Middlebury database, including the original stereo image pair, the ground-truth disparity map, and disparity maps using computed Bayesian stereopsis under different NSS models.

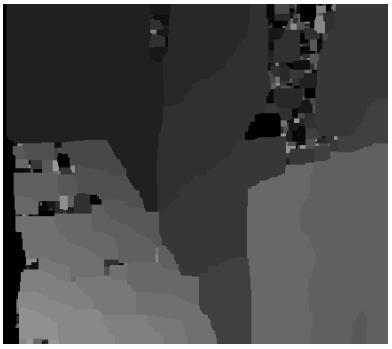


(a) Left image

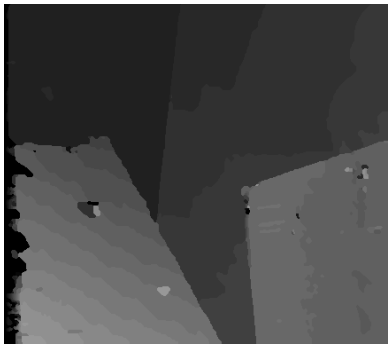
(b) Right image



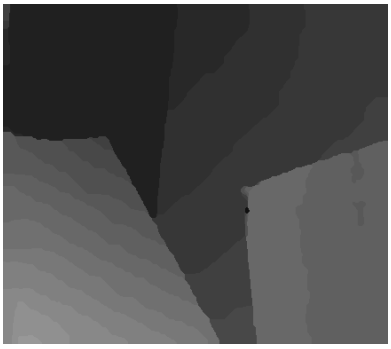
(c) Ground-truth disparity map



(d) Canonical formulation



(e) Previous NSS model

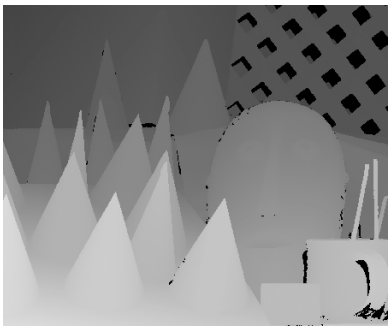


(f) Proposed luminance-chrominance-range NSS model with L*, a*, and b*

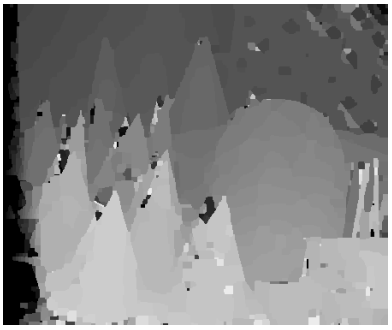
Figure 4.8: Simulation results on *Venus* from the Middlebury database, including the original stereo image pair, the ground-truth disparity map, and disparity maps using computed Bayesian stereopsis under different NSS models.

(a) Left image

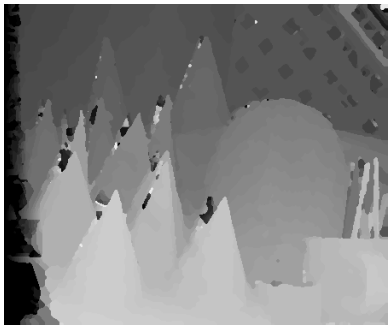
(b) Right image



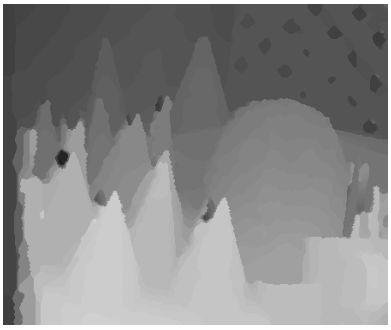
(c) Ground-truth disparity map



(d) Canonical formulation



(e) Previous NSS model

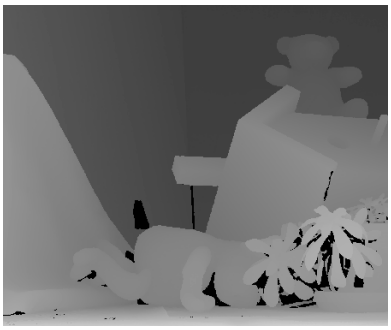


(f) Proposed luminance-chrominance-range NSS model with L*, a*, and b*

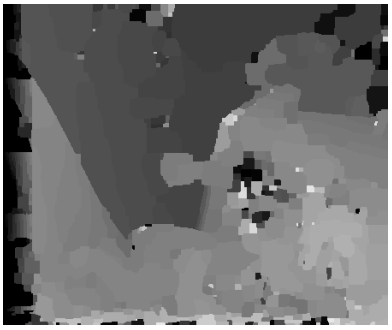
Figure 4.9: Simulation results on *Cones* from the Middlebury database, including the original stereo image pair, the ground-truth disparity map, and disparity maps using computed Bayesian stereopsis under different NSS models.

(a) Left image

(b) Right image

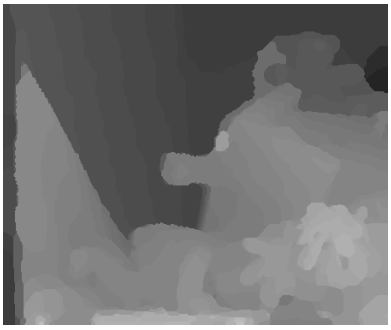


(c) Ground-truth disparity map



(d) Canonical formulation

(e) Previous NSS model



(f) Proposed luminance-chrominance-range NSS model with L*, a*, and b*

Figure 4.10: Simulation results on *Teddy* from the Middlebury database, including the original stereo image pair, the ground-truth disparity map, and disparity maps using computed Bayesian stereopsis under different NSS models.

Chapter 5

Bivariate and Spatial Oriented Correlation Statistical Models

5.1 Introduction

Most prevalent statistical models of natural, photographic images characterize only the univariate distributions of divisively normalized bandpass responses or wavelet-like decompositions of them. However, the higher-order dependencies between spatially neighboring responses are not yet well understood. Towards filling this gap, we propose new bivariate and spatial oriented correlation models that capture statistical regularities between perceptually decomposed natural image luminance samples. We validate the new model on a variety of natural images. As a demonstration of its usefulness, we deploy the new correlation model to solve the problem of image interpolation. The experimental results show that the proposed interpolation algorithm utilizing our new statistical model achieves comparable performance with bicubic image interpolation.

5.2 Univariate Natural Scene Statistical Models

A variety of natural scene statistical models have been developed in the vision science literature, both in the spatial [15] and wavelet domain [14]. Early on, Ruderman [15] showed that a simple non-linear operation of local mean subtraction followed by variance divisive normalization on natural image luminance results in a decorrelating and Gaussianizing effect. While the statistics, i.e., marginal distributions, of natural image pixels exhibit non-Gaussian behavior, after projection onto appropriate multi-scale spaces, e.g., using wavelet bases [17] or 2D Gabor filter banks [14], the resulting coefficients are found to obey regular statistical models, such as Gaussian scale mixtures [22]. These natural scene statistical models have been deployed in perceptual and computational image/video applications with great success, such as image denoising and restoration [23], and image/video quality assessment [25, 66, 100, 113].

However, efforts to date have focused on the use of first-order univariate statistical models, although there certainly exist significant dependencies between spatially neighboring bandpass image responses, which are not yet well understood or modeled. Specifically, little work has been conducted on modeling joint/bivariate relationships embedded in spatially oriented natural image luminances. In [21], Simoncelli found that the coefficients of orthonormal wavelet decompositions of natural images are fairly well-decorrelated; however, they are not independent. He also showed that the empirical joint histograms of adjacent coefficients produce contour plots having distinct 'bowtie' shapes. This was observed on coefficient pairs separated by different spatial offsets,

across adjacent scales, and at orthogonal orientations. However, quantitative models that characterize these bivariate distributions are not available.

Here we make progress towards filling this gap by introducing new bivariate and correlation models of spatially neighboring bandpass image responses. The models use a versatile multivariate generalized Gaussian distribution combined with a new exponentiated sine function model of correlation. We statistically validate the robustness of the new NSS models, and demonstrate their usefulness by application to image interpolation.

5.3 Bivariate and Correlation Natural Scene Statistical Models

Human vision systems (HVS) extract abundant information from natural environments by processing visual stimuli through different levels of decomposition and interpretation. Since we want to learn and explore the statistical relationships that are embedded in natural images, and how these statistics might be implicated in visual processing and used for practical image processing, we apply certain perceptually relevant pre-processing steps on natural image luminance, and develop our new bivariate and correlation models from the empirical response distributions.

The basic resources on which we perform bivariate and correlation statistical modeling are the pristine images from the popular and widely used LIVE IQA Database [114].

5.3.1 Perceptual Decomposition

We acquire luminance by transforming pristine color images into the perceptually relevant CIELAB color space, which is optimized to quantify perceptual color differences and better corresponds to human color perception than does the perceptually nonuniform RGB space [89]. Each luminance image (L*) is then transformed by the steerable pyramid decomposition, which is an over-complete wavelet transform that allows for increased orientation selectivity [115]. The use of the wavelet transform is motivated by the fact that its space-scale-orientation decomposition is similar to the bandpass filtering that occurs in area V1 of primary visual cortex [14, 116].

After applying the multi-scale, multi-orientation decomposition, we perform the perceptually significant process of divisive normalization on the luminance wavelet coefficients of all of the sub-bands [117]. The divisive normalization transform (DNT) used in our work is implemented as follows [118]:

$$u(x_i, y_i) = \frac{w(x_i, y_i)}{\sqrt{s + \mathbf{w_g}^\top \mathbf{w_g}}} = \frac{w(x_i, y_i)}{\sqrt{s + \sum_j g(x_j, y_j) w(x_j, y_j)^2}} \tag{5.1}$$

where (x_i, y_i) are spatial coordinates, w are the wavelet coefficients, u are the coefficients after DNT, s is a semi-saturation constant, the weighted sum occurs over neighborhood pixels indexed by j, and $\{g(x_j, y_j)\}$ is a finite-extent Gaussian weighting function.

5.3.2 Bivariate Statistical Model

We study the joint distribution of spatially adjacent luminance wavelet coefficients subjected to DNT, i.e., u in Eq. (5.1). Specifically, we use the steerable pyramid decomposition with five scales, indexed from 1 (finest) to 5 (coarsest), and twelve frequency-tuning orientations: 0, $\frac{1}{12}\pi$, ..., $\frac{11}{12}\pi$.

Here we mainly focus on the bivariate distributions and correlations of horizontally and vertically adjacent pixels. Specifically, for horizontally adjacent pixels, we sample pairs from locations (x, y) and $(x+1, y)$ in an image. Since we have observed that very similar statistics arise from horizontally and vertically adjacent pixels, we will only discuss the results for the horizontal case.

To model the bivariate joint histogram of spatially adjacent band-pass responses, we utilize a multivariate generalized Gaussian distribution (MGGD), which includes both the multivariate Gaussian and Laplace distributions as special cases. The probability density function of a multivariate generalized Gaussian distribution that we use is:

$$p(\mathbf{x}; \mathbf{M}, \alpha, \beta) = \frac{1}{|\mathbf{M}|^{\frac{1}{2}}} g_{\alpha,\beta}(\mathbf{x}^\top \mathbf{M}^{-1} \mathbf{x}) \tag{5.2}$$

where $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{M}$ is an $N \times N$ symmetric scatter matrix, α and β are scale and shape parameters, respectively, and $g_{\alpha,\beta}(\cdot)$ is the density generator:

$$g_{\alpha,\beta}(y) = \frac{\beta \Gamma(\frac{N}{2})}{(2^{\frac{1}{\beta}} \pi \alpha)^{\frac{N}{2}} \Gamma(\frac{N}{2\beta})} e^{-\frac{1}{2}(\frac{y}{\alpha})^\beta} \tag{5.3}$$

where $y \in \mathbb{R}^+$. Note that when $\beta = 0.5$, Eq. (5.2) yields the multivariate Laplacian distribution, and when $\beta = 1$, Eq. (5.2) corresponds to the multivariate Gaussian distribution. Moreover, when $\beta \to \infty$, the MGGD converges to a multivariate uniform distribution.

To fit an MGGD model to the bivariate histogram of spatially adjacent sub-band coefficients of a natural image and to find the corresponding model parameters, we adopt the maximum likelihood estimator (MLE) algorithm [50, 119]. Specifically, when the shape parameter, β, of the MGGD model is unknown, the MLEs of parameters $\mathbf{M}$, α, and β can be obtained by differentiating the log-likelihood of $p(\{\mathbf{x}_1, \cdots, \mathbf{x}_K\}|\mathbf{M}, \alpha, \beta)$, where $\{\mathbf{x}_1, \cdots, \mathbf{x}_K\}$ are K independent and identically distributed (i.i.d.) MGGD random vectors, with respect to $\mathbf{M}$, α, and β. This yields the MLEs of the parameters $\mathbf{M}$, α, and β, as below.

$$\mathbf{M} = \frac{1}{K} \sum_{k=1}^{K} \left[\frac{NK}{y_k + y_k^{1-\beta} \sum_{j \neq k}^{K} y_j^{\beta}} \mathbf{x}_k \mathbf{x}_k^{\top} \right] \tag{5.4}$$

$$\alpha = \left[\frac{\beta}{NK} \sum_{k=1}^{K} y_k^{\beta} \right]^{\frac{1}{\beta}} \tag{5.5}$$

$$f(\beta) = \frac{NK}{2 \sum_{k=1}^{K} y_k^{\beta}} \sum_{k=1}^{K} \left[y_k^{\beta} \ln(y_k) \right] - \frac{NK}{2\beta} \left[\Psi\left(\frac{N}{2\beta} \right) + \ln\left(\frac{2\beta}{NK} \sum_{k=1}^{K} y_k^{\beta} \right) \right] - K = 0 \tag{5.6}$$

where $y_k = \mathbf{x}_k^{\top} \mathbf{M}^{-1} \mathbf{x}_k$ and $\Psi(\cdot)$ is the digamma function, which is the logarithmic derivative of the gamma function, i.e., $\Psi(x) = \frac{d}{dx} \ln(\Gamma(x))$.

Note that the MLEs of $\mathbf{M}$ and β depend on each other, while α can be

Algorithm 1 Estimate the MGGD parameters using the MLEs

Initialize $\mathbf{M}$ and β
for $i = 1$ **to** *max_num_iter* **do**
  Estimate $\mathbf{M}$ using Eq. (5.4).
  Estimate β using Eq. (5.7) via the Newton-Raphson method:

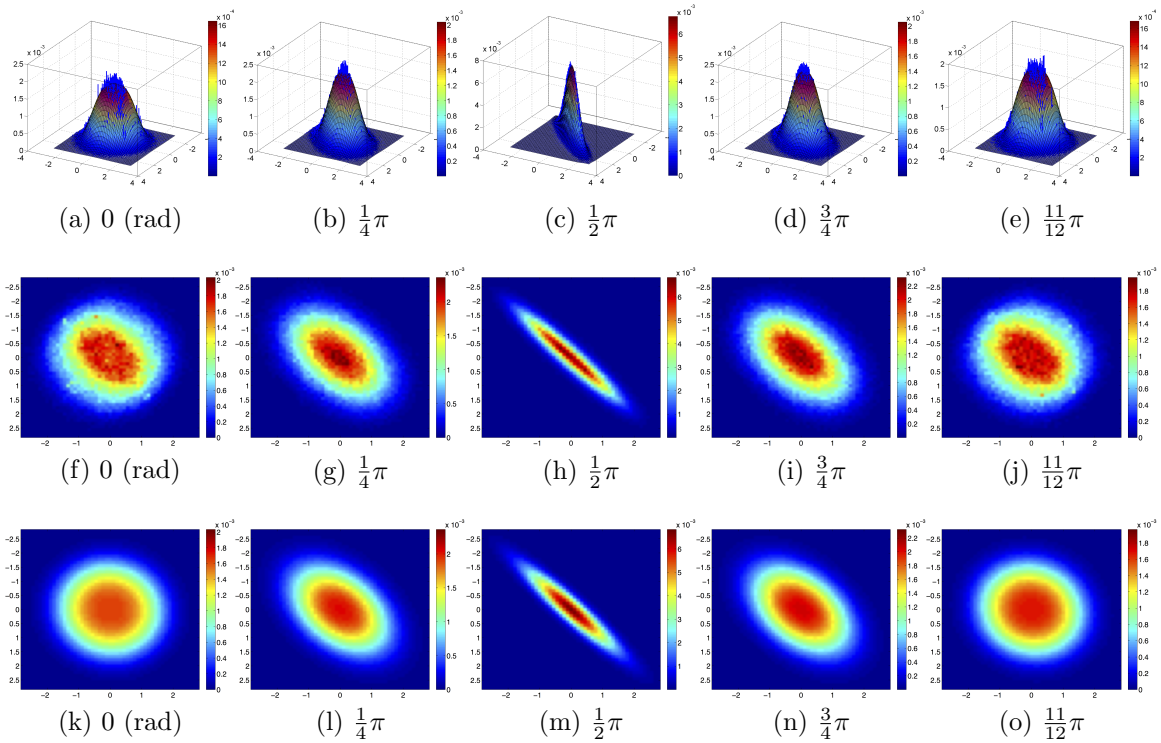
$$\beta_i = \beta_{i-1} - \frac{f(\beta_{i-1})}{f'(\beta_{i-1})} \tag{5.7}$$

  if $|\beta_i - \beta_{i-1}| \leq$ *fitting_error* **then**
    break
  end if
end for
Estimate α using Eq. (5.5).

estimated directly from β. The iterative algorithm as shown in Algorithm 1 yields MLEs of the MGGD model parameters.

We model the bivariate empirical histograms of horizontally adjacent sub-band coefficients in natural images as following a bivariate generalized Gaussian distribution (BGGD), *viz.*, using Eq. (5.2) with $N = 2$. The BGGD parameters are obtained using the maximum likelihood estimator (MLE) algorithm described above.

Figure 5.1 shows the empirical joint distributions of horizontally adjacent sub-band responses and their corresponding BGGD fits on pristine image 'building2' from the LIVE IQA Database [114]. As may be seen in the three-dimensional illustrations shown in the top row, where the blue bars represent the actual histograms and the colored meshes represent the BGGD fits, the joint distributions of L* sub-band responses are well modeled as bivariate gen-



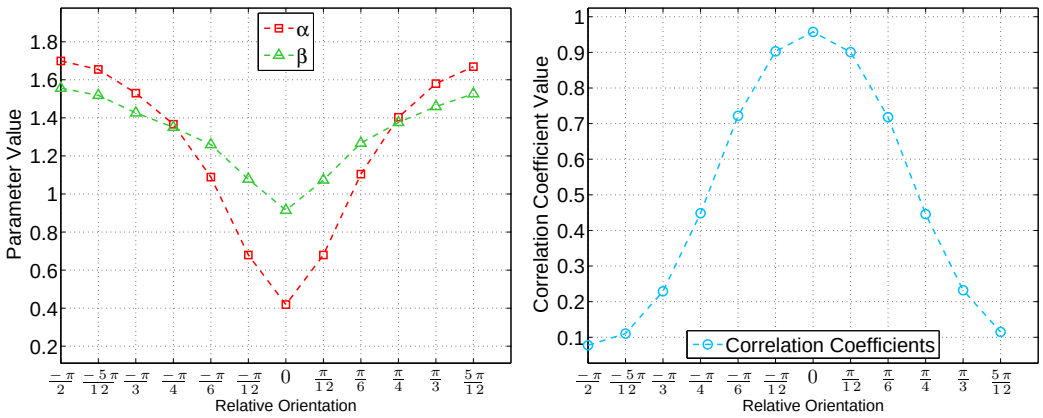
(a) 0 (rad) (b) $\frac{1}{4}\pi$ (c) $\frac{1}{2}\pi$ (d) $\frac{3}{4}\pi$ (e) $\frac{11}{12}\pi$

(f) 0 (rad) (g) $\frac{1}{4}\pi$ (h) $\frac{1}{2}\pi$ (i) $\frac{3}{4}\pi$ (j) $\frac{11}{12}\pi$

(k) 0 (rad) (l) $\frac{1}{4}\pi$ (m) $\frac{1}{2}\pi$ (n) $\frac{3}{4}\pi$ (o) $\frac{11}{12}\pi$

Figure 5.1: Joint histograms of horizontally adjacent bandpass coefficients from a pristine image and the corresponding BGGD fits at the finest scale with different orientations. From left column to right column: 0 (rad), $\frac{1}{4}\pi$, $\frac{1}{2}\pi$, $\frac{3}{4}\pi$, and $\frac{11}{12}\pi$. Top row: 3D illustration of bivariate histogram and BGGD fit, middle row: 2D iso-probability contour plot of histogram, and bottom row: 2D iso-probability contour plot of BGGD fit.

eralized Gaussian. The 2D illustrations, which depict iso-probability contour maps of the joint distributions and the fits in the middle and bottom rows, respectively, also demonstrate the close fits of the BGGD model. The most important observation here is that both the shape and height of the bivariate distributions and fits vary with the tuning orientations of the sub-band responses. In particular, when the spatial relationship between bandpass sam-



(a) BGGD parameters (b) Correlation Coefficients

Figure 5.2: Plots of the two BGGD model parameters and the correlation coefficients as a function of relative orientation.

ples, e.g., horizontal, matches the sub-band tuning orientation, e.g., $\frac{1}{2}\pi$, then the joint distribution becomes peaky and extremely elliptical, meaning the horizontally adjacent bandpass responses are highly correlated at sub-band orientation $\frac{1}{2}\pi$. Conversely, when the spatial relationship and the sub-band tuning orientation are orthogonal, e.g., horizontal and 0 (rad), then the joint distribution becomes nearly a circular Gaussian, implying almost uncorrelated sub-band responses.

To further examine this spatial orientation dependency, in Figure 5.2 (a) we plotted the BGGD model parameters, i.e., α and β, as a function of relative orientation at the same scale as in Figure 5.1. Here we define relative orientation as the difference between the sub-band tuning orientation and the spatial orientation of adjacent responses. For example, if the sub-band tuning orientation is 0 (rad), and the pixels are horizontally adjacent, i.e., the spatial

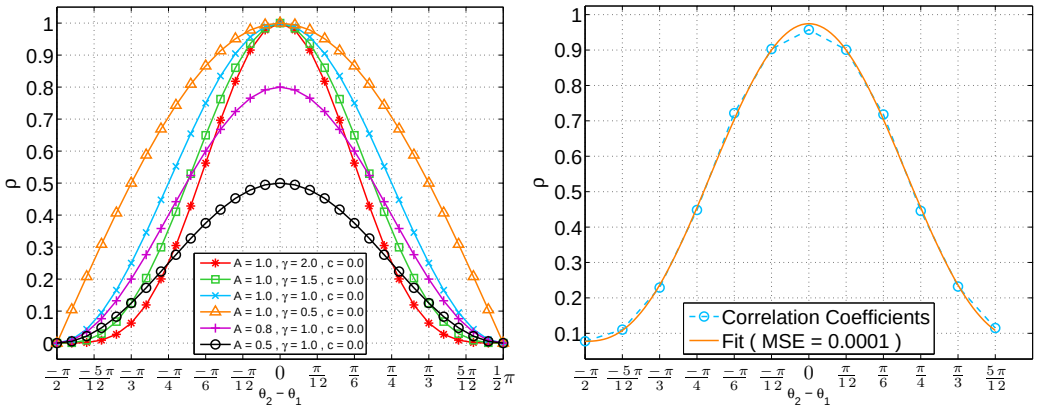
orientation is $\frac{1}{2}\pi$, meaning that they are orthogonal, then the corresponding relative orientation is equal to $0-\frac{1}{2}\pi = -\frac{1}{2}\pi$. Figure 5.2 (a) clearly shows that there is strong orientation dependency of both parameters. We have also studied the behavior of the correlation coefficients of spatially adjacent responses as a function of relative orientation. These are contained in the scatter matrix $\mathbf{M}$ of the BGGD model (Eq. (5.2) with $N = 2$). Figure 5.2 (b) shows the correlation coefficients between horizontally adjacent bandpass responses as a function of relative orientation. The horizontally adjacent bandpass responses are most correlated when the sub-band tuning orientation aligns at $\frac{1}{2}\pi$, and become nearly uncorrelated at orientations 0 (rad) and π, substantiating the spatial relative orientation dependency observed in Figure 5.1.

5.3.3 Spatial Oriented Correlation Model

Motivated by this observed regular, periodic behavior, we have deployed an exponentiated sine function to model the correlation coefficients as a function of relative orientation:

$$\rho = f(\theta_1, \theta_2) = A \left[\frac{1 + \sin\left(\frac{2\pi(\theta_2 - \theta_1)}{T} + \varphi\right)}{2} \right]^{\gamma} + c \tag{5.8}$$

where ρ is the correlation coefficients between spatially adjacent bandpass responses, θ_1 and θ_2 represent spatial and sub-band tuning orientations, respectively, A is the amplitude, T is the period, φ is the phase, γ is the exponent, and c is the offset. Since the correlation coefficient is period-π in relative orientation and reaches maximum when $\theta_2 - \theta_1 = k\pi, k \in \mathbb{Z}$, we obtain a



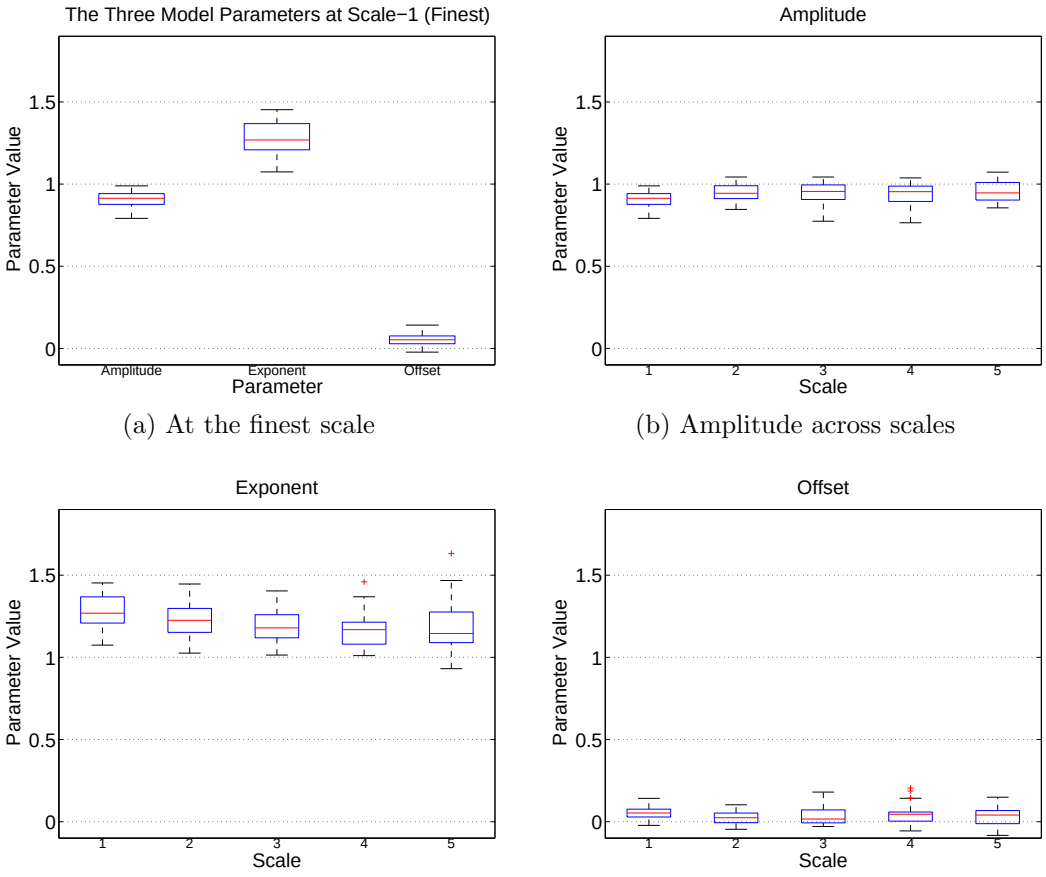
(a) Exponentiated sine function (b) Fit to correlation coefficients

Figure 5.3: The exponentiated sine function and its fit to correlation coefficients as a function of relative orientation.

three-parameter exponentiated sine model by fixing $T = \pi$ and $\varphi = \frac{\pi}{2}$:

$$\begin{aligned} \rho = f(\theta_1, \theta_2) &= A\left[\frac{1 + \cos\left(2(\theta_2 - \theta_1)\right)}{2}\right]^{\gamma} + c \\ &= A\left[\cos\left(\theta_2 - \theta_1\right)\right]^{2\gamma} + c \end{aligned} \tag{5.9}$$

Figure 5.3 (a) shows exemplar exponentiated sine curves for different sets of parameters. The exponentiated sine model is able to capture a wide range of periodic curves having bell-shaped lobes of varying relative slopes. Figure 5.3 (b) plots an empirical correlation coefficient curve as a function of relative orientation and its overlaid exponentiated sine fit for horizontally adjacent bandpass responses, i.e., $\theta_1 = \frac{1}{2}\pi$. From both the curve overlap and associated mean squared error (MSE), it is apparent that the exponentiated sine model fits the spatial oriented correlations between adjacent bandpass luminance responses extremely well.



(a) At the finest scale

(b) Amplitude across scales

(c) Exponent across scales

(d) Offset across scales

Figure 5.4: The box plots of the exponentiated sine model parameters.

To gain more insight of this exponentiated sine model, we computed the correlation coefficients between horizontally adjacent bandpass responses as a function of sub-band tuning orientation for all 29 pristine images in the LIVE IQA Database, and found the corresponding exponentiated sine model parameters, i.e., amplitude A, exponent γ, and offset c. In Figure 5.4 (a), we present the box plots of the three model parameters at the finest scale

across all pristine images with whiskers expressing the 1.5 interquartile range (IQR). Figure 5.4 (b) to (d) show the box plots of amplitude, exponent, and offset obtained from all pristine images across different scales, respectively. Clearly, both the amplitude and offset parameters hold fairly consistent values across image content and scales, i.e., $A \approx 0.9$ and $c \approx 0.05$, while the exponent parameter varies roughly within the range of $[1, 1.4]$. Based on these statistics, a succinct one-parameter model may be arrived at:

$$\rho = f(\theta_1, \theta_2; \gamma) = 0.9 \left[\cos\left(\theta_2 - \theta_1\right)\right]^{2\gamma} + 0.05 \tag{5.10}$$

Indeed, little is lost and simplicity gained by taking $A = 1$ and $c = 0$, wherein Eq. (5.10) becomes

$$\rho = f(\theta_1, \theta_2; \gamma) = \left[\cos\left(\theta_2 - \theta_1\right)\right]^{2\gamma} \tag{5.11}$$

In each of (5.8)–(5.11), the model parameters are estimated with non-linear least squares using the Levenberg-Marquardt algorithm [120].

5.4 Validation of the Exponentiated Sine Model

To validate the robustness of the new spatial-oriented correlation model (Eq. (5.11)), we performed a statistical hypothesis test on the 29 pristine images in the LIVE IQA Database. In particular, we used a chi-squared test for goodness of fit. First, we computed the exponentiated cosine model parameter γ at each scale by fitting the mean correlation coefficients between horizontally adjacent bandpass responses as a function of sub-band tuning

orientation for all LIVE pristine images. Then, we obtained the corresponding exponentiated cosine function, i.e., $\boldsymbol{\rho}_\gamma \in \mathbb{R}^D$ where D is the number of sub-band tuning orientations, using Eq (5.11). Finally, we computed the chi-squared statistic χ^2 to determine if the null hypothesis H_0, i.e., the correlation coefficients as a function of sub-band tuning orientation are drawn from a population with mean equal to $\boldsymbol{\rho}_\gamma$, is supported. Specifically, if H_0 is rejected, it means that the exponentiated cosine function is not a statistically robust model for natural spatial-oriented correlations; otherwise, we can conclude that the spatial-oriented correlations of all LIVE pristine images can be statistically represented by the exponentiated cosine model $\boldsymbol{\rho}_\gamma$. The chi-squared statistic χ^2 is computed as:

$$\chi^2 = \sum_{i=1}^{N} \sum_{j=1}^{D} \frac{(\rho_{i_j} - \rho_{\gamma_j})^2}{\rho_{\gamma_j}} \tag{5.12}$$

where $\{\rho_{\gamma_j}\} = \boldsymbol{\rho}_\gamma \in \mathbb{R}^D$ is the model, $\{\rho_{i_j}\} = \boldsymbol{\rho}_i \in \mathbb{R}^D$ are the correlation coefficients as a function of sub-band tuning orientation for the i-th pristine image, and N is the number of pristine images. We repeated this procedure to perform chi-squared statistical tests on all five scales, from 1 (finest) to 5 (coarsest). The test results are summarized in Table 5.1. We can see that the p-values for all five scales are larger than a significance level $\alpha = 0.05$, indicating that the new spatial-oriented exponentiated cosine correlation model holds well for the tested natural images. In addition, the model parameter γ estimated for each scale varies slightly around 1.2, which supports the box plot of γ in Figure 5.4 (c).

Table 5.1: Chi-Squared Statistical Test Results

Scale	Model Parameter γ	χ^2	p-value	$> \alpha = 0.05$?
1	1.2515	14.6661	0.1983	Yes
2	1.2691	8.3057	0.6857	Yes
3	1.1846	16.4710	0.1245	Yes
4	1.1491	16.1336	0.1362	Yes
5	1.1759	9.9386	0.5359	Yes

5.5 Application to Image Interpolation

Image interpolation/up-sampling is a common operation on digital photographs. As an example application, we developed a simple image interpolation algorithm using the new spatial-oriented correlation model. Assume we want to up-sample an image by a factor of 2 along both dimensions. We first compute the correlation coefficients between spatially adjacent bandpass responses and acquire the exponentiated cosine model of the given image. We record the corresponding correlation coefficients as a function of sub-band tuning orientation as the target model. Then, we insert zeros between neighboring pixels to achieve factor-2 up-sampling. Next, at each pixel location, we compute the correlation coefficients as a function of sub-band tuning orientation between spatially adjacent pixels within a local neighborhood, e.g., an n-by-n window. Finally, we interpolate the best pixel value that generates a spatial-oriented correlation model closest to the recorded target model in terms of the L_2 norm.

Figure 5.5 shows an example of an interpolated image by a factor of 4 along both dimensions utilizing the new spatial-oriented correlation model,

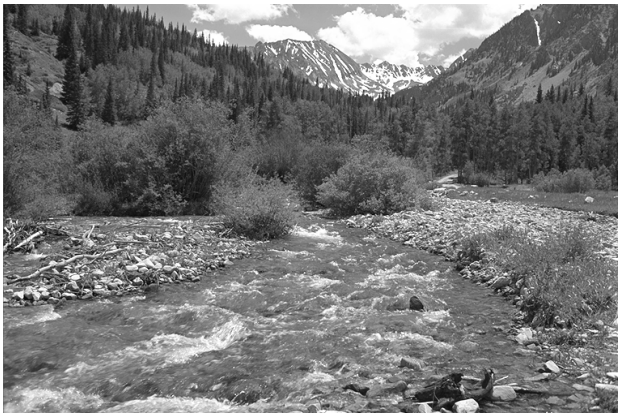
Table 5.2: Image Interpolation Results

Method	PSNR	
	Median	Standard Deviation
Bicubic	22.19	3.28
Proposed	22.62	3.20

along with the results of the commonly used method, bicubic interpolation. We can see that the proposed interpolation method using the new correlation model generates a better result than the bicubic interpolation both visually and in terms of the PSNR error metric. Table 5.2 summarizes the image interpolation results for all 29 pristine images in the LIVE IQA Database, where the new spatial-oriented correlation model achieves higher PSNR with better consistency across image content.

5.6 Summary

We have proposed new natural scene statistical models that express the bivariate joint distributions and correlations between spatially neighboring bandpass responses of natural images. The new model was statistically validated as able to model the relative oriented correlations of natural luminance images. A simple application to image interpolation demonstrated the effectiveness of the new spatial oriented correlation NSS model.



(a) Pristine

(b) Bicubic (PSNR = 18.71)

(c) Proposed (PSNR = 19.25)

Figure 5.5: Example image interpolation result.

Chapter 6

Depth Estimation from Monocular Natural Images

6.1 Introduction

Estimating an accurate and naturalistic dense depth map from a single monocular photographic image is a difficult problem. Nevertheless, 2D images of the real-world environment contain significant statistical information regarding the 3D structure of the world. Towards exploiting this information to solve the problem, we propose a Bayesian model, termed Natural3D, that recovers detailed 3D scene structures by extracting reliable and robust statistical features. These features are derived from standard marginal univariate natural scene statistics (NSS) models as well as new bivariate/correlation NSS models that describe the relationships between 2D photographic images and their associated depth maps. This is accomplished by building a dictionary of canonical depth patterns from which NSS features are extracted as prior information. The dictionary is used to create a multivariate Gaussian mixture (MGM) likelihood model that associates local image features with depth patterns. The resulting Bayesian model is then used to form spatial depth predictions. As compared with state-of-the-art depth estimation methods, superior performance is obtained in terms of correlations with ground-truth

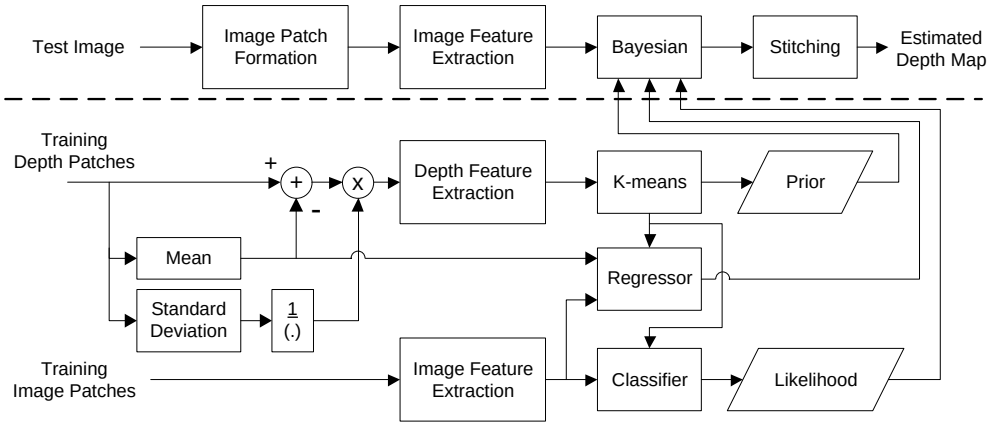


Figure 6.1: Block diagram of Natural3D.

depths and estimated depth errors.

6.2 Proposed Bayesian Depth Estimation Model

We begin by summarizing Natural3D and the contributions we make. As shown in Fig. 6.1, Natural3D is partitioned into two phases, training and testing. In the training phase, shown below the dotted line in Fig. 6.1, patches of size $M \times M$ are collected from a set of real-world photographic images and their corresponding ground-truth depth maps. From each patch pair, a vector of natural scene statistics (NSS) features is extracted and stored. Then, to capture valuable statistical relationships that are embedded in the luminances and depths of the natural images, we learn priors and likelihoods from these perceptually relevant image and depth features. In the testing phase, an input image is also partitioned into overlapping $M \times M$ patches, and the same vector

of features is extracted from each image patch. A corresponding depth patch is then estimated for each image patch using a Bayesian model driven by the learned priors and likelihoods. Finally, all the depth patches are stitched together to form an estimated depth map. The details of each element of Natural3D are explained in the following subsections.

Our contributions are three-fold. First, Natural3D is the first NSS model-based depth-from-2D estimation algorithm. It employs both marginal univariate and new bivariate/correlation NSS models to extract 'depth-aware' features. Bayesian inference is used to achieve one-shot depth recovery without using high-level semantics, imposing smoothness constraints or iterative optimization methods. Second, we have created a new, high-quality 3D database, the LIVE Color+3D Database Release-2 [8], which contains high-resolution stereoscopic color image pairs with accurately co-registered dense depth maps. It provides a rich source of information regarding natural depth statistics, as well as an excellent resource for developing and evaluating a variety of stereoscopic/3D image processing and vision algorithms, such as depth estimation and quality assessment. We are making this database publicly available free of charge. Finally, Natural3D delivers superior performance relative to top-performing state-of-the-art depth estimation algorithms. We have made code for Natural3D available for independent evaluation and further academic research at [121].

6.2.1 Perceptual Decomposition

Human vision systems (HVS) extract abundant information from natural environments by processing visual stimuli through massively parallel and pipelined levels of decomposition and interpretation. By analyzing the natural statistics of the 2D and 3D visual world, and by learning how the HVS processes natural image and depth information, a variety of statistical models have been proposed that capture the behavior of perceptually motivated bandpass responses of luminance/chrominance and depth/disparity on natural scenes [30, 32, 33]. Since the philosophy underlying our approach is to learn and employ good models of the statistical laws that describe the relationships between depth perception and the structure of natural images, we apply certain perceptually relevant preprocessing steps to the recorded image data, including biologically-motivated linear bandpass decompositions and nonlinear divisive normalization processes. Then, 'depth-aware' NSS features are extracted from the univariate and bivariate empirical distributions of these responses.

Our work is therefore admittedly perceptually motivated, and hence is defensible as a type of image engineering to create 3D presentations suitable for human viewing, as for example in the creation of 3D cinematic or television content from archived 2D movies. However, as we show in the sequel, the method delivers highly competitive objective results, and we envision that, given its conceptual and computational simplicity, it could find other (e.g. 'robotic') applications. In its current form, we utilize only luminance informa-

tion in our model, although there are definite statistical relationships between image color and depth [33] as well as in the perception of depth on color images [86].

We acquire luminance from color images by transforming them into the perceptually uniform CIELAB color space [89]. Each luminance image (L*) is then decomposed by a steerable pyramid decomposition, which is an over-complete wavelet transform that allows for increased orientation selectivity [115]. The use of the wavelet transform is motivated by the fact that its space-scale-orientation decomposition is similar to the bandpass filtering that occurs in area V1 of primary visual cortex [14, 116]. In our experimental implementation, we deploy a steerable pyramid decomposition with five scales, indexed from 1 (finest) to 5 (coarsest), and twelve frequency-tuning orientations: 0, $\frac{1}{12}\pi$, ..., $\frac{11}{12}\pi$ (rad).

After applying the multi-scale, multi-orientation decomposition, we perform the perceptually significant process of divisive normalization on the luminance wavelet coefficients of all of the sub-bands [117]. Divisive normalization, also termed adaptive gain control, has been developed in the psychovisual literature to account for the nonlinear behavior of certain cortical neurons [122]. The divisive normalization transform (DNT) that we use is defined as [118]:

$$u(x_i, y_i) = \frac{w(x_i, y_i)}{\sqrt{s + \mathbf{w_g}^\top \mathbf{w_g}}} = \frac{w(x_i, y_i)}{\sqrt{s + \sum_j g(x_j, y_j) w(x_j, y_j)^2}} \tag{6.1}$$

where (x_i, y_i) are spatial coordinates, w are the wavelet coefficients, u are the coefficients following the DNT, s is a semi-saturation constant, $\{g(x_j, y_j)\}$ is a finite-extent Gaussian weighting function, and the weighted sum occurs over neighborhood pixels indexed by j.

In the following subsections, we explain the details of how the image and depth features are extracted from the divisively normalized sub-band responses, and how features are used to learn the prior and likelihood models for depth estimation.

6.2.2 Image Feature Extraction

It is well-established that there exist statistical relationships between image luminances and depths information in natural scenes [30], and a variety of univariate statistical models have been proposed to fit the bandpass responses of luminance/chrominance and disparity [32, 33]. Very recently, new closed-form bivariate and correlation statistical models have been developed that effectively capture spatial dependencies between neighboring sub-band responses in natural images [51]. Natural3D exploits these NSS features to learn the relationships that exist between projected image luminances and depths information that is embedded in them.

6.2.2.1 Univariate NSS Feature

Considerable work has been conducted on modeling the statistics of natural images that have been passed through multi-scale, multi-orientation

bandpass transforms, e.g., Gabor filters, wavelets, etc [16, 18]. A common and well-accepted model of the empirical histograms of divisively normalized luminance sub-band responses, i.e., u in Eq. (6.1), is the univariate generalized Gaussian distribution (GGD). The probability density function of a univariate GGD with zero mean is:

$$p(x;\alpha_u,\beta_u) = \frac{\beta_u}{2\alpha_u\Gamma(\frac{1}{\beta_u})}e^{-(\frac{|x|}{\alpha_u})^{\beta_u}} \tag{6.2}$$

where $\Gamma(\cdot)$ is the ordinary gamma function and α_u and β_u are scale and shape parameters, respectively. For pristine, undistorted natural images, it is commonly assumed that $\beta_u = 2$ (i.e. Gaussian), while distortions tend to create structural degradations that modify β_u (typically $\beta_u < 2$). We estimate the GGD parameters on small $M \times M$ patches, so β_u locally varies. The two GGD parameters of each sub-band (scale and shape) are estimated from each bandpass patch histogram (using the method in [123]) and are included in the feature set of the patch.

6.2.2.2 Bivariate NSS Feature

We also capture dependencies that exist between spatially neighboring bandpass image responses by modeling the bivariate distributions of horizontally adjacent sub-band responses sampled from all locations (x, y) and $(x + 1, y)$ in each image patch. Since we have observed similar statistics from both horizontally and vertically neighboring responses [50], and used sub-band orientations covering 0 to π (rad), we exploit only horizontally adjacency to achieve the same efficacy with reduced computational complexity. This also

applies to the correlation NSS feature, which will be detailed in Sec. 6.2.2.3. To model these empirical joint histograms, we utilize a multivariate generalized Gaussian distribution (MGGD), which includes both the multivariate Gaussian and Laplace distributions as special cases. The probability density function of a multivariate generalized Gaussian distribution is defined as:

$$p(\mathbf{x}; \mathbf{M}, \alpha_b, \beta_b) = \frac{1}{|\mathbf{M}|^{\frac{1}{2}}} g_{\alpha_b, \beta_b}(\mathbf{x}^\top \mathbf{M}^{-1} \mathbf{x}) \tag{6.3}$$

where $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{M}$ is an $N \times N$ symmetric scatter matrix, α_b and β_b are scale and shape parameters, respectively, and $g_{\alpha_b, \beta_b}(\cdot)$ is the density generator:

$$g_{\alpha_b, \beta_b}(y) = \frac{\beta_b \Gamma(\frac{N}{2})}{(2^{\frac{1}{\beta_b}} \pi \alpha_b)^{\frac{N}{2}} \Gamma(\frac{N}{2\beta_b})} e^{-\frac{1}{2}(\frac{y}{\alpha_b})^{\beta_b}} \tag{6.4}$$

where $y \in \mathbb{R}^+$. Note that when $\beta_b = 0.5$, Eq. (6.3) becomes the multivariate Laplacian distribution, and when $\beta_b = 1$, Eq. (6.3) corresponds to the multivariate Gaussian distribution. When $\beta_b \to \infty$, the MGGD converges to a multivariate uniform distribution, and when $\beta_b < 0.5$, it becomes a 2D heavy-tailed 'sparsity' density. In our implementation, we model the bivariate empirical histograms of horizontally adjacent sub-band coefficients of each image patch using a bivariate generalized Gaussian distribution (BGGD) with $N = 2$ in Eq. (6.3). The parameters of the BGGD can be obtained on the bandpass coefficients of image patches using the maximum likelihood estimator (MLE) algorithm described in [50]. The scale and shape parameters, α_b and β_b, are included in each image patch's feature set.

6.2.2.3 Correlation NSS Feature

We also model the correlation structure that exists between spatially neighboring bandpass luminance responses. In particular, we have found that the correlation coefficients between spatially adjacent bandpass responses possess strong orientation dependencies [51]. For example, horizontally adjacent bandpass responses are most correlated when the sub-band tuning orientation aligns at $\frac{1}{2}\pi$ (rad), and become nearly uncorrelated at orientation 0 and π (rad). The correlation is periodic in relative orientation between spatial and sub-band tuning orientation. This relative orientation regularity of correlation implies that there exist powerful constraints on spatially neighboring bandpass image responses.

Indeed, the periodic relative orientation dependency of the correlation coefficients between spatially adjacent bandpass responses can be well modeled in a closed form by an exponentiated sine function:

$$\rho = f(\theta_1, \theta_2) = A \left[\frac{1 + \sin\left(\frac{2\pi(\theta_2 - \theta_1)}{T} + \varphi\right)}{2} \right]^{\gamma} + c \tag{6.5}$$

where ρ is the correlation coefficient between spatially adjacent bandpass responses, θ_1 and θ_2 are spatial and sub-band tuning orientations, respectively, A is amplitude, T is the period, φ is the phase, γ is an exponent, and c is the offset. When measured on naturalistic photographic images, the correlation coefficient is found to be π-periodic, reaching maximum when $\theta_2 - \theta_1 = k\pi, k \in \mathbb{Z}$,

yielding a three-parameter exponentiated sine model:

$$\rho = f(\theta_1, \theta_2) = A\left[\frac{1+\cos\left(2(\theta_2-\theta_1)\right)}{2}\right]^{\gamma} + c$$
$$= A\left[\cos\left(\theta_2-\theta_1\right)\right]^{2\gamma} + c \tag{6.6}$$

In our implementation, we compute the correlation coefficients between all horizontally adjacent sub-band responses within each image patch over all scales, fit each with the exponentiated sine model, and include all three fitting parameters, A, γ, and c, into the feature set. The fitting parameters are estimated via non-linear least squares using the Levenberg-Marquardt algorithm [120].

At this point, all of the NSS-based features that drive Natural3D have been described. As a result, the 'depth-aware' image feature vector $\mathbf{f}_I$ that is used to characterize each image patch is formed as:

$$\mathbf{f}_I = \left[\{\alpha_{u,s,r}, \beta_{u,s,r}\}, \{\alpha_{b,s,r}, \beta_{b,s,r}\}, \{A_s, \gamma_s, c_s\}\right]^{\top} \tag{6.7}$$

where $s \in \{1, 2, \ldots, S\}$, S is the number of scales, and $r \in \{1, 2, \ldots, R\}$, R is the number of sub-band orientations.

6.2.3 Depth Feature Extraction

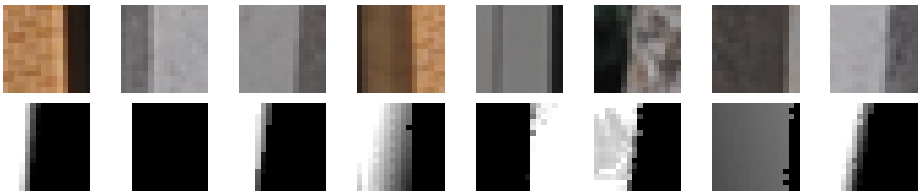
To a much greater degree than image luminances, range/depth maps captured from natural environments are smooth surfaces with relatively few textures. Based on this observation, and since we are interested in capturing depth differentials, we use the histograms of the gradient magnitudes [124] of debiased and normalized patch depths extracted from ground-truth depth

maps as features in the depth prior model. The center-left part of Fig. 6.1 summarizes this process. The mean value is subtracted from each depth patch, then the result is divisively normalized by the depth patch standard deviation. The gradient magnitudes of the resulting normalized depth patches are computed and the corresponding histograms found. Specifically, the histograms of depth gradient magnitudes are computed along eight canonical orientations: $0, \frac{1}{8}\pi, \ldots, \frac{7}{8}\pi$ (rad), resulting in an eight-bin histogram for each depth patch.

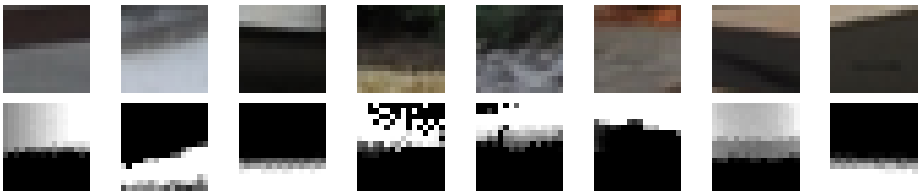
In addition to the gradient histograms, the histograms of bandpass response magnitudes are also computed on the perceptually decomposed depth patches. Specifically, we compute the divisively normalized wavelet responses using the same steerable pyramid and the same normalization (Eq. (6.1)), and obtain histograms by binning the responses within each depth patch along the same eight canonical sub-band tuning orientations as used to define the image gradient magnitude histograms. In sum, a 16-dimensional depth feature vector $\mathbf{f}_D$ characterizing each depth patch is arrived at. These are used to create the prior and likelihood models, as explained next.

6.2.4 Prior

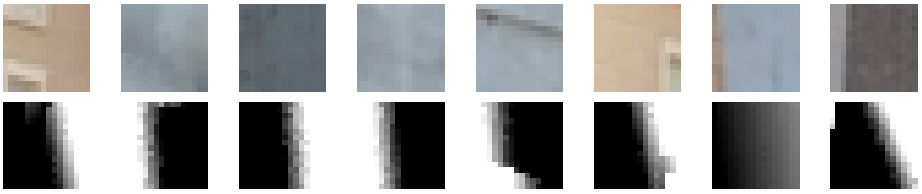
It has been observed that discontinuities in depth maps are usually co-located with luminance edges occurring in the corresponding optical images [125]. Depth patches having similar depth patterns may be expected to exhibit similar luminance distributions [32]. Moreover, depth maps tend to possess simpler, more regular patterns than natural luminance images. Based on these



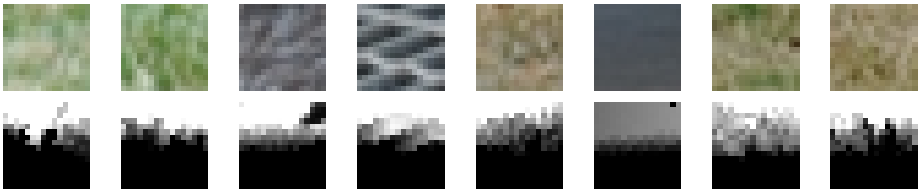
(a) Pattern-1



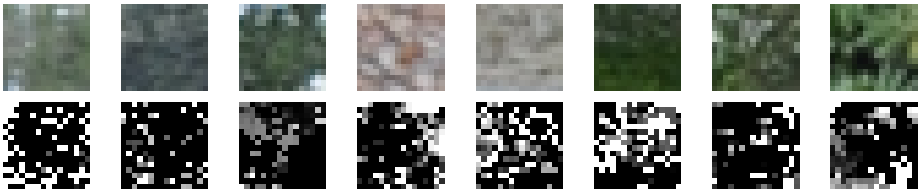
(b) Pattern-2



(c) Pattern-3



(d) Pattern-4



(e) Pattern-5

Figure 6.2: Examples of different canonical depth patterns.

observations, we build a dictionary of canonical depth patterns by clustering the processed, characteristic depth features that are extracted from depth patches as explained in the preceding section. As a simple method of data reduction, we employ the centroid-based k-means algorithm.

Figure 6.2 shows examples of several canonical depth patterns (near cluster centroids) extracted by the k-means algorithm assuming five clusters, each with eight examples. For each canonical depth pattern, the top row shows the clustered depth patches (normalized residues) using the extracted features, while the bottom row shows the co-registered image patches. The depicted canonical depth patterns contain a variety of geometric structures, including depth discontinuities along the horizontal direction (pattern-1), and along the vertical direction (pattern-2), smoother variation of depth along the horizontal direction (pattern-3), and along the vertical direction (pattern-4), and a busier, more complex pattern of depth changes (pattern-5). Complex depth patterns like pattern-5 are relatively uncommon, and appear in scenes containing rough objects, such as trees and grass. As the number of clusters is increased, these five canonical depth patterns still exist in similar form, although other clusters of depth patches emerge having similar structures that differ in some ways, such as orientation. In sum, the depth prior of Natural3D consists of the normalized residual depth patch $\mathbf{d}_n$, i.e., the cluster centroid associated with each canonical depth pattern, and the ratio $p(n)$ of each canonical depth pattern among all processed depth patches, where $n \in \{1, 2, \ldots, N\}$ and N is the number of canonical depth patterns, i.e., the number of clusters

used by the k-means algorithm.

The above procedure may be viewed as a way of finding a 'sparse' set of representative depth patches. This suggests that a more sophisticated 'sparse basis' might be found from which depth estimates could be computed. As a proof of concept, here we use the k-means algorithm owing to its simplicity and efficacy.

6.2.5 Likelihood

As may be observed from the canonical depth patterns shown in Fig. 6.2, depth discontinuities in range maps consistently align with luminance edges in co-registered natural images of the same scene [32, 125]. However, textured areas in photographic images that present significant variations in luminance/chrominance may not necessarily correspond to depth changes. In other words, there exist high correlations between image edges and depth discontinuities, although the relationship is asymmetric. If the bandpass response to an image patch contains significant energy, then there is a relatively high likelihood of co-located variations, i.e. large depth gradients, in the corresponding range map. Conversely, if the range map contains large variations, then the co-located image bandpass response is even more likely to be large. To better utilize these relationships between image and depth variations in naturalistic settings, we derive a likelihood model which associates image patches with appropriate canonical depth patterns.

Assume that N canonical depth patterns have been obtained defining

the prior using k-means clustering. Assign each image patch a label indicating its associated canonical depth pattern (cluster centroid) for its corresponding depth patch. Then, using these labeling results, the depth-aware feature vectors, i.e., $\mathbf{f}_I$ in Eq. (6.7), that are extracted from each image patch are used to train a classifier using a multivariate Gaussian mixture (MGM) model. The reason that the MGM model is well suited to this classification task is that, as may be observed in Fig. 6.2, image patches presenting different appearances and/or textured surfaces may yet be associated with the same canonical depth pattern. Therefore, we exploit the multi-modal Gaussian mixture model trained on each canonical depth pattern to be able to handle the heterogeneity of its image patches. An MGM model is defined as:

$$p(\mathbf{x};\boldsymbol{\theta}) = \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m) \tag{6.8}$$

where $\boldsymbol{\theta}$ is the model parameter vector, $\mathbf{x}$ is a multi-dimensional data vector, e.g., some measurement or a feature, $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)$ is the m-th Gaussian component, and w_m is the m-th mixture weight with the constraint that $\sum_{m=1}^{M} w_m = 1$. Note that the complete MGM model is parametrized by $\boldsymbol{\theta} = \{w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}, m \in \{1,\ldots,M\}$, which includes the mean vectors, covariance matrices, and mixture weights from all Gaussian components. Finally, the m-th Gaussian component density function is given by:

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m) = \frac{1}{(2\pi)^{K/2}|\boldsymbol{\Sigma}_m|^{1/2}} e^{\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)^\top \boldsymbol{\Sigma}_m^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)\right]} \tag{6.9}$$

where K is the dimensionality of $\mathbf{x}$. Here the depth-aware image feature vector is modeled: $\mathbf{x} = \mathbf{f}_I \in \mathbb{R}^K$. Therefore, the likelihood of encountering an image

patch with a specific extracted feature $\mathbf{f}_I$ given a particular canonical depth pattern indexed by n can be expressed as:

$$p(\mathbf{f}_I; \boldsymbol{\theta}_n) = \sum_{m=1}^{M} w_{n,m} \mathcal{N}(\mathbf{f}_I; \boldsymbol{\mu}_{n,m}, \boldsymbol{\Sigma}_{n,m}) \tag{6.10}$$

where $\boldsymbol{\theta}_n = \{w_{n,m}, \boldsymbol{\mu}_{n,m}, \boldsymbol{\Sigma}_{n,m}\}, m \in \{1, \ldots, M\}$.

6.2.6 Regression on Mean Depth

As discussed in Sec. 6.2.3 and 6.2.4, a preprocessing step is performed prior to the extraction of features from depth patches to learn the prior, whereby each depth patch is normalized by removing the mean and standard deviation to homogenize the depth patterns and to better reveal their essentially distinguishing characteristics. In order to be able to add the mean value of each depth patch back when estimating the true range values of test image patches, it is necessary to learn a mapping from the image feature space using a regression model. In other words, given an input image patch, the trained regressor can be used to estimate the mean range of the corresponding depth patch using the extracted depth-aware image feature vector $\mathbf{f}_I$. Since we observed negligible influences, both numerically and visually, of patch standard deviations on the estimated depth maps, Natural3D is able to attain the same degree of performance without recovering depth patch standard deviations.

In addition to $\mathbf{f}_I$, we exploit two other useful monocular depth cues to assist with recovery of true range values. As shown in [30], there exists a general dependency between the intrinsic image brightness and co-located

distance in natural scenes. We use this "the brighter the nearer" law to further guide the estimation of the mean patch depth value using the average luminance of the corresponding image patch. Moreover, in natural scenes, the distance from the nodal point to any point in the scene tends to increase as its height increases. Specifically, if we assume that the y-coordinate of a pixel increases from the bottom to the top of an image, the range values of pixels with larger y-coordinates are generally larger than those with smaller y-coordinates. Thus, we introduce as a second additional feature into the regressor on mean depth values, the normalized y-coordinate of each patch in the image:

$$f_y = \frac{p_y}{I_h} \tag{6.11}$$

where p_y is the y-coordinate of the image patch, and I_h is the height of the image. Thus, in sum, the aggregate feature vector characterizing each image patch used in the regression model to learn mean patch depth values includes the depth-aware feature set $\mathbf{f}_I$, the average patch luminance, and the normalized y-coordinate f_y. In Natural3D, we utilize a standard support vector regressor (SVR) [126] to implement the training and testing processes, using multiple train-test sets as described in Sec. 6.3. SVR is generally noted for being able to effectively handle high dimensional data [127]. We implement the SVR model with a radial basis function (RBF) kernel using the LIBSVM package [128].

6.2.7 Bayesian Model

We now describe how Natural3D incorporates the canonical depth pattern prior model, the likelihood model that associates image patches with different canonical depth patterns, and the regression model that recovers mean patch depth values. Given a test image, the model algorithm first divides it into overlapped patches of size $M \times M$ as in the training phase, where a $\frac{1}{4}$ overlap (stride) is used, i.e., the patches overlap each other by $\frac{M}{4}$ pixels along both dimensions.[1] Next, the depth-aware feature vector $\mathbf{f}_I$ is extracted from each image patch, as well as the average luminance and the normalized y-coordinate, which are used, as described earlier, for mean depth regression. Then, the extracted feature vector $\mathbf{f}_I$ is fed into the trained prior, likelihood, and regression models to form a Bayesian inference of the corresponding estimated depth patch. Specifically, the estimated depth patch $\mathbf{D}$ of an image patch is formed as follows:

$$\mathbf{D} = \mathbf{d}_n + \mu_n \tag{6.12}$$

where $\mathbf{d}_n$ (obtained in Sec. 6.2.4) is the normalized residual depth patch associated with the estimated canonical depth pattern n, μ_n is the corresponding mean depth value obtained from the regression model, and n is the index of the estimated canonical depth pattern derived from the prior and likelihood

[1] In the results described later, we chose $M = 32$, although we have found the model to be robust to this choice.

models, which is given by:

$$\begin{aligned} n &= \underset{n'}{\operatorname{argmax}} \{p(n'|\mathbf{f}_I)\} = \underset{n'}{\operatorname{argmax}} \{p(\mathbf{f}_I|n')p(n')\} \\ &= \underset{n'}{\operatorname{argmax}} \{p(\mathbf{f}_I; \boldsymbol{\theta}_{n'})p(n')\} \end{aligned} \tag{6.13}$$

where $p(\mathbf{f}_I|n') = p(\mathbf{f}_I; \boldsymbol{\theta}_{n'})$ is the likelihood (Eq. (6.10)) of encountering an image patch having the extracted feature vector $\mathbf{f}_I$ given a canonical depth pattern n', as was derived in Sec. 6.2.5, and where $p(n')$ is the corresponding prior probability (ratio) of the estimated canonical depth pattern n', as was obtained in Sec. 6.2.4.

6.2.8 Stitching

The last stage of the overall depth estimation system is to stitch all of the depth patches together to create a final estimated depth map using only the monocular test image as input. Seeking simplicity, we define the stitching operation to simply average the estimated range values of overlapped pixels across the assembled depth patches.

6.3 Experimental Results

To evaluate the performance of Natural3D, we trained and tested the proposed Bayesian model on two publicly accessible databases, the LIVE Color+3D Database Release-2 [8] and the Make3D Laser+Image Dataset-1 [10–12].

6.3.1 Databases

The LIVE Color+3D Database Release-2 consists of 99 pairs of color images and accurately co-registered ground-truth depth maps, all with a high-definition resolution of 1920×1080. The database was constructed using an advanced range scanner, RIEGL VZ-400 [9], with a Nikon D700 digital single-lens reflex camera mounted on top of it. Careful and precise calibration was executed before data acquisition, and a transformation was applied to achieve an accurate depth-image registration [84]. The dense depth maps in the database provide a rich source of information regarding natural depth statistics, and are also an excellent resource for evaluating depth estimation algorithms (including binocular, since co-registered stereopairs are included). To avoid overlap between training and testing image/depth content, we split the entire database into 80% training and 20% testing subsets at each train-test iteration with no content shared between the training and testing subsets. This train-test procedure was repeated 50 times to ensure that there was no bias introduced due to image/depth training content.

The Make3D Laser+Image Dataset-1 contains a total of 534 pairs of color images and corresponding ground-truth depth maps, where 400 are used for training and 134 for testing with no content overlap. The color images are high resolution 2272×1704, while the ground-truth depth maps are only available at a very low resolution of 55×305. These sparse ground-truth depth maps with the unmatched aspect ratio to color images make the Make3D Laser+Image less than ideal for developing and testing contemporary dense

depth estimation algorithms. However, due to its early availability, it has been widely used for evaluating monocular depth estimation methods. To make a complete comparison, we also trained and tested Natural3D on the Make3D database.

We compared Natural3D with a top-performing state-of-the-art depth estimation method, Depth Transfer [47], which has delivered the best recently reported performance on the Make3D Laser+Image Dataset-1. Depth Transfer first selects candidates from a database by matching a high-level image feature, GIST [129], then it optimizes an energy function to generate the most likely depth map by considering all of the warped candidate depth maps under a set of spatial regularization constraints.

6.3.2 Quantitative Evaluation

We performed a quantitative evaluation on the two examined monocular depth estimation algorithms in terms of four different objective metrics. We report the results obtained using two common error metrics, the relative error (Rel.):

$$\sum_{i=1}^{P} \frac{|\mathbf{D}(x_i, y_i) - \mathbf{D}^*(x_i, y_i)|/\mathbf{D}^*(x_i, y_i)}{P} \tag{6.14}$$

and the root-mean-square error (RMS):

$$\sqrt{\sum_{i=1}^{P} \frac{[\mathbf{D}(x_i, y_i) - \mathbf{D}^*(x_i, y_i)]^2}{P}} \tag{6.15}$$

where $\mathbf{D}(x_i, y_i)$ and $\mathbf{D}^*(x_i, y_i)$ represent the estimated and ground-truth depth map at pixel location (x_i, y_i), respectively, and P is the number of pixels. In

addition, to examine how well a depth estimation method is able to recover relative distances from natural scenes, we also report two different correlation coefficients between the estimated and ground-truth depth values: the Pearson's linear correlation coefficient ρ_p and the Spearman's rank order correlation coefficient ρ_s. These metrics are complementary: ρ_p and ρ_s measure the accuracy and monotonicity, respectively, of the estimated range values by a depth estimation algorithm against the ground-truth range values, where a value of 1 indicates perfect correlation.

As is evident from Table 6.1 and 6.2, which show the median metric performance on the LIVE Color+3D Database Release-2 (across train-test splits) and the Make3D Laser+Image Dataset-1 (across test scenes), respectively, Natural3D outperforms Depth Transfer in terms of all four objective metrics. The higher correlation performance of Natural3D indicates that it is capable of recovering more accurate relative distances between the distinct objects and regions that occur in natural scenes. In addition, Natural3D achieves both lower relative and RMS errors than Depth Transfer, meaning that the depth maps estimated by Natural3D are closer to the ground-truth values. Table 6.3 and 6.4 show the standard deviations of the different performance metrics on the two 3D databases, which reflect the performance consistencies of the two examined depth estimation algorithms. Clearly, Natural3D delivers more consistent performance in terms of both linear and rank order correlation, while providing similar or better performance than Depth Transfer.

Table 6.1: Performance Comparison of Monocular Depth Estimation Algorithms on the LIVE Color+3D Database Release-2 (Median across 50 Train-Test Splits)

Algorithm	Metric			
	ρ_p	ρ_s	Rel.	RMS
Depth Transfer	0.4196	0.5197	0.6399	13.0671
Natural3D	0.4404	0.5654	0.5969	12.8417

Table 6.2: Performance Comparison of Monocular Depth Estimation Algorithms on the Make3D Laser+Image Dataset-1 (Median across 134 Test Scenes)

Algorithm	Metric			
	ρ_p	ρ_s	Rel.	RMS
Depth Transfer	0.5184	0.7572	0.6840	17.3187
Natural3D	0.5892	0.7670	0.5548	16.6251

Table 6.3: Performance Comparison of Monocular Depth Estimation Algorithms on the LIVE Color+3D Database Release-2 (Standard Deviation across 50 Train-Test Splits)

Algorithm	Metric			
	ρ_p	ρ_s	Rel.	RMS
Depth Transfer	0.2205	0.2461	0.3891	8.3052
Natural3D	0.1987	0.2253	0.3858	8.3921

Table 6.4: Performance Comparison of Monocular Depth Estimation Algorithms on the Make3D Laser+Image Dataset-1 (Standard Deviation across 134 Test Scenes)

Algorithm	Metric			
	ρ_p	ρ_s	Rel.	RMS
Depth Transfer	0.2016	0.1882	0.4929	7.8353
Natural3D	0.1976	0.1563	0.4524	5.6753

6.3.3 Visual Comparison

In addition to the quantitative comparison, we also supply a visual comparison by showing examples of the depth maps estimated by the two

compared depth estimation algorithms along with the corresponding ground-truth depth maps, as shown in Figs. 6.3 to 6.6 (from the LIVE Color+3D Database Release-2) and Figs. 6.7 to 6.10 (from the Make3D Laser+Image Dataset-1). Note that for the Make3D Laser+Image Dataset-1, the ground-truth and estimated depth maps are scaled to match the image resolution for display purposes. We also supply scatter plots between the estimated and the ground-truth range values to gain a broader perspective of performance.

Generally, Depth Transfer tends to over-smooth the estimated depth maps due to its smoothness constraint, while Natural3D is able to capture more detailed depth structures in the scene. For example, in Fig. 6.3, Depth Transfer is not able to capture the tree trunks in the foreground, while it also incorrectly merges the tree trunks in the background. By comparison, Natural3D creates a clearer representation of the foreground tree trunks, achieving much higher linear and rank order correlations against the ground-truth depth map. Figure 6.5 shows a number of human objects and a tree branch, posing more challenging content for monocular depth estimation algorithms. Natural3D successfully captures details such as the intersection of the human hand and the tree branch, while Depth Transfer fails to recover such complicated structures due to over-smoothing. Similarly, Figure 6.6 shows that Natural3D accurately reconstructs the main tree structures, while Depth Transfer incorrectly combines separate tree trunks.

In Fig. 6.7, Depth Transfer incorrectly extends the ground over the building and the sky, while Natural3D is able to separate the ground and the

Table 6.5: Computational Complexity of Monocular Depth Estimation Algorithms

Algorithm	Runtime per Estimated Depth Map (s)
Depth Transfer	1490.53
Natural3D	161.05

building, as well as most of the sky. Similarly, the trees and sky in the background of Fig. 6.8 are missing in the estimated range map by Depth Transfer, but Natural3D successfully reconstructs most of them. In both Fig. 6.9 and 6.10, Natural3D is capable of recovering the tree depth structures, as well as identifying the sky in the background, while Depth Transfer only captures the ground.

The correlation coefficients shown in all of the figures support the results of visual inspection, confirming that Natural3D achieves superior performance at recovering relative distances in natural scenes. This is accomplished without introducing a smoothness constraint or other iterative procedure into the Bayesian model.

The complete experimental results, including quantitative and visual comparison, of the two examined monocular depth estimation algorithms on every image in both databases can be found at [130].

6.3.4 Computational Complexity

Another advantage of Natural3D is that there is no need for an iterative solution process, resulting in greatly reduced computational complexity. Table 6.5 shows the runtime per estimated depth map for the two examined

algorithms. Both algorithms were implemented using the MATLAB programming language, and the simulations were run on an Intel Core i7 quad-core processor with 16GB memory. Since Natural3D utilizes trained prior and likelihood models, it runs almost 10 times faster than Depth Transfer, which uses an iterative procedure to solve an optimization function.

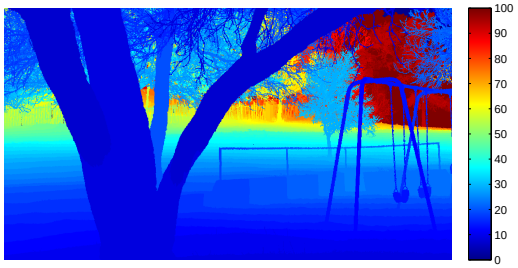
6.4 Summary

By exploiting reliable and robust statistical models describing the relationships between luminances and depths in natural scenes, we have created a Bayesian model, termed Natural3D, for recovering depth information from monocular (photographic) natural images. Two component models are learned from ground-truth depth maps: a prior model, including a dictionary of canonical depth patterns, and a likelihood model, which embeds co-occurrences of image and depth characteristics in natural scenes. When compared to a top-performing state-of-the-art method, it delivers superior performance in the estimation of both absolute and relative depths from natural images.

The superior performance attained by Natural3D implies that a biological visual system might be able to capture coarse depth estimates of the environment using the statistical information computed from retinal images at hand and the associations between image textures and true 3D geometric structures. We believe that the prior and likelihood models developed in Natural3D not only yield insights into how 3D structures in the environment might be recovered from image data, but could be used to benefit a variety of

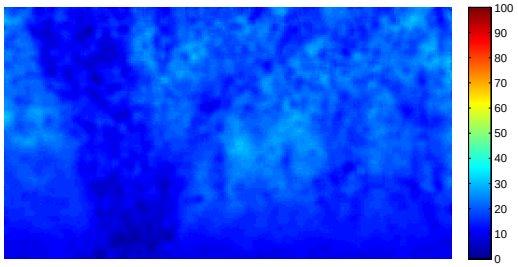
3D image/video and vision algorithms. We envision that our future work will involve introducing deeper statistical models relating image and range data to recover more accurate and detailed depth information.

(a) Natural image

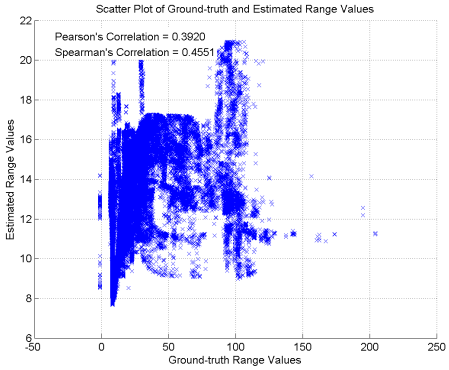


(b) Ground-truth depth map

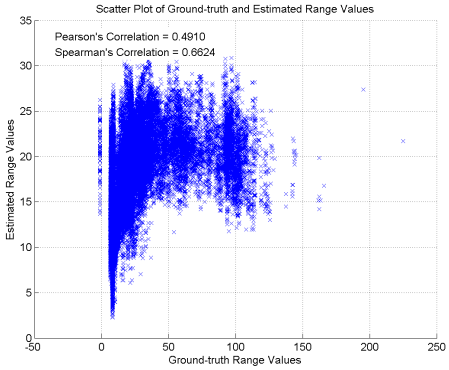
(c) Estimated depth map by Depth Transfer



(d) Estimated depth map by Natural3D



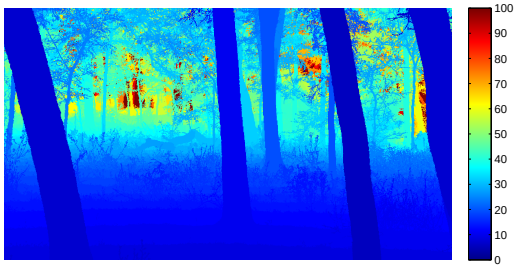
(e) Scatter plot of Depth Transfer result



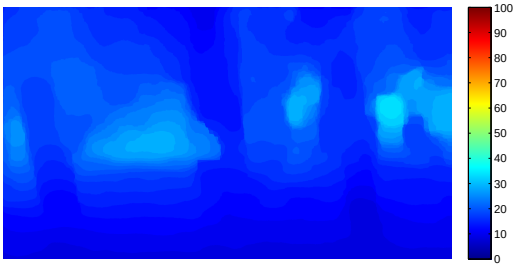
(f) Scatter plot of Natural3D result

Figure 6.3: Example result of estimated depth maps along with the ground-truth depth map on the LIVE Color+3D Database Release-2.

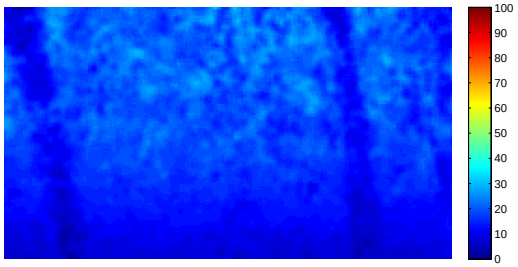
(a) Natural image



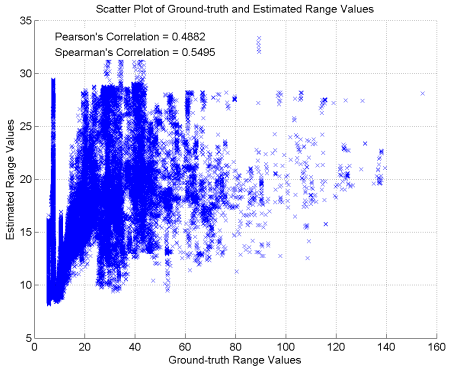
(b) Ground-truth depth map



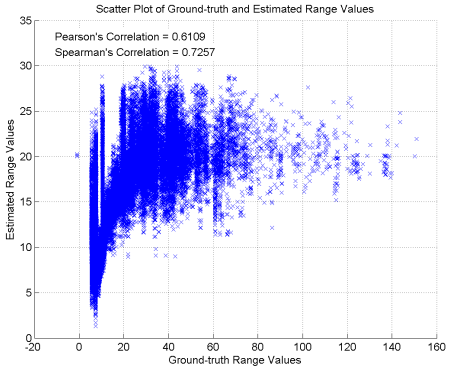
(c) Estimated depth map by Depth Transfer



(d) Estimated depth map by Natural3D



(e) Scatter plot of Depth Transfer result



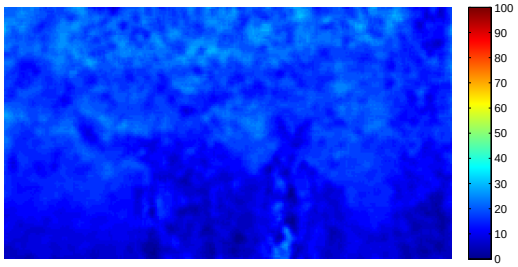
(f) Scatter plot of Natural3D result

Figure 6.4: Example result of estimated depth maps along with the ground-truth depth map on the LIVE Color+3D Database Release-2.

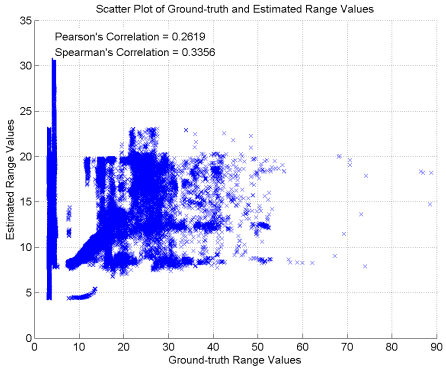
(a) Natural image

(b) Ground-truth depth map

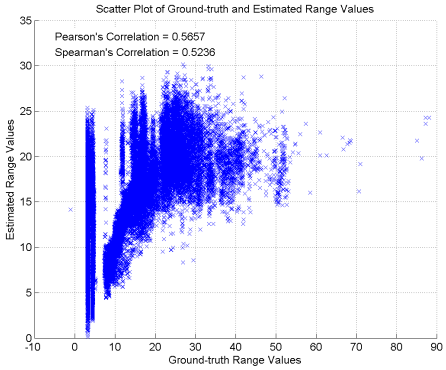
(c) Estimated depth map by Depth Transfer



(d) Estimated depth map by Natural3D



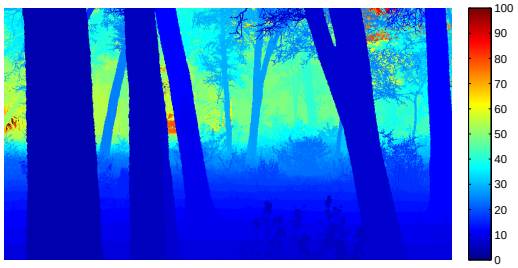
(e) Scatter plot of Depth Transfer result



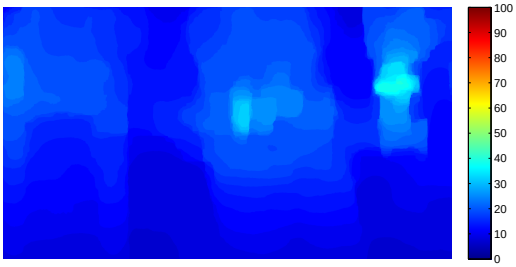
(f) Scatter plot of Natural3D result

Figure 6.5: Example result of estimated depth maps along with the ground-truth depth map on the LIVE Color+3D Database Release-2.

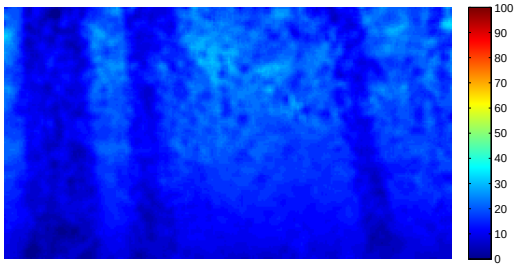
(a) Natural image



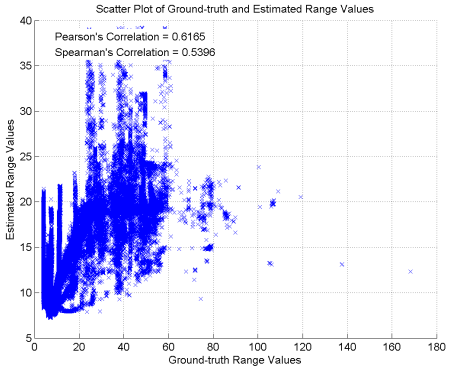
(b) Ground-truth depth map



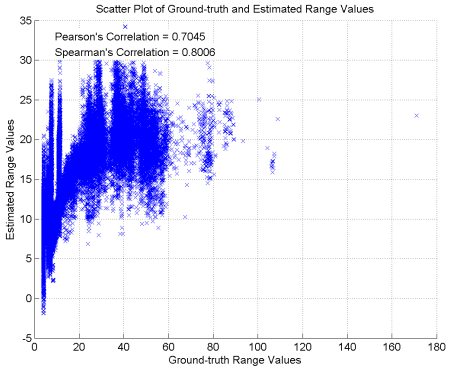
(c) Estimated depth map by Depth Transfer



(d) Estimated depth map by Natural3D



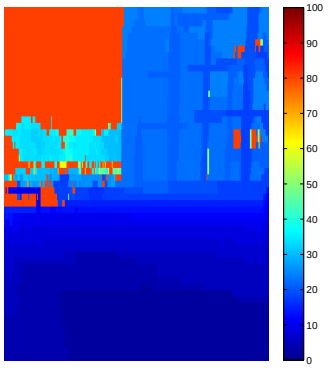
(e) Scatter plot of Depth Transfer result



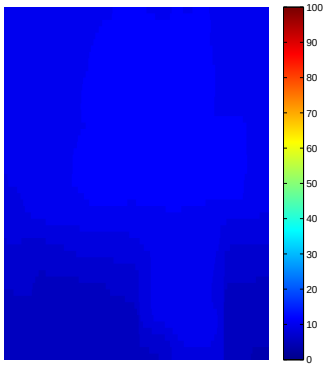
(f) Scatter plot of Natural3D result

Figure 6.6: Example result of estimated depth maps along with the ground-truth depth map on the LIVE Color+3D Database Release-2.

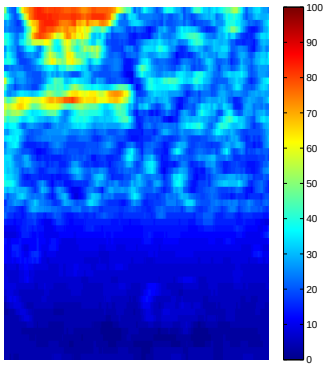
(a) Natural image



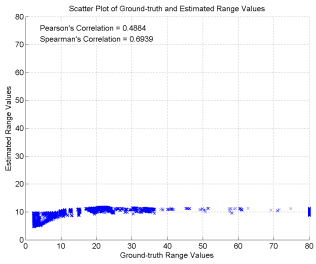
(b) Ground-truth depth map



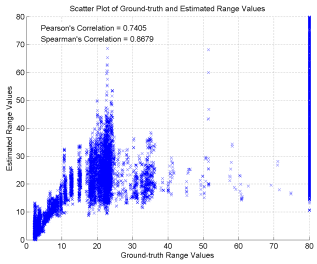
(c) Estimated depth map by Depth Transfer



(d) Estimated depth map by Natural3D



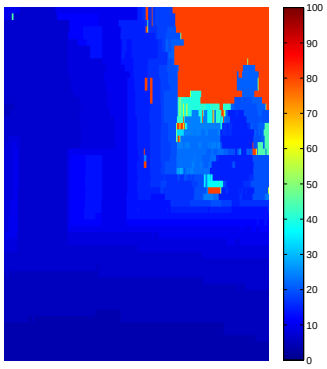
(e) Scatter plot of Depth Transfer result



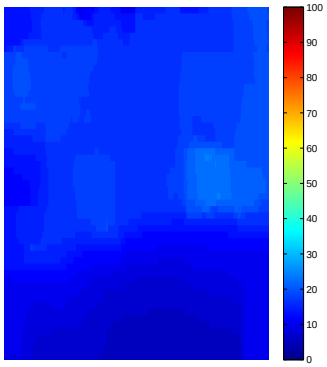
(f) Scatter plot of Natural3D result

Figure 6.7: Example result of estimated depth maps along with the ground-truth depth map on the Make3D Laser+Image Dataset-1.

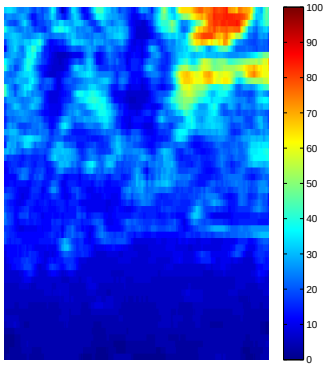
(a) Natural image



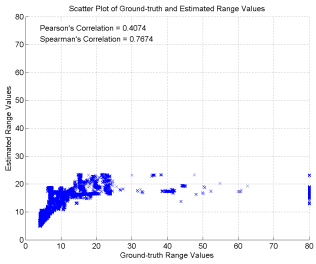
(b) Ground-truth depth map



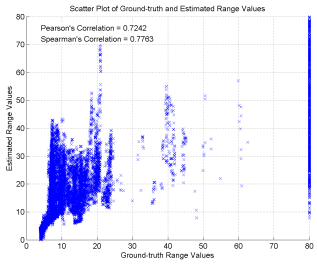
(c) Estimated depth map by Depth Transfer



(d) Estimated depth map by Natural3D



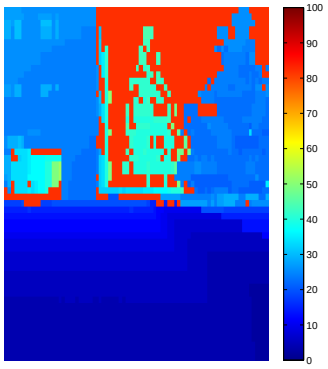
(e) Scatter plot of Depth Transfer result



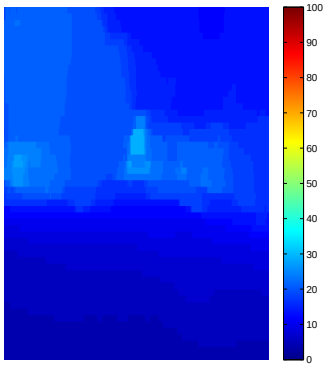
(f) Scatter plot of Natural3D result

Figure 6.8: Example result of estimated depth maps along with the ground-truth depth map on the Make3D Laser+Image Dataset-1.

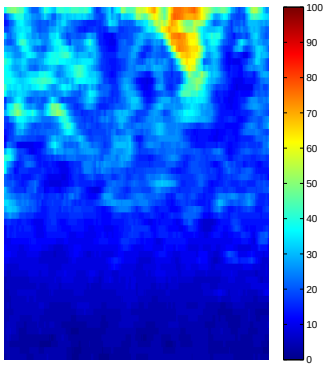
(a) Natural image



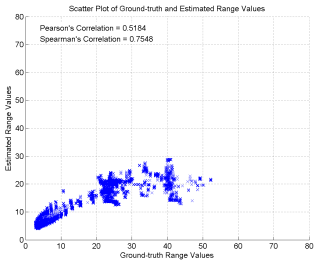
(b) Ground-truth depth map



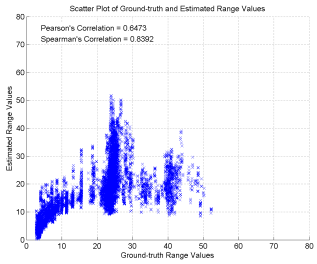
(c) Estimated depth map by Depth Transfer



(d) Estimated depth map by Natural3D



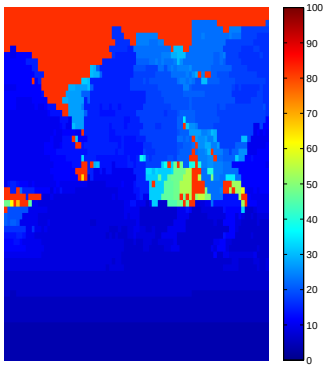
(e) Scatter plot of Depth Transfer result



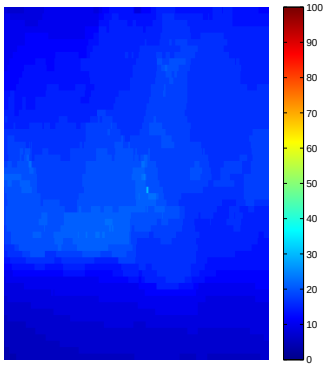
(f) Scatter plot of Natural3D result

Figure 6.9: Example result of estimated depth maps along with the ground-truth depth map on the Make3D Laser+Image Dataset-1.

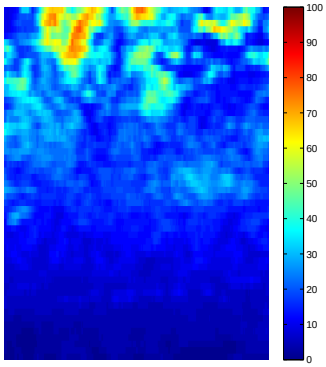
(a) Natural image



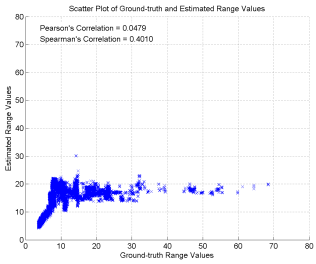
(b) Ground-truth depth map



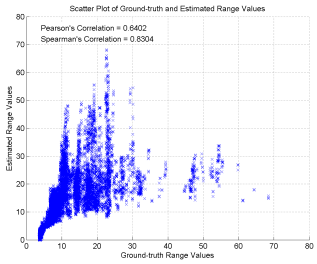
(c) Estimated depth map by Depth Transfer



(d) Estimated depth map by Natural3D



(e) Scatter plot of Depth Transfer result



(f) Scatter plot of Natural3D result

Figure 6.10: Example result of estimated depth maps along with the ground-truth depth map on the Make3D Laser+Image Dataset-1.

Chapter 7

No-Reference Stereoscopic/3D Image Quality Assessment

7.1 Introduction

In recent years, bandpass statistical models of natural, photographic images of the world have been used with great success to solve highly diverse problems involving image representation, image repair, image quality assessment, and image compression. One missing element has been a reliable and generic model of spatial image correlation that reflects the distributions of oriented spatial structures. We have developed such a model for bandpass pristine images [51] and have generalized it here to also capture the spatial correlation structure of bandpass distorted images. The model applies well to both luminance and depth images [50]. As a demonstration of the usefulness of the new model, we develop a new no-reference stereoscopic/3D image quality assessment (IQA) framework, dubbed Stereoscopic/3D BLind Image Naturalness Quality (S3D-BLINQ) Index, which utilizes both univariate and new bivariate natural scene statistics (NSS) models. We first validate the robustness and effectiveness of the new bivariate and correlation NSS features extracted from distorted stereopairs, then demonstrate that they are predictive of distortion severity. Our experimental results show that the resulting 3D

image quality predictor based in part on the new model outperforms state-of-the-art full- and no-reference 3D IQA algorithms on both symmetrically and asymmetrically distorted stereoscopic image pairs.

7.2 Bivariate and Correlation NSS Models

It has been demonstrated that the distributions of divisive-normalized bandpass responses to natural, photographic images possess strong Gaussian-like regularities. Natural scene statistical models of this type have been widely deployed in numerous image and video applications with success [23, 25, 100]. However, less progress has been made on the development of accurate and general statistical models of the higher-order dependencies that exist between spatially neighboring bandpass image responses.

Towards developing such models of the bivariate, correlation statistics of oriented, bandpass images, we utilize the steerable pyramid decomposition of images, which is an over-complete wavelet transform that allows for increased orientation selectivity [115]. The wavelet transform model is motivated by the fact that its space-scale-orientation decomposition broadly resembles the band-pass filtering that occurs in area V1 of primary visual cortex [14, 116]. After applying the multi-scale, multi-orientation decomposition, the perceptually significant process of divisive normalization is applied to the image wavelet coefficients of all sub-bands [117]. The divisive normalization transform (DNT)

used here is implemented as follows [118]:

$$S(x_i, y_i) = \frac{s(x_i, y_i)}{\sqrt{c_s + \mathbf{s_g}^\top \mathbf{s_g}}} = \frac{s(x_i, y_i)}{\sqrt{c_s + \sum_j g(x_j, y_j) s(x_j, y_j)^2}} \tag{7.1}$$

where (x_i, y_i) are spatial coordinates, s represents the sub-band wavelet coefficients, S represents the coefficients after DNT, and c_s is a semi-saturation constant. The sum occurs over neighborhood pixels indexed by j, where $\{g(x_j, y_j)\}$ is a Gaussian weighting function.

Previous work by others and ourselves [51, 131] showed that the empirical joint histograms of spatially adjacent sub-band coefficients of natural images can be well fitted by multivariate generalized Gaussian distribution (MGGD) models, which include both multivariate Gaussian and Laplace distributions as special cases. The probability density function of a multivariate generalized Gaussian distribution (MGGD) is defined as:

$$p(\mathbf{x}; \mathbf{M}, \alpha_b, \beta_b) = \frac{1}{|\mathbf{M}|^{\frac{1}{2}}} g_{\alpha_b, \beta_b}(\mathbf{x}^\top \mathbf{M}^{-1} \mathbf{x}) \tag{7.2}$$

where $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{M}$ is an $N \times N$ symmetric scatter matrix, α_b and β_b are the scale and shape parameters, respectively, and $g_{\alpha_b, \beta_b}(\cdot)$ is a density generator defined as:

$$g_{\alpha_b, \beta_b}(y) = \frac{\beta_b \Gamma(\frac{N}{2})}{(2^{\frac{1}{\beta_b}} \pi \alpha_b)^{\frac{N}{2}} \Gamma(\frac{N}{2\beta_b})} e^{-\frac{1}{2}(\frac{y}{\alpha_b})^{\beta_b}} \tag{7.3}$$

where $y \in \mathbb{R}^+$. Note that when $\beta_b = 0.5$, Eq. (7.2) becomes a multivariate Laplacian distribution, and when $\beta_b = 1$, Eq. (7.2) corresponds to a multivariate Gaussian distribution. Moreover, when $\beta_b \to \infty$, the MGGD converges to

a multivariate uniform distribution. In our recent work, we have become interested in modeling the bivariate empirical histograms of horizontally adjacent sub-band coefficients of both 2D and 3D (cyclopean) images using the bivariate generalized Gaussian distribution (BGGD) with $N = 2$. The parameters of the BGGD can be obtained on the bandpass coefficients of images using the maximum likelihood estimator (MLE) algorithm described in [50].

By examining the fitted BGGD models of pristine bandpass images, i.e., not artificially subjected to, or containing any noticeable distortions, we have found orientation dependencies between spatially adjacent sub-band image coefficients [51]. In particular, when the spatial relationship of adjacent responses, e.g., horizontal, matches the sub-band orientation, e.g., $\frac{1}{2}\pi$, the joint distribution of the responses becomes peaky and extremely elliptical. On the other hand, when the spatial relationship and the sub-band orientation approach orthogonality, e.g., horizontal vs. 0 (rad), the joint distribution becomes nearly circular and more Gaussian-like.

We can seek to quantitatively capture this statistical dependency on relative orientation from an interesting, systematic and potentially useful perspective by directly modeling the correlation entries embedded in the scatter matrix $\mathbf{M}$ of the BGGD model. We define relative orientation to be the difference between the sub-band tuning orientation and the spatial orientation of adjacent responses. For example, if the sub-band tuning orientation is 0 (rad), and the pixels are horizontally adjacent, i.e., the spatial orientation is $\frac{1}{2}\pi$, meaning that they are orthogonal, then the corresponding relative ori-

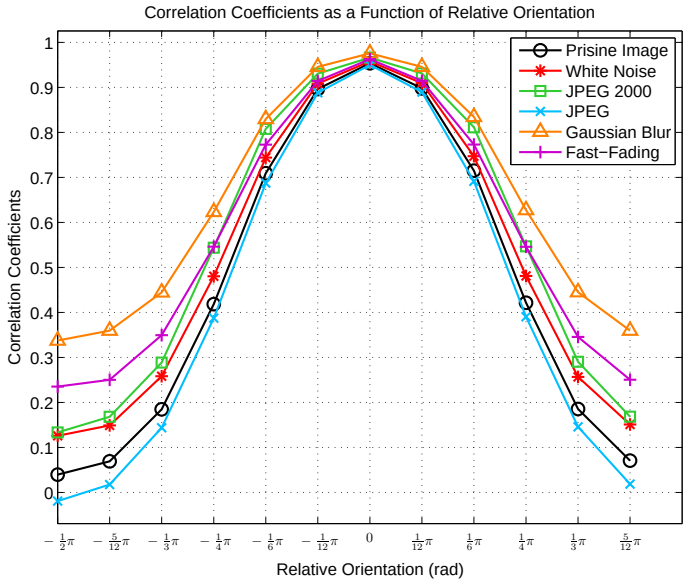


Figure 7.1: The correlation coefficients between spatially adjacent sub-band responses as a function of relative orientation.

entation is equal to $0 - \frac{1}{2}\pi = -\frac{1}{2}\pi$. We have observed that the correlation coefficient between horizontally adjacent bandpass responses reaches a maximum when the sub-band orientation is $\frac{1}{2}\pi$, whereas horizontal neighboring responses become nearly uncorrelated as the sub-band tuning orientation approaches 0 (rad) and π. In addition, the correlation takes symmetric values around sub-band orientation equal to $\frac{\pi}{2}$ and is periodic with respect to all possible relative orientations between the pairs of horizontally adjacent sub-band responses and their orientation tuning. In particular, we observed that the correlation coefficients, when plotted as a function of relative orientation, have a roughly sinusoidal shape but with narrowed lobes. We also observed that when images are distorted by commonly occurring impairments, such as

blur, noise, compression or those found in publicly available image quality databases, the correlation plots as a function of relative orientation also take roughly sinusoidal shapes but with different degrees of lobe narrowness. Figure 7.1 illustrates an example of the correlation coefficient plots as a function of relative orientation for both pristine and impaired images afflicted by different types of distortions from the LIVE IQA Database [114]. We can clearly see that all curves possess sinusoidal-like shapes, but with different degrees of lobe narrowness depending on the distortion type. This strongly suggested to us that a successful model of the sub-band correlations could enable us to capture the effects of distortions on sub-band correlations.

We have found that the periodic relative orientation dependency of the correlation coefficients can be well modeled as an exponentiated sine function given by:

$$y = f(x_1, x_2) = A \left[\frac{1 + \sin\left[\frac{2\pi(x_2 - x_1)}{T} + \theta\right]}{2} \right]^{\gamma} + c_e \tag{7.4}$$

where y is the correlation coefficients between spatially adjacent bandpass responses, x_1 and x_2 represent spatial and sub-band tuning orientations, respectively, A is the amplitude, T is the period, θ is the phase, γ is the exponent, and c_e is the offset. The correlation coefficient is period-π in relative orientation and reaches a maximum when $x_2 - x_1 = k\pi, k \in \mathbb{Z}$, yielding a three-parameter exponentiated sine model with amplitude A, exponent γ, and offset c_e, by

fixing $T = \pi$ and $\theta = \frac{\pi}{2}$:

$$y = f(x_1, x_2) = A \left[\frac{1 + \cos\left[2(x_2 - x_1)\right]}{2} \right]^{\gamma} + c_e \tag{7.5}$$

The model holds well for both undistorted images and images corrupted by common distortions. On undistorted images, $A \approx 1$, $c_e \approx 0$, and γ falls within a certain range [51].

Since we developed this new relative orientation correlation model in the context of our work on natural 3D statistics and their applications, we validate the generalized BGGD and exponentiated sine models on perceptually relevant 'cyclopean' images, formed from the left and right images of a stereopair, as discussed in Section 7.3.2. In the sequel, features based on these natural scene statistical models are used to quantify the degree of perceptual impairment of viewed stereo images.

7.3 Natural Stereopair Quality Index

Inspired by the success of 2D image/video quality assessment algorithms that use 2D natural scene statistics, we have developed a no-reference natural stereopair quality index (S3D-BLINQ Index), which achieves high correlations with human subjective judgments of S3D image quality using the new bivariate and correlation NSS models explained in Section 7.2, along with a symmetrically defined model of the cyclopean image, to extract robust, effective features for S3D image quality prediction.

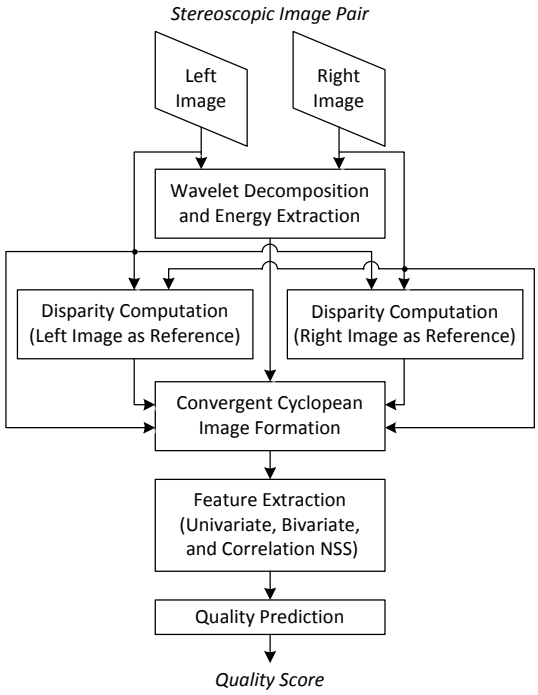


Figure 7.2: S3D-BLINQ Index framework.

7.3.1 Framework Overview

Figure 7.2 diagrams the processing flow of the proposed S3D-BLINQ Index framework. S3D-BLINQ Index first forms a convergent cyclopean image using disparity maps computed from both left- and right-view images as references. Next, both spatial-domain and wavelet-domain univariate NSS features, as well as the bivariate and correlation NSS features introduced in Section 7.2, are extracted from the convergent cyclopean image. Finally, the perceptual quality of S3D images is predicted by mapping the extracted features to human

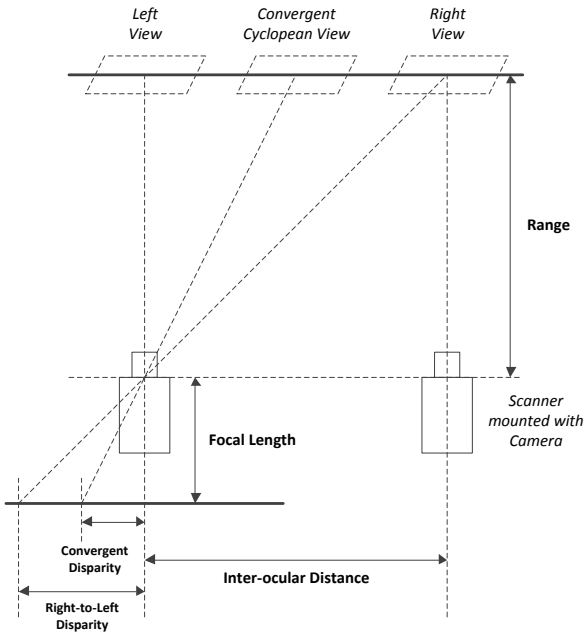


Figure 7.3: Parallel-viewing geometry for generating the convergent cyclopean image from the left and right images.

opinion scores. Each step is detailed in the following sub-sections.

7.3.2 Convergent Cyclopean Image Formation

The synthesized cyclopean image model adopted in prior work was formed by fixing the left-view image, then warping the right-view image onto the corresponding left-view image coordinates using a disparity map computed by a stereo algorithm, using the left-view image as reference. Of course, the binocular human vision system does not form a cyclopean 3D precept using the left- or right-view image as reference; instead, it synergistically fuses the

two views into an intermediate image on a coordinate frame defined relative to fixation [78]. A biased cyclopean image model may fail to capture certain parts of the 3D world accurately, e.g., near depth discontinuities, when the S3D image pair is asymmetrically distorted. To address this distinction, we deploy a more complete and hence perceptually relevant model of the convergent cyclopean image using a general parallel-viewing geometric model of practical S3D image display scenarios.

Admittedly, simulating the true cyclopean image associated with a given stereoscopic image pair is a daunting task, since it requires accounting for a variety of issues, including the display geometry, vergent gaze direction, fixation, accommodation, etc. It is still generally unclear how human vision systems form a cyclopean image from the two visual stimuli received via the retinas of the two eyes. However, the simple linear model proposed by Levelt [132], which remains the most widely-used cyclopean image model, explains the formation of the perceived cyclopean image I_C experienced when a stereoscopic stimulus is presented as a linear combination of neural representations of the stimuli I_L and I_R to the left and right eyes:

$$I_C = w_L \cdot I_L + w_R \cdot I_R \tag{7.6}$$

where w_L and w_R are the weighting coefficients on the corresponding stimuli with the constraint $w_L + w_R = 1$. Levelt further hypothesized that the duration of a dominance period of an eye depends on the stimulus strength in the other eye, making the weighting coefficients positively correlated with the relative

stimulus strengths between the two eyes. Therefore, given a stereoscopic image pair with an associated disparity map computed from them, a synthesized cyclopean image can be obtained by disparity-compensating and mapping the two images of a stereopair onto the same coordinate system. Assuming the disparity map is computed using the left image as reference to match the right image, the synthesized cyclopean image may be generated as [79, 81, 82]:

$$\begin{aligned} I_C(x,y) = {} & w_L(x,y) \cdot I_L(x,y) \\ & + w_R(x - D_L(x,y), y) \cdot I_R(x - D_L(x,y), y) \end{aligned} \tag{7.7}$$

where (x, y) are spatial pixel coordinates, I_L and I_R are left and right image representations (e.g., luminance or bandpass luminance), respectively, w_L and w_R are the weighting maps, and D_L is the disparity map computed by matching elements of I_R to those in I_L, i.e., using I_L as reference. However, this synthesized cyclopean image may fail to capture certain characteristics that may affect the perception of an asymmetrically distorted stereopair. For example, consider two asymmetrically distorted stereopairs having the same content, one of them a pristine left-view image with a distorted right-view image, and the other a pristine right-view image with a similarly impaired left-view image. A synthesized cyclopean image arrived at using Eq. (7.7) will possibly generate very different results on these two asymmetrically distorted stereoscopic image pairs, whereas we would ordinarily expect the perception of these stereopairs to be similar. To address this possible bias in the synthesized cyclopean image, we propose a more perceptually relevant model that we call

the convergent cyclopean image. This model has the virtue of, even for symmetrically distorted or undistorted images, providing a larger and collectively consistent set of constraints on this difficult, ill-posed problem.

Without loss of generality and towards practical applications, we adopt a simple parallel-viewing geometry to generate a convergent cyclopean image given a stereoscopic image pair, as illustrated in Fig. 7.3. In principle, a convergent cyclopean image may be formed as a linear combination of both the disparity-compensated left- and right-view images, yielding a coherent, symmetrically defined representation. The simplest approach is to model the convergent disparity as equal to half of the right-to-left or left-to-right disparities computed by a canonical stereo algorithm. Then, the convergent cyclopean image I_{CC} becomes:

$$\begin{aligned} I_{CC}(x,y) = {} & w_L(x + D'_R(x,y), y) \cdot I_L(x + D'_R(x,y), y) \\ & + w_R(x - D'_L(x,y), y) \cdot I_R(x - D'_L(x,y), y) \end{aligned} \tag{7.8}$$

where $D'_R(x,y) = \frac{D_R(x,y)}{2}$ and $D'_L(x,y) = \frac{D_L(x,y)}{2}$ are convergent disparity maps computed using the right and left images as references, respectively, and D_R and D_L are the canonical disparity maps computed using the right and left images as references, respectively.

The stimulus strengths, i.e., the weighting maps w_L and w_R in (7.8), are modeled as the sum of the energies of wavelet coefficients computed using a steerable pyramid, followed by a DNT taken across sub-bands as described in Section 7.2. As a result, the convergent cyclopean image given a stereoscopic

(a) Pristine right view (b) Pristine left view (c) Pristine convergent cyclopean image

(d) JPEG-compressed right view (e) Pristine left view (f) Asymmetrically JPEG-compressed convergent cyclopean image

(g) Gaussian blurred right view (h) Pristine left view (i) Asymmetrically Gaussian blurred convergent cyclopean image

Figure 7.4: Examples of convergent cyclopean images formed by pristine, asymmetrically JPEG-compressed, and asymmetrically Gaussian blurred stereopairs.

image pair is formed as

$$I_{CC}(x,y) = \frac{E_L[x + D'_R(x,y), y]}{E_U(x,y)} \cdot I_L[x + D'_R(x,y), y] + \frac{E_R[x - D'_L(x,y), y]}{E_U(x,y)} \cdot I_R[x - D'_L(x,y), y] \tag{7.9}$$

$$E_L(x,y) = \frac{\sum_{k=1}^{K} S^2_{L_k}(x,y)}{K} \tag{7.10}$$

$$E_R(x,y) = \frac{\sum_{k=1}^{K} S^2_{R_k}(x,y)}{K} \tag{7.11}$$

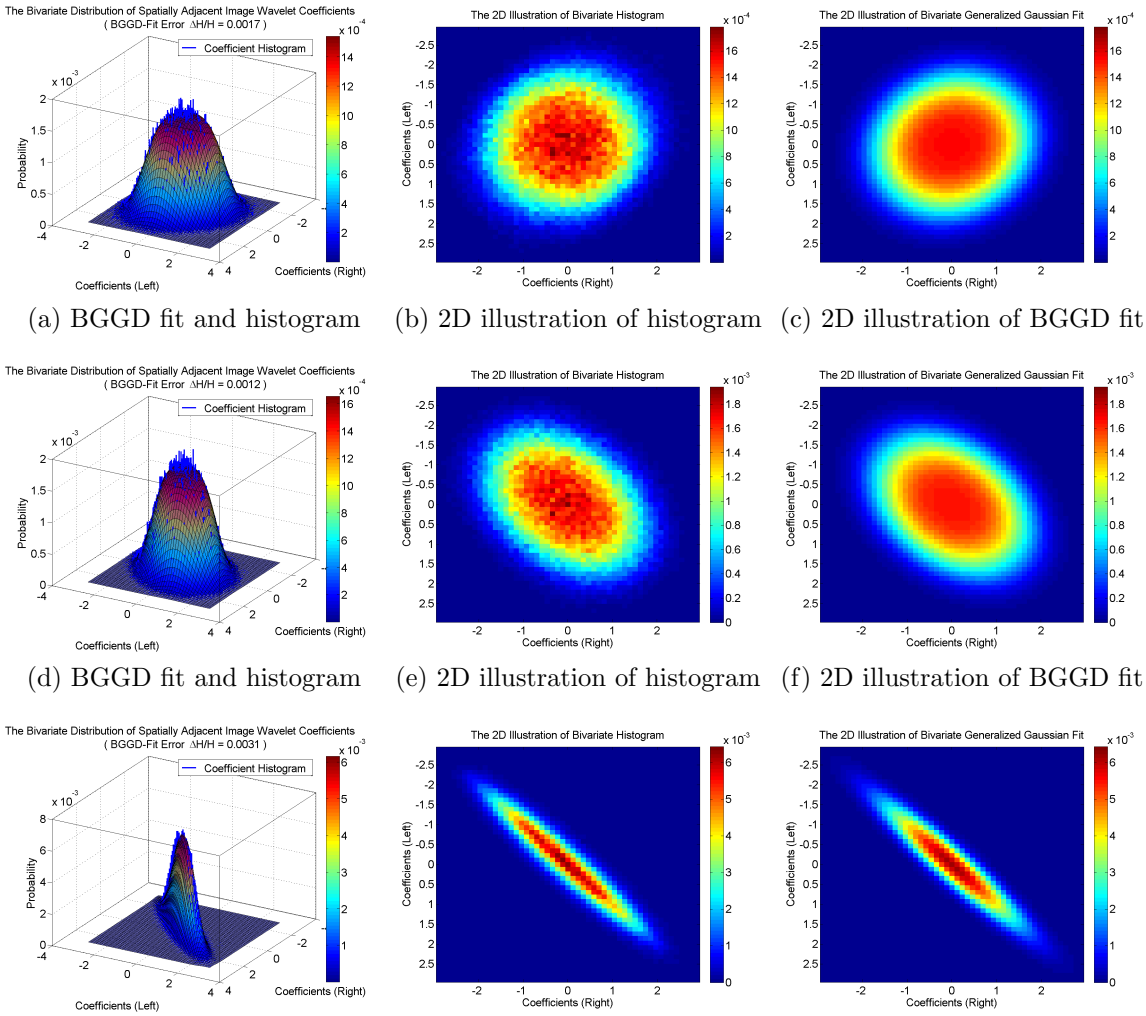
$$E_U(x,y) = E_L[x + D'_R(x,y), y] + E_R[x - D'_L(x,y), y] \tag{7.12}$$

where E_L and E_R are the left and right energy maps, S_{L_k} and S_{R_k} are the left and right sub-band coefficients at sub-band k, K is the number of sub-bands, and E_U serves to achieve unit-sum weighting as in (7.6).

Figure 7.4 shows some examples of convergent cyclopean images formed by pristine and asymmetrically distorted stereopairs. All left-view images are pristine, while the right-view image examples include pristine, JPEG-compressed, and Gaussian blurred images. When a human observes such stereopairs, the deep question arises whether, upon successfully free-fusing left- and right-view images, one would be able to construct a clear 3D percept by some process of distortion masking or whether the view might have an appearance of heightened distortion, perhaps owing to a facilitation of the asymmetric impairments. For example, blocky distortions on the a JPEG-compressed right-view image might be more or less apparent on 3D viewing, while Gaussian blur of one of the images may not reduce the sharpness of the overall 3D percept. Such questions have been explored deeply in the experiments reported in [75] and in more focused studies in the references cited there. In any case, the generated convergent cyclopean images render a means of capturing these perceptual effects.

7.3.3 Spatial-Domain Univariate NSS Feature Extraction

It has been demonstrated that natural scene statistics (NSS) models provide powerful and robust tools for gauging human judgments of visual distortions on 2D images and videos [25, 65, 100]. Early on, Ruderman [133]



(a) BGGD fit and histogram (b) 2D illustration of histogram (c) 2D illustration of BGGD fit

(d) BGGD fit and histogram (e) 2D illustration of histogram (f) 2D illustration of BGGD fit

(g) BGGD fit and histogram (h) 2D illustration of histogram (i) 2D illustration of BGGD fit

Figure 7.5: Joint histograms of horizontally adjacent bandpass coefficients from a pristine convergent cyclopean image and corresponding BGGD fits at the finest scale along different sub-band tuning orientations. Top row: orientation = 0, middle row: orientation = $\frac{\pi}{4}$, and bottom row: orientation = $\frac{\pi}{2}$.

showed that a simple non-linear operation of local mean subtraction and divisive variance normalization on natural image luminances results in a decor-

related, Gaussianized 'contrast' image. This spatial-domain NSS model has been extended in various ways and successfully deployed in no-reference 2D quality assessment algorithms [66, 113] which deliver highly competitive performance relative to top-performing full-reference metrics. We utilize a similar decomposition to extract spatial-domain univariate features from each convergent cyclopean image. First, the luminance of the convergent cyclopean image, I_{CC}, is transformed as:

$$\hat{I}_{CC}(x,y) = \frac{I_{CC}(x,y) - \mu(x,y)}{\sigma(x,y) + c} \tag{7.13}$$

where (x, y) are spatial pixel coordinates, μ and σ are locally weighted spatial means and standard deviations computed using a Gaussian window superimposed over the spatial neighborhood, and $c = 1$ is a constant that ensures stability. To capture a broader spectrum of distortion statistics than Gaussian on convergent cyclopean images, we use the univariate generalized Gaussian distribution (GGD) model to fit the empirical histograms of the contrast images $\hat{I}_{CC}$. The probability density function of a univariate GGD with zero mean is:

$$p(x;\alpha_u,\beta_u) = \frac{\beta_u}{2\alpha_u\Gamma(\frac{1}{\beta_u})} e^{-(\frac{|x|}{\alpha_u})^{\beta_u}} \tag{7.14}$$

where $\Gamma(\cdot)$ is the ordinary gamma function and α_u and β_u are scale and shape parameters, respectively.

We use the moment-matching based approach proposed in [123] to estimate the parameters of the univariate GGD fit. The two extracted univariate

GGD parameters, $[\alpha_u, \beta_u]^\top$, are deployed as spatial-domain 'quality-aware' features.

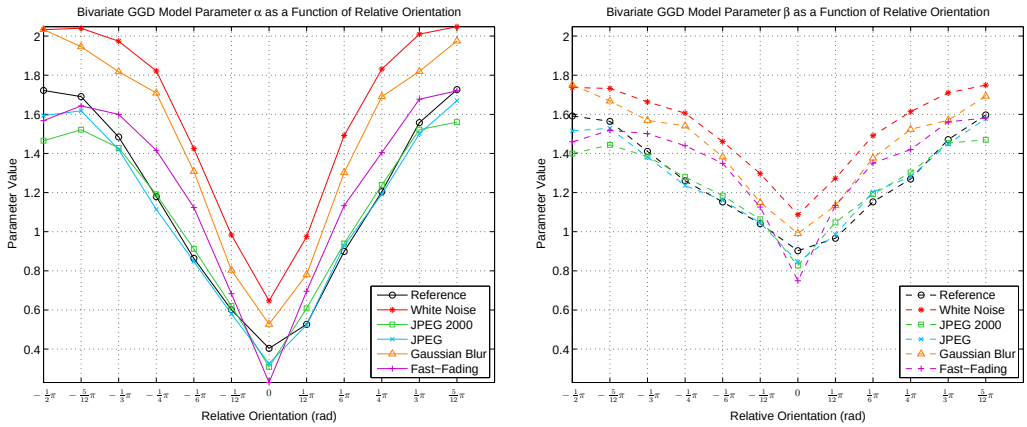
Since the normalizing operation (7.13) is isotropic, we also model the statistical relationships between neighboring pixels along different orientations using the very general univariate asymmetric generalized Gaussian distribution (AGGD) [66, 134]. Specifically, we fit the empirical histograms of pairwise products of adjacent (cardinal and diagonal) coefficients of the convergent cyclopean contrast image, $\hat{I}_{CC}$, using the multi-parameter univariate AGGD probability density function with zero mean:

$$p(x; \alpha_l, \alpha_r, \beta_a) = \begin{cases} \frac{\beta_a}{(\alpha_l+\alpha_r)\Gamma(\frac{1}{\beta_a})} e^{-(\frac{-x}{\alpha_l})^{\beta_a}}, & x < 0 \\ \frac{\beta_a}{(\alpha_l+\alpha_r)\Gamma(\frac{1}{\beta_a})} e^{-(\frac{x}{\alpha_r})^{\beta_a}}, & x \leq 0 \end{cases} \tag{7.15}$$

where β_a is the shape parameter, and α_l and α_r are scale parameters that control the spread of the AGGD to the left and right of the origin. The parameters of the AGGD fits are also estimated using the moment-matching based approach in [123]. All three AGGD parameters, $[\alpha_l, \alpha_r, \beta_a]^\top$, extracted from each S3D convergent cyclopean contrast image are employed as spatial-domain quality features.

7.3.4 Wavelet-Domain Univariate NSS Feature Extraction

Considerable work has focused on modeling the statistics of natural images using multi-scale, multi-orientation transforms, e.g., Gabor filters, wavelets, etc [16, 18]. Success has also been attained by utilizing transform-domain NSS models to create 2D image and video quality assessment models [25, 65, 100].



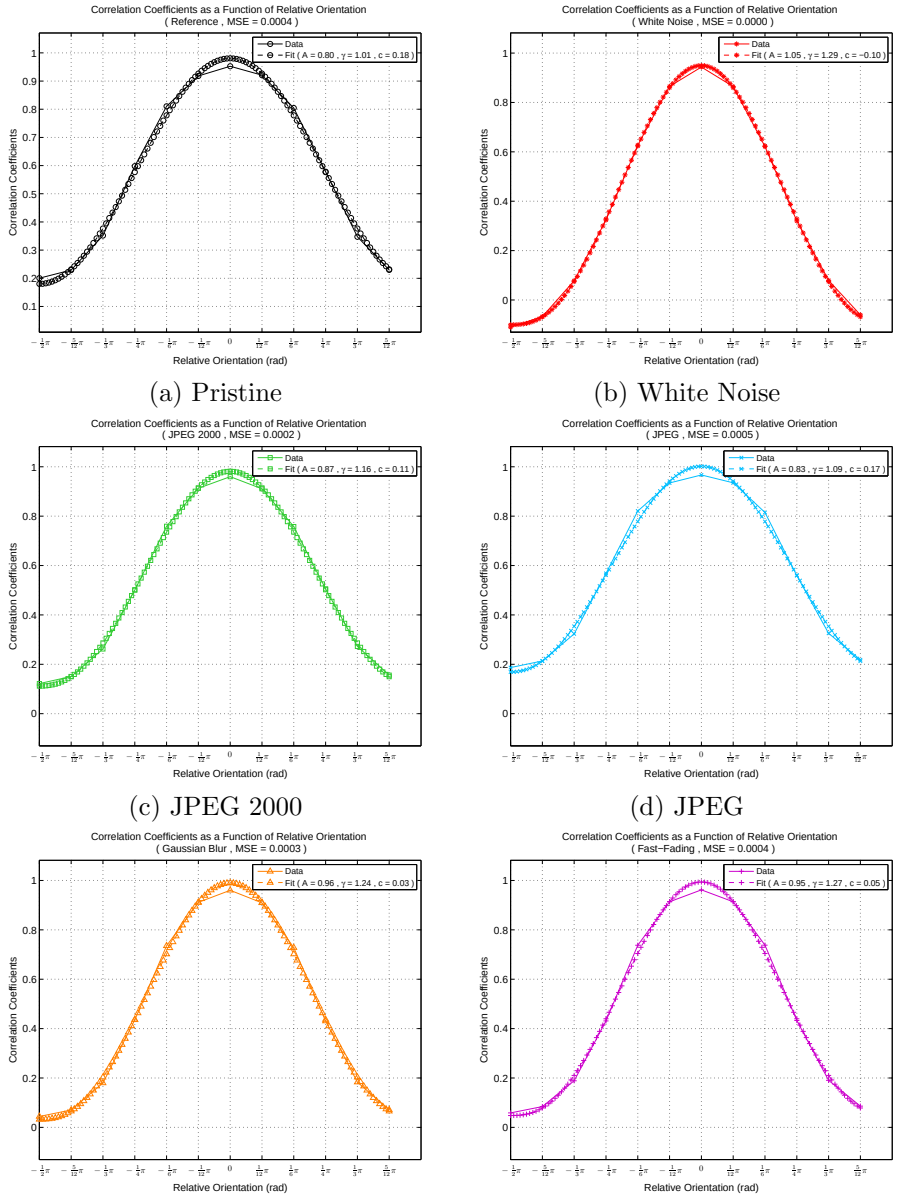
(a) scale parameter α_b (b) shape parameter β_b

Figure 7.6: Plots of the two BGGD model parameters as a function of relative orientation from pristine and distorted convergent cyclopean images.

In these kinds of IQA models, perceptually relevant transform-domain features are computed via area V1-like band-pass filtering. Likewise, we process the convergent cyclopean image I_{CC} using the same steerable pyramid wavelet decomposition as earlier, followed by the divisive normalization transform (Section 7.2). We again use the univariate generalized Gaussian distribution (GGD) to fit the empirical histograms of these sub-band coefficients using (7.14). The two resulting GGD parameters from each sub-band, scale and shape, are included in the wavelet-domain feature set.

7.3.5 Bivariate Density and Correlation NSS Feature Extraction

We employ the new bivariate density and correlation NSS models introduced in Section 7.2 to extract wavelet-domain features from the convergent cyclopean images. First, we validate the efficacy of these new NSS models, by



(a) Pristine (b) White Noise

(c) JPEG 2000 (d) JPEG

(e) Gaussian Blur (f) Fast-Fading

Figure 7.7: Exponentiated sine fits to the curves of correlation coefficients between spatially adjacent wavelet coefficients as a function of relative orientation for distorted convergent cyclopean images.

fitting the BGGD to the empirical histograms of spatially adjacent bandpass coefficients after DNT on convergent cyclopean images formed by undistorted stereopairs. Figure 7.5 shows the joint empirical histograms of horizontally adjacent bandpass coefficients and the corresponding BGGD fits at the finest scale over several different orientations $(0, \frac{\pi}{4}, \frac{\pi}{2})$. In the plots, blue bars represent the true histograms while colored meshes represent the fits. Clearly, the bivariate joint distributions of horizontally adjacent wavelet coefficients are well modeled by BGGD. The 2D figures, which are iso-probability contour maps of the joint distributions, also illustrate the high accuracy of the fits obtained using the BGGD models. The most important observation here is that the shapes and heights of the joint distributions both vary with sub-band orientation. This matches our early findings of BGGD fits on pristine 2D images [51], i.e., there exist much higher dependencies between spatially adjacent pixels after being decomposed by bandpass filters when the orientations are similar.

To portray a clear picture of this relative orientation dependency, we plot the two BGGD model parameters α_b and β_b as a function of relative orientation on convergent cyclopean images afflicted by different types of distortion in Fig. 7.6. Clearly there is a strong relative orientation dependency of both parameters, each reaching minimum value when the spatial orientation matches the sub-band tuning orientation, i.e., $x_2 - x_1 = 0$. Thus, horizontally adjacent sub-band coefficients share the highest correlation when their tuning orientation is $\frac{\pi}{2}$, with the correlation declining away from $\frac{\pi}{2}$. These

plots also show that different types of distortion cause different degrees of relative orientation dependency, which we shall also exploit to develop additional quality-aware correlation features. Specifically, the scale and shape parameters $[\alpha_b, \beta_b]^\top$ are deployed as bivariate NSS quality features.

As discussed in Section 7.2, the relative orientation dependency of spatially adjacent bandpass responses of images is reflected by a systematic behavior of the correlation coefficients. Figure 7.7 shows exponentiated sine function fits to the correlation coefficient plots of horizontally adjacent wavelet coefficients, as a function of relative orientation, on convergent cyclopean images afflicted by several different types of distortion. In agreement with the low mean squared errors (MSE), the appearances of the exponentiated sine model fits are quite close to the orientation-dependent correlation coefficient curves. Note that the model varies with distortion type. We also conducted a multivariate statistical hypothesis test as we describe in the next subsection that illustrates the discriminative power of the exponentiated sine model. Based on these amicable results, we include the exponentiated sine model parameters, $[A, \gamma, c_e]^\top$, as distortion sensitive IQA features.

7.3.6 Validation of the Exponentiated Sine Model

The exponentiated sine model of the correlation coefficients between spatially adjacent sub-band responses was shown to be a reliable model of natural photographic images in [51]. Here we validate the applicability of the more general exponentiated sine model (7.5) for convergent cyclopean images

formed from a diversity of distorted stereoscopic image pairs taken from the LIVE 3D Image Quality Database Phase II [135]. Specifically, we performed a statistical hypothesis test on the three parameters of the exponentiated sine model across all computed convergent cyclopean images for each type of distortion, including the pristine convergent cyclopean images. First, we computed the orientation-dependent correlation coefficient curves at a particular scale for all N_D images afflicted by distortion type D in the database, and obtained corresponding exponentiated sine fits for each distorted cyclopean image. Denote each exponentiated sine fit by a vector $\mathbf{x} = [A, \gamma, c]^\top \in \mathbb{R}^3$, where A, γ, and c are the three model parameters, amplitude, exponent, and offset, respectively. For brevity, we tabulate the results only for the finest scale; however, we obtained similar results for other scales as well (see [136]). Next, we computed the mean model parameter vector across all convergent cyclopean images having distortion type D, denoted $\overline{\mathbf{x}}_D = \sum_{i=1}^{N_D} \mathbf{x}_{D_i}$, where $\mathbf{x}_{D_i} = [A_{D_i}, \gamma_{D_i}, c_{D_i}]^\top$. Then, we applied a two-sample multivariate t-test to determine whether the null hypothesis H_0, i.e., the two mean vectors $\overline{\mathbf{x}}_{D_1}$ and $\overline{\mathbf{x}}_{D_2}$ of two different distortion types D_1 and D_2 are equal, is supported. If the null hypothesis is supported, then the two exponentiated sine models of distortion types D_1 and D_2 are statistically identical. However, if H_0 is rejected, we can conclude that distortion types D_1 and D_2 possess significantly different exponentiated sine models. In particular, we computed Hotelling's two-sample T-squared statistic T^2, which generalizes the Student's two-sample t statistic:

$$T^2 = \frac{N_{D_1} N_{D_2}}{N_{D_1} + N_{D_2}} (\overline{\mathbf{x}}_{D_1} - \overline{\mathbf{x}}_{D_2})^\top \mathbf{S_p}^{-1} (\overline{\mathbf{x}}_{D_1} - \overline{\mathbf{x}}_{D_2}) \tag{7.16}$$

where $\mathbf{S_p}$ is an unbiased estimate of the pooled covariance matrix:

$$\mathbf{S_p} = \frac{\sum_{D_{1_i}=1}^{N_{D_1}} (\mathbf{x}_{D_{1_i}} - \overline{\mathbf{x}}_{D_1})(\mathbf{x}_{D_{1_i}} - \overline{\mathbf{x}}_{D_1})^\top}{N_{D_1} + N_{D_2} - 2} + \frac{\sum_{D_{2_i}=1}^{N_{D_2}} (\mathbf{x}_{D_{2_i}} - \overline{\mathbf{x}}_{D_2})(\mathbf{x}_{D_{2_i}} - \overline{\mathbf{x}}_{D_2})^\top}{N_{D_1} + N_{D_2} - 2} \tag{7.17}$$

Finally, T^2 can be related to the F-distribution as:

$$\frac{N_{D_1} + N_{D_2} - P - 1}{(N_{D_1} + N_{D_2} - 2)P} T^2 \sim F_{P, N_{D_1} + N_{D_2} - P - 1} \tag{7.18}$$

where P is the dimension of $x_{D_{1_i}}$ and $x_{D_{2_i}}$.

Therefore, we are able to compute the p-value of our null hypothesis test as:

$$p = 1 - C_{F_{P, N_{D_1} + N_{D_2} - P - 1}} \left(\frac{N_{D_1} + N_{D_2} - P - 1}{(N_{D_1} + N_{D_2} - 2)P} T^2 \right) \tag{7.19}$$

where $C_{F_{P, N_{D_1} + N_{D_2} - P - 1}}$ represents the cumulative distribution function of the F-distribution. Note that in our test, $P = 3$ and $N_{D_1} = N_{D_2} = 72$. We repeated this null hypothesis test on all different pairs of distortion types, including the pristine, to examine the robustness of the exponentiated sine model.

Table 7.1 records the results of all distortion type pairs, wherein each entry represents whether the null hypothesis between the row and column distortion types is rejected or not. Table 7.2 shows the corresponding p-values for each null hypothesis test. Note that since Hotelling's two-sample T-squared statistic T^2 commutes, the entries in Tables 7.1 and 7.2 are diagonally symmetric. Clearly, at the one-sided significance level $\alpha = 0.05$, all hypothesis

Table 7.1: Statistical Hypothesis Test Results. A Value of '1' Indicates That the Null Hypothesis is Rejected While a Value of '0' Indicates Supported.

	Pristine	WN	JP2K	JPEG	Blur	FF
Pristine	0	1	1	1	1	1
WN	1	0	1	1	1	1
JP2K	1	1	0	1	1	1
JPEG	1	1	1	0	1	1
Blur	1	1	1	1	0	1
FF	1	1	1	1	1	0

Table 7.2: Computed p-Values From Statistical Hypothesis Tests

	Pristine	WN	JP2K	JPEG	Blur	FF
Pristine	-	0	9.12×10^{-08}	8.24×10^{-08}	1.11×10^{-16}	3.21×10^{-09}
WN	0	-	0	0	0	0
JP2K	9.12×10^{-08}	0	-	0	3.05×10^{-13}	3.10×10^{-03}
JPEG	8.24×10^{-08}	0	0	-	4.74×10^{-11}	2.99×10^{-14}
Blur	1.11×10^{-16}	0	3.05×10^{-13}	4.74×10^{-11}	-	3.92×10^{-06}
FF	3.21×10^{-09}	0	3.10×10^{-03}	2.99×10^{-14}	3.92×10^{-06}	-

tests were rejected with p-value $< \alpha$, indicating that every distortion type, including the pristine, can be characterized by a distinct exponentiated sine model. These results not only support the validity of the proposed exponentiated sine model of the relative orientation-dependent correlation coefficients between spatially adjacent sub-band responses across different types of distorted convergent cyclopean images, but also substantiate its relevance for developing quality-predictive features on S3D images.

7.3.7 Quality Prediction

The last step in the proposed S3D-BLINQ Index framework is to predict the quality of stereoscopic image pairs using the aforedescribed NSS features extracted from the corresponding convergent cyclopean images. A mapping

is learned from the feature space to human subjective quality scores using a regression model. The proposed framework is generically amenable to the application of any kind of regressor. The implementation of S3D-BLINQ Index described here utilizes a support vector machine (SVM) regressor (SVR) [126] using multiple train-test sequences as described in the next section. SVR is generally noted for being able to handle high dimensional data [127], and has also been used to create a variety of 2D IQA models [25, 137]. We implement the SVR model with a radial basis function (RBF) kernel using the LIBSVM package [128].

7.4 Experimental Results and Discussion

7.4.1 Performance Evaluation of S3D-BLINQ Index

In the previous section, we motivated and developed a statistics-based S3D IQA framework, dubbed S3D-BLINQ Index, that incorporates old and new models of the univariate and bivariate statistics of natural photographic S3D images. We next evaluated the efficacy of the new S3D IQA model against state-of-the-art 2D and S3D IQA models on the LIVE 3D Image Quality Database Phase II [135], which consists of both symmetrically and asymmetrically distorted stereopairs. There are five different types of distortions in the LIVE 3D Image Quality Database Phase II: JPEG and JPEG2000 (JP2K) compression, additive white Gaussian noise (WN), Gaussian blur (Blur), and a Rayleigh fast-fading channel distortion (FF). The severities of each of the degradations vary significantly and with good overall perceptual separations

between distortion levels. For full-reference algorithms, we used all of the available reference and distorted stereopairs, while for no-reference algorithms, we divided the whole database into 80% training and 20% testing subsets at each train-test iteration so that there was no overlap between training and testing image content. This train-test procedure was repeated 1000 times to ensure that there was no bias introduced due to the image content used for training. We report the median performance across all iterations as the final performance score.

We computed two correlation measures, Spearman's rank-order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (LCC), along with the root-mean-squared error (RMSE) between the predicted quality scores and the recorded subjective opinion scores (DMOS) to evaluate the performance of the quality assessment models [138]. Since both LCC and RMSE are accuracy measures, all algorithm scores were passed through a logistic non-linear function to map to DMOS space before computing LCC and RMSE [138]. The SROCC, LCC, and RMSE values of the tested 2D and S3D IQA models evaluated on the LIVE 3D Image Quality Database Phase II are summarized and tabulated in Tables 7.3 – 7.5. Higher values of the two correlation measures, SROCC and LCC, and lower values of RMSE indicate better performance.

From Tables 7.3 and 7.4, it can be seen that the highest attained performance by any "pure 2D" IQA model on the LIVE S3D image pairs reached about 0.8 correlation against the subjective opinion scores. Again, the quality

Table 7.3: Comparison (SROCC) of Different 2D and 3D Image Quality Assessment Models on Different Distortion Types in the LIVE 3D Image Quality Database Phase II

	Algorithm†	WN	JP2K	JPEG	Blur	FF	Overall
2D	PSNR	0.919	0.597	0.491	0.690	0.730	0.665
	SSIM [63]	0.922	0.704	0.678	0.838	0.834	0.792
	MS-SSIM [64]	0.946	0.798	0.847	0.801	0.833	0.777
	BRISQUE [66]	0.846	0.593	0.769	0.862	0.935	0.770
3D	Benoit [72]	0.923	0.751	0.867	0.455	0.773	0.728
	You [76]	0.909	0.894	0.795	0.813	0.891	0.786
	Yasakethu [67]	0.880	0.598	0.736	0.028	0.684	0.501
	Cyclopean MS-SSIM [81]	0.940	0.814	0.843	0.908	0.884	0.889
	Sazzad [77]	0.714	0.724	0.649	0.682	0.559	0.543
	Chen [82]	0.950	0.867	0.867	0.900	0.933	0.880
	S3D-BLINQ Index	0.946	0.845	0.818	0.903	0.899	0.905

† *Italics* indicate no-reference IQA models. Others are full-reference IQA models.

Table 7.4: Comparison (LCC) of Different 2D and 3D Image Quality Assessment Models on Different Distortion Types in the LIVE 3D Image Quality Database Phase II

	Algorithm†	WN	JP2K	JPEG	Blur	FF	Overall
2D	PSNR	0.917	0.627	0.459	0.706	0.762	0.680
	SSIM [63]	0.928	0.723	0.650	0.848	0.858	0.802
	MS-SSIM [64]	0.950	0.820	0.856	0.798	0.842	0.783
	BRISQUE [66]	0.845	0.681	0.795	0.951	0.931	0.782
3D	Benoit [72]	0.926	0.784	0.853	0.535	0.807	0.748
	You [76]	0.912	0.905	0.830	0.784	0.915	0.800
	Yasakethu [67]	0.891	0.664	0.734	0.450	0.746	0.558
	Cyclopean MS-SSIM [81]	0.957	0.834	0.862	0.963	0.901	0.900
	Sazzad [77]	0.722	0.776	0.786	0.795	0.674	0.568
	Chen [82]	0.947	0.899	0.901	0.941	0.932	0.895
	S3D-BLINQ Index	0.953	0.847	0.888	0.968	0.944	0.913

† *Italics* indicate no-reference IQA models. Others are full-reference IQA models.

scores predicted by these models were obtained by simply averaging the scores computed on the left- and right-view images. Among these 2D quality metrics,

Table 7.5: Comparison (RMSE) of Different 2D and 3D Image Quality Assessment Models on Different Distortion Types in the LIVE 3D Image Quality Database Phase II

	Algorithm†	WN	JP2K	JPEG	Blur	FF	Overall
2D	PSNR	4.269	7.674	6.514	9.865	7.456	8.275
	SSIM [63]	3.988	6.783	5.572	7.370	5.910	6.741
	MS-SSIM [64]	3.334	5.621	3.792	8.397	6.212	7.025
	BRISQUE [66]	5.731	7.193	4.448	4.323	4.206	7.038
3D	Benoit [72]	4.028	6.096	3.787	11.76	6.894	7.490
	You [76]	4.396	4.186	4.086	8.649	4.649	6.772
	Yasakethu [67]	10.71	7.343	4.976	12.43	7.667	9.364
	Cyclopean MS-SSIM [81]	3.368	5.562	3.865	3.747	4.966	4.987
	Sazzad [77]	7.416	6.189	4.535	8.450	8.505	9.294
	Chen [82]	3.513	4.298	3.342	4.725	4.180	5.102
	S3D-BLINQ Index	3.547	5.482	4.169	4.453	4.199	4.657

† *Italics* indicate no-reference IQA models. Others are full-reference IQA models.

the full-reference SSIM index achieved the best performance.

The best S3D image quality prediction models that utilize 3D information were able to deliver a significantly higher 0.9 correlation level of performance. In particular, utilizing a synthesized cyclopean image boosts the performance of simple 2D IQA models, such as MS-SSIM, by more than 0.1 correlation level. By combining a synthesized cyclopean image with statistical models of disparity statistics, the no-reference S3D IQA model proposed by Chen *et al.* [82] was able to deliver performance comparable to the best full-reference models. However, as mentioned earlier, perceptual issues arise when forming cyclopean images with left-right (or right-left) bias from an S3D image pair. Since all of the asymmetrically distorted stereoscopic image pairs in the LIVE 3D Image Quality Database Phase II [135] were created using a pristine

left-view image and a right-view image impaired by different types and degrees of distortions. Therefore, the two synthesized left- and right-view cyclopean images with bias may present differing perceptual characteristics, possibly resulting in biased performance of S3D IQA models utilizing cyclopean images, such as cyclopean MS-SSIM [81] and Chen [82].

To further investigate how much bias these perceptually distinct cyclopean images can introduce, we examined three different implementations of cyclopean MS-SSIM using the two possible left- and right-view cyclopean images. Implementation M_1 computed the MS-SSIM score between the undistorted and distorted left-view cyclopean images generated using the disparity maps computed using the left-view images, which are always pristine, as references. This is the same implementation adopted in [81]. Implementation M_2 computed the MS-SSIM score from right-view cyclopean images with disparity maps computed using the right-view images, which always contain some types of distortion, as references. In the last implementation M_3, we generated the final quality score by simply averaging the two above MS-SSIM scores. The last implementation could be used in practical scenarios because real-world stereoscopic image pairs can be impaired with either asymmetry. We tabulate the performance of these three different cyclopean MS-SSIM implementations in Table 7.6. It can be seen that the performance drops dramatically for implementation M_2. The seemingly more natural implementation (M_3) also suffers with reduced performance not significantly different than the 2D MS-SSIM index. Regarding the no-reference S3D IQA model proposed by Chen *et al.* [82],

since the features are extracted asymmetrically, their algorithm would require modification to be applied on arbitrary asymmetries. In the performance comparisons, the original implementations in [81] and [82] were used.

By synthesizing a more perceptually relevant and consistent convergent cyclopean image and by utilizing robust, effective bivariate and correlation natural image statistical models, S3D-BLINQ Index is able to achieve better than 0.9 correlation using both SROCC and LCC. It not only outperforms other state-of-the-art 2D and S3D IQA algorithms in terms of correlation monotonicity and accuracy, but also predicts the perceptual quality of stereoscopic image pairs with the lowest RMSE, as shown in Table 7.5.

Tables 7.3 – 7.5 also detail the performance of each quality assessment algorithm on different types of distorted stereopairs. We can see that almost all 2D and 3D algorithms are able to predict quality scores that correlate well with human opinions for stereoscopic image pairs affected by the WN distortion. However, several quality metrics perform poorly when predicting the perceptual quality of stereopairs impaired by JPEG, JP2K, and Blur distortions. These poor performances may be explained as a result of binocular facilitation [75, 139] whereby distortions co-located with high depth variations are more easily found by human subjects. This observed effect is not yet well understood or properly modeled.

To examine the capability of different 2D and S3D IQA models when dealing with unequally distorted stereopairs, which may be more common in practice, we list in Table 7.7 the performance of the same algorithms on both

Table 7.6: Comparison of Different Cyclopean MS-SSIM Implementations on the LIVE 3D Image Quality Database Phase II

Implementation	SROCC	LCC	RMSE
M_1	0.889	0.900	4.987
M_2	0.704	0.743	7.559
M_3	0.778	0.787	6.971

Table 7.7: Comparison (SROCC) of Different 2D and 3D Image Quality Assessment Algorithms on Symmetrically and Asymmetrically Distorted Stimuli in the LIVE 3D Image Quality Database Phase II

	Algorithm†	Symmetric	Asymmetric	Overall
2D	PSNR	0.776	0.587	0.665
	SSIM [63]	0.828	0.733	0.792
	MS-SSIM [64]	0.912	0.684	0.777
	BRISQUE [66]	0.849	0.667	0.770
3D	Benoit [72]	0.860	0.671	0.728
	You [76]	0.914	0.701	0.786
	Yasakethu [67]	0.656	0.496	0.501
	Cyclopean MS-SSIM [81]	0.923	0.842	0.889
	Sazzad [77]	0.420	0.517	0.543
	Chen [82]	0.918	0.834	0.880
	S3D-BLINQ Index	0.937	0.849	0.905

† *Italics* indicate no-reference IQA models. Others are full-reference IQA models.

symmetrically and asymmetrically distorted stereoscopic image pairs in the LIVE 3D Image Quality Database Phase II [135]. It can be seen that most of the examined quality models are capable of predicting scores that correlate well with human judgments on symmetrically distorted stereopairs. However, almost all of them perform poorly on asymmetrically distorted stereopairs, except for those utilizing cyclopean images. Among these, S3D-BLINQ In-

dex afforded the best performance on both symmetrically and asymmetrically distortions, resulting in the best overall correlation numbers as well.

7.4.2 Augmentation of the Bivariate and Correlation Models

The solid performance of S3D-BLINQ Index can be attributed to utilizing the perceptually relevant, convergent cyclopean image and robust and descriptive bivariate and correlation NSS models. Here we analyze the performance boost provided by the new bivariate and correlation NSS features underlying the S3D-BLINQ Index learning process. Specifically, we incorporated three different sets of features extracted from the convergent cyclopean image, following the same framework as described in Section 7.3. The three feature sets include the spatial-domain univariate NSS features, the wavelet-domain univariate NSS features, and the bivariate and correlation NSS features. We tabulate the performance of these three different feature sets, as well as the combination of all, i.e., S3D-BLINQ Index, in Table 7.8. It can be seen that using only the spatial-domain univariate NSS features is able to achieve 0.9 level of correlation performance on symmetrically distorted stereopairs, while the wavelet-domain univariate NSS features improve performance on asymmetric distortions. The bivariate and correlation NSS features further augment performance on asymmetrically distorted stereopairs, resulting in an overall 0.9 SROCC score when combining all feature sets.

Table 7.8: Comparison (SROCC) of the Proposed S3D-BLINQ Index framework using Different Feature Sets on Symmetrically and Asymmetrically Distorted Stimuli in the LIVE 3D Image Quality Database Phase II

Feature Set	Symmetric	Asymmetric	Overall
Spatial-Domain Univariate	0.911	0.808	0.873
Wavelet-Domain Univariate	0.852	0.815	0.854
Bivariate and Correlation	0.877	0.826	0.868
All	0.937	0.849	0.905

7.5 Summary

We generalized our new bivariate and correlation NSS models to capture the spatial oriented structure in bandpass distorted 2D and S3D images. These bivariate and correlation models are validated to be able to robustly and reliably quantify the statistical regularities embedded in spatially adjacent luminance pixels, and preliminarily yet systematically address one of the most important issues on NSS modeling of higher-order dependencies, which has not been well explored in literature. To demonstrate the efficacy of these new models, we deploy them to develop a new no-reference S3D IQA framework – the Stereoscopic/3D BLind Image Naturalness Quality (S3D-BLINQ) Index. Two important contributions are presented in this chapter. First, we defined a new and powerful set of quality-discriminative features by exploiting the new bivariate and correlation NSS models. Second, we proposed a convergent cyclopean image model to address bias encountered by earlier cyclopean image models.

Chapter 8

Conclusion and Future Work

In this dissertation we addressed the multidisciplinary problem of statistical modeling of natural image and depth/range information embedded in natural environments. A high-quality data set of accurately co-registered color images and depth/range maps, the LIVE Color+3D Database [7, 8], has been constructed and made publicly available in this regard. This database provides abundant and valuable resources ready for a diversity of research in visual psychophysics, image/video processing, computer vision, etc.

By utilizing this high-resolution, high-quality 3D image and depth/range database, we developed marginal and conditional priors relating natural luminance/chrominance and disparity, and demonstrate their efficacy with application to a chromatic Bayesian stereo algorithm. The statistical analysis we performed and the color-depth priors we derived yield insight into how 3D structures in the environment might be recovered from color image data.

While extensive research has been conducted on modeling marginal distributions of bandpass natural image responses with univariate functions, there exist higher-order dependencies between spatially neighboring bandpass responses that are not yet well understood or utilized in literature. Towards

filling this gap, we developed new bivariate and spatial oriented correlation models that capture statistical regularities between perceptually decomposed natural luminance and depth samples.

As a demonstration of the efficacy and effectiveness of our new bivariate and correlation natural scene statistical models, we exploited them to address two challenging, yet very important problems, depth estimation from monocular images and no-reference stereoscopic/3D (S3D) image quality assessment. With the aid of these reliable and robust statistical models, both the proposed Bayesian framework of depth estimation and the proposed S3D image quality index attain superior performance to state-of-the-art algorithms.

Understanding how human vision systems perceive binocular visual stimuli to reconstruct 3D natural environments is of extreme importance to a variety of science and engineering disciplines. In this dissertation, we approached this problem by exploring and modeling the statistics embedded in natural images and depth maps. We believe that our new bivariate and spatial oriented correlation models not only form a foundation of higher-order statistical exploration in natural scenes, but also have great potential to be used in a broad spectrum of 3D vision and image/video processing algorithms, such as de-noising, super-resolution, shape-from-X, quality assessment, etc. In particular, we expect a 'completely' blind S3D quality evaluator by appropriately incorporating these new statistical models that better quantify stereopair naturalness, and by incorporating perceptual measurements of visual discomfort into S3D quality of experience (QoE) models. Other future work

involves exploiting more psychophysical knowledge of human vision systems and introducing more complete higher-level statistical models that describe the interactions between natural image and depth/range information.

Bibliography

[1] L. Meesters, W. IJsselsteijn, and P. Seuntiens, "A survey of perceptual evaluations and requirements of three-dimensional TV," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 381–391, March 2004.

[2] H. Kalva, L. Christodoulou, L. Mayron, O. Marques, and B. Furht, "Challenges and opportunities in video coding for 3D TV," *IEEE International Conference on Multimedia and Expo*, pp. 1689–1692, Jul. 2006.

[3] F. L. Kooi and A. Toet, "Visual comfort of binocular and 3D displays," *Displays*, vol. 25, pp. 99–108, Aug. 2004.

[4] M. Lambooij, W. IJsselsteijn, M. Fortuin, and I. Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: A review," *Journal of Imaging Science and Technology*, vol. 53, no. 3, pp. 1–14, 2009.

[5] S. Reichelt, R. Haussler, G. Futterer, and N. Leister, "Depth cues in human visual perception and their realization in 3D displays," *SPIE Int. Conf. on Three-Dimensional Imaging, Visualization, and Display*, vol. 7690, no. 1, 2010.

[6] Z. Wang and A. C. Bovik, "Reduced- and no-reference image quality assessment: The natural scene statistic model approach," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29–40, Nov. 2011.

[7] C.-C. Su, L. K. Cormack, and A. C. Bovik, "LIVE Color+3D Database Release-1," http://live.ece.utexas.edu/research/3dnss/live_color_plus_3d.html.

[8] ——, "LIVE Color+3D Database Release-2," http://live.ece.utexas.edu/research/3dnss/live_color_plus_3d.html.

[9] RIEGL, "RIEGL VZ-400 3D Terrestrial Laser Scanner," http://rieglusa.com/products/terrestrial/vz-400/index.shtml.

[10] A. Saxena, M. Sun, and A. Y. Ng, "Make3D Laser+Image Dataset-1," http://make3d.cs.cornell.edu/data.html.

[11] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," *Neural Information Processing Systems*, 2005.

[12] A. Saxena, M. Sun, and A. Ng, "Make3D: Learning 3D scene structure from a single still images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, May 2009.

[13] B. A. Olshausen and D. J. Field, "Vision and the coding of natural images," *American Scientist*, vol. 88, pp. 238–245, 2000.

[14] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of the Optical Society of America A*, vol. 4, no. 12, pp. 2379–2394, 1987.

[15] D. L. Ruderman and W. Bialek, "Statistics of natural images: Scaling in the woods," *Physical Review Letters*, vol. 73, pp. 814–817, Aug. 1994.

[16] D. J. Field, "Wavelets, vision and the statistics of natural scenes," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 357, no. 1760, p. 2527, 1999.

[17] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *Journal of Mathematical Imaging and Vision*, vol. 18, no. 1, pp. 17–33, 2003.

[18] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 1193–1216, Mar. 2001.

[19] W. S. Geisler, "Visual perception and the statistical properties of natural scenes," *Annual Review of Psychology*, vol. 59, no. 1, pp. 167–192, Jan. 2008.

[20] A. van der Schaaf and J. H. van Hateren, "Modelling the power spectra of natural images: Statistics and information," *Vision Research*, vol. 36, no. 17, pp. 2759–2770, 1996.

[21] E. P. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," *Proc. SPIE*, vol. 3813, no. 1, pp. 188–195, 1999.

[22] S. Lyu and E. P. Simoncelli, "Statistical modeling of images with fields of gaussian scale mixtures," *Advances in Neural Information Processing Systems*, vol. 19, p. 945, 2007.

[23] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Transactions on Image Processing*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[24] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[25] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[26] M. A. Saad, A. C. Bovik, and C. Charrier, "Model-based blind image quality assessment using natural DCT statistics," *IEEE Transactions on Image Processing*, to appear.

[27] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[28] B. Olshausen and D. Field, "Natural image statistics and efficient coding," *Network: Computation in Nerual Systems*, vol. 7, no. 2, pp. 333–339, 1996.

[29] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[30] B. Potetz and T. S. Lee, "Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes," *Journal of the Optical Society of America A*, vol. 20, no. 7, pp. 1292–1303, Jul. 2003.

[31] ——, "Scaling laws in natural scenes and the inference of 3D shape," *Advances in Neural Information Processing Systems*, vol. 18, pp. 1089–1096, 2006.

[32] Y. Liu, L. K. Cormack, and A. C. Bovik, "Statistical modeling of 3-D natural scenes with application to bayesian stereopsis," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2515–2530, Sep. 2011.

[33] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Color and depth priors in natural images," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2259 – 2274, June 2013.

[34] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of*

Computer Vision, vol. 47, pp. 7–42, 2002.

[35] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference, 2002.

[36] S. Das and N. Ahuja, "Performance analysis of stereo, vergence, and focus as depth cues for active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 12, pp. 1213–1219, Dec. 1995.

[37] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 690–706, Aug. 1999.

[38] A. Maki, M. Watanabe, and C. Wiles, "Geotensity: Combining motion and lighting for 3D surface reconstruction," *International Journal of Computer Vision*, vol. 48, no. 2, pp. 75–90, Jul. 2002.

[39] T. Lindeberg and J. Garding, "Shape from texture from a multi-scale perspective," in *Proceedings of the 4th International Conference on Computer Vision*, May 1993, pp. 683–691.

[40] J. Malik and R. Rosenholtz, "Computing local surface orientation and shape from texture for curved surfaces," *International Journal of Computer Vision*, vol. 23, no. 2, pp. 149–168, June 1997.

[41] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 577–584, Jul. 2005.

[42] E. Delage, H. Lee, and A. Ng, "A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor images," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2418–2428, 2006.

[43] T. Nagai, T. Naruse, M. Ikehara, and A. Kurematsu, "HMM-based surface reconstruction from single images," in *Proceedings of the International Conference on Image Processing*, vol. 2, Sept. 2002, pp. 561–564.

[44] T. Hassner and R. Basri, "Example based 3D reconstruction from single 2D images," in *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*, June 2006, pp. 15–15.

[45] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 24, pp. 1226–1238, 2002.

[46] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1253 –1260, June 2010.

[47] K. Karsch, C. Liu, and S. Kang, "Depth extraction from video using non-parametric sampling," *European Conference on Computer Vision*, vol. 7576, pp. 775–788, Oct. 2012.

[48] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition, June 2011, pp. 305–312.

[49] A. Bovik, "Automatic prediction of perceptual image and video quality," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2008–2024, Sept. 2013.

[50] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Bivariate statistical modeling of color and range in natural scenes," in *Proceedings of SPIE, Human Vision and Electronic Imaging XIX*, vol. 9014, 2014.

[51] ——, "New bivariate and correlation statistical models of natural images," *IEEE Signal Processing Letters*, 2014, submitted.

[52] ——, "Oriented correlation models of distorted natural images with application to natural stereopair quality evaluation," *IEEE Transactions on Image Processing*, 2014, submitted.

[53] Motion Picture Association of America (MPAA), "Theatrical market statistics," http://www.mpaa.org/policy/industry, 2012.

[54] BBC News - Technology, "James Cameron: All entertainment 'inevitably 3D'," http://www.bbc.co.uk/news/entertainment-arts-23790877, Aug. 2013.

[55] J. Baltes, S. McCann, and J. Anderson, "Humanoid robots: Abarenbou and daodan," *RoboCup-Humanoid League Team Description*, 2006.

[56] A. M. William and D. L. Bailey, "Stereoscopic visualization of scientific and medical content," in *ACM SIGGRAPH 2006 Educators Program*, no. 26, 2006.

[57] C.-F. Westin, "Extracting brain connectivity from diffusion MRI [life sciences]," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 124–152, Nov. 2007.

[58] Y. A. W. De Kort and W. A. IJsselsteijn, "Reality check: the role of realism in stress reduction using media technology," *Cyberpsychology & Behavior*, vol. 9, no. 2, pp. 230–233, 2006.

[59] A. Puri, R. V. Kollarits, and B. G. Haskell, "Basics of stereoscopic video, new compression results with MPEG-2 and a proposal for MPEG-4," *Signal Processing: Image Communication*, vol. 10, no. 1, pp. 201–234, 1997.

[60] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 993–1008, 2003.

[61] International Telecommunication Union (ITU), "Methodology for the subjective assessment of the quality of television pictures," ITU-R Rec. BT.500-11, Sept. 2009.

[62] Z. Wang and A. C. Bovik, "Mean squared error: love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[63] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality

assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[64] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2003*, vol. 2, Nov. 2003, pp. 1398–1402.

[65] M. Saad, A. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[66] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[67] S. L. P. Yasakethu, C. T. E. R. Hewage, W. Fernando, and A. Kondoz, "Quality analysis for 3D video using 2D video quality models," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 1969–1976, Nov. 2008.

[68] P. Gorley and N. Holliman, "Stereoscopic image quality metrics and compression," in *Proceedings of SPIE, Stereoscopic Displays and Applications XIX*, vol. 6803, Jan. 2008.

[69] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, "Depth is encoded in the visual cortex by a specialized receptive field structure," *Nature*, vol.

352, no. 6331, pp. 156–159, 1991.

[70] D. J. Fleet, H. Wagner, and D. J. Heeger, "Neural encoding of binocular disparity: energy models, position shifts and phase shifts," *Vision Research*, vol. 36, no. 12, pp. 1839–1857, 1996.

[71] B. G. Cumming, "An unexpected specialization for horizontal disparity in primate primary visual cortex," *Nature*, vol. 418, no. 6898, pp. 633–636, 2002.

[72] A. Benoit, P. L. Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–13, 2009.

[73] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Proceedings of the 7th European Conference on Computer Vision*, vol. 2352, May 2002, pp. 82–96.

[74] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 41–54, Oct. 2006.

[75] M.-J. Chen, L. K. Cormack, and A. C. Bovik, "Distortion conspicuity on stereoscopically viewed 3D images may correlate to scene content and distortion type," *Journal of the Society for Information Display*, 2014, to appear.

[76] J. You, L. Xing, A. Perkis, and X. Wang, "Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis," in *Proceedings of the International Workshop on Video Processing and Quality Metrics*, 2010.

[77] Z. Sazzad, R. Akhter, J. Baltes, and Y. Horita, "Objective no-reference stereoscopic image quality prediction based on 2D image features and relative disparity," *Advances in Multimedia*, vol. 2012, no. 8, pp. 1–16, Jan. 2012.

[78] B. Julesz, *Foundations of Cyclopean Perception*. The University of Chicago Press, 1971.

[79] A. Maalouf and M.-C. Larabi, "CYCLOP: a stereo color image quality assessment metric," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2011, pp. 1161–1164.

[80] S. J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *Proceedings of SPIE, Human Vision, Visual Processing, and Digital Display III*, vol. 1666, Feb. 1992, pp. 2–15.

[81] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1143–1155, Oct. 2013.

[82] M.-J. Chen, L. K. Cormack, and A. C. Bovik, "No-reference quality assessment of natural stereopairs," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3379–3391, Sept. 2013.

[83] RIEGL, "RIEGL RiSCAN PRO Software for 3D Terrestrial Laser Scanner," http://rieglusa.com/products/terrestrial/vz-400/software.shtml.

[84] Intel Corporation, "OpenCV: Camera Calibration and 3D Reconstruction," http://opencv.willowgarage.com/documentation/camera_calibration_and_3d_reconstruction.html.

[85] I. Fine, D. I. A. MacLeod, and G. M. Boynton, "Surface segmentation based on the luminance and color statistics of natural scenes," *Journal of the Optical Society of America A*, vol. 20, no. 7, pp. 1283–1291, 2003.

[86] J. R. Jordan, W. S. Geisler, and A. C. Bovik, "Color as a source of information in the stereo correspondence process," *Vision Research*, vol. 30, no. 12, pp. 1955–1970, 1990.

[87] J. R. Jordan and A. C. Bovik, "Using chromatic information in edge-based stereo correspondence," *Computer Vision, Graphics, and Image Processing: Image Understanding*, vol. 54, no. 1, pp. 98–118, Jul. 1991.

[88] ——, "Using chromatic information in dense stereo correspondence," *Pattern Recognition*, vol. 25, no. 4, pp. 367–383, Apr. 1992.

[89] U. Rajashekar, Z. Wang, and E. P. Simoncelli, "Perceptual quality assessment of color images using adaptive signal representation," *SPIE*

Int. Conf. on Human Vision and Electronic Imaging, vol. 7527, no. 1, Jan. 2010.

[90] B. G. Cumming and G. C. DeAngelis, "The physiology of stereopsis," *Neuroscience*, vol. 24, no. 1, p. 203, 2001.

[91] I. P. Howard and B. J. Rogers, *Binocular Vision and Stereopsis*. New York, USA: Oxford University Press, 1995.

[92] C. Blakemore, "The range and scope of binocular depth discrimination in man," *Journal of Physiology*, vol. 211, no. 3, pp. 599–622, Dec. 1970.

[93] Y. Liu, A. C. Bovik, and L. K. Cormack, "Disparity statistics in natural scenes," *Journal of Vision*, vol. 8, no. 11, pp. 1–14, Aug. 2008.

[94] S. Prince, A. D. Pointon, B. G. Cumming, and A. J. Parker, "Quantitative analysis of the responses of V1 neurons to horizontal disparity in dynamic random-dot stereograms," *Journal of Neurophysiology*, vol. 87, no. 1, pp. 191–208, 2002.

[95] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of Physiology*, vol. 195, pp. 215–243, Mar. 1968.

[96] S. Marcelja, "Mathematical description of the responses of simple cortical cells," *Journal of the Optical Society of America A*, vol. 70, pp. 1297–1300, 1980.

[97] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 55–73, Jan. 1990.

[98] D. H. Hubel, "The visual cortex of the brain," *Scentific American*, vol. 209, no. 5, pp. 54–63, Nov. 1963.

[99] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America A*, vol. 2, no. 7, pp. 1160–1169, Jul. 1985.

[100] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[101] R. L. D. Valois, D. G. Albrecht, and L. G. Thorell, "Spatial frequency selectivity of cells in macaque visual cortex," *Vision Research*, vol. 22, no. 5, pp. 545–559, 1982.

[102] R. L. D. Valois, E. W. Yund, and N. Hepler, "The orientation and direction selectivity of cells in macaque visual cortex," *Vision Research*, vol. 22, no. 5, pp. 531–544, 1982.

[103] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *Journal of Neurophysiology*, vol. 58, no. 6, pp. 1233–1258, 1987.

[104] G. J. Burton and I. R. Moorhead, "Color and spatial structure in natural scenes," *Applied Optics*, vol. 26, pp. 157–170, 1987.

[105] W. Bialek, D. L. Ruderman, and A. Zee, "Optimal sampling of natural images: A design principle for the visual system?" *Advances in Neural Information Processing Systems*, vol. 3, pp. 363–369, 1991.

[106] J. H. van Hateren, "Real and optimal neural images in early vision," *Nature*, vol. 360, pp. 68–70, Nov. 1992.

[107] ——, "Spatiotemporal contrast sensitivity of early vision," *Vision Research*, vol. 33, no. 2, pp. 257–267, 1993.

[108] P. N. Belhumeur, "A Bayesian approach to binocular steropsis," *International Journal of Computer Vision*, vol. 19, pp. 237–260, 1996.

[109] S. T. Barnard, "A stochastic approach to stereo vision," *Proc. of the 5th Nat. Conf. on Artificial Intelligence, AAAI*, vol. 1, pp. 676–680, Aug. 1986.

[110] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, pp. 43–77, 1994.

[111] R. Szeliski and R. Zabih, "An experimental comparison of stereo algorithms," *Vision Algorithms: Theory and Practice*, vol. 1883, pp. 1–19, 2000.

[112] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," *IEEE International Conference on Computer Vision*, vol. 2, pp. 508–515, Jul. 2001.

[113] A. Mittal, R. Soundararajan, and A. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, March 2013.

[114] H. R. Sheikh, Z. Wang, L. K. Cormack, and A. C. Bovik, "LIVE Image Quality Assessment Database," http://live.ece.utexas.edu/research/quality/subjective.htm.

[115] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," *IEEE International Conference on Image Processing*, vol. 3, pp. 444–447, Oct. 1995.

[116] B. A. Olshausen and D. J. Field, "How close are we to understanding V1?" *Neural Computation*, vol. 17, no. 8, pp. 1665–1699, Aug. 2005.

[117] M. J. Wainwright, O. Schwartz, and E. P. Simoncelli, "Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons," *Probabilistic Models of the Brain: Perception and Neural Function*, pp. 203–222, Feb. 2002.

[118] S. Lyu, "Dependency reduction with divisive normalization: justification and effectiveness," *Neural Computation*, vol. 23, pp. 2942–2973, 2011.

[119] L. Bombrun, F. Pascal, J.-Y. Tourneret, and Y. Berthoumieu, "Performance of the maximum likelihood estimators for the parameters of multivariate generalized gaussian distributions," *IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 3525–3528, 2012.

[120] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the Society for Industrial & Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[121] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Natural3D software release," http://live.ece.utexas.edu/research/3dnss/index.html.

[122] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature Neuroscience*, vol. 4, pp. 819–825, Aug. 2001.

[123] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 52–56, Feb. 1995.

[124] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, Sept. 1999, pp. 1150–1157.

[125] J.-Y. Jou and A. C. Bovik, "Improved initial approximation and intensity-guided discontinuity detection in visible-surface reconstruction," *Com-*

puter Vision, Graphics, and Image Processing, vol. 47, no. 3, pp. 292–326, Sept. 1989.

[126] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[127] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[128] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[129] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, May 2001.

[130] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Experimental results of monocular depth estimation algorithms," http://live.ece.utexas.edu/research/3dnss/index.html.

[131] F. Pascal, L. Bombrun, J.-Y. Tourneret, and Y. Berthoumieu, "Parameter estimation for multivariate generalized Gaussian distributions,"

IEEE Transactions on Signal Processing, vol. 61, no. 23, pp. 5960–5971, Dec. 2013.

[132] W. J. M. Levelt, *On Binocular Rivalry*. Mouton, The Hague, 1968.

[133] D. L. Ruderman, "The statistics of natural images," *Network: Computation in Neural Systems*, vol. 5, no. 4, pp. 517–548, 1994.

[134] N.-E. Lasmar, Y. Stitou, and Y. Berthoumieu, "Multiscale skewed heavy tailed model for texture analysis," in *Proceedings of the IEEE International Conference on Image Processing*, Nov. 2009, pp. 2281–2284.

[135] A. Moorthy, C.-C. Su, M.-J. Chen, A. Mittal, L. K. Cormack, and A. C. Bovik, "LIVE 3D Image Quality Database Phase I and Phase II," http://live.ece.utexas.edu/research/quality/live_3dimage.html.

[136] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Validation of the exponentiated sine model," http://live.ece.utexas.edu/research/3dnss/modeling.html.

[137] M. Narwaria and W. Lin, "Objective image quality assessment based on support vector regression," *IEEE Transactions on Neural Networks*, vol. 21, no. 3, pp. 515–519, March 2010.

[138] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[139] M.-J. Chen, A. C. Bovik, and L. K. Cormack, "Study on distortion conspicuity in stereoscopically viewed 3D images," in *Proceedings of the IEEE IVMSP Workshop*, June 2011, pp. 24–29.

Vita

Che-Chun Su received the Bachelor of Science (B.S.) degree in electrical engineering, and the Master of Science (M.S.) degree in communication engineering from National Taiwan University, Taipei, Taiwan (R.O.C.), in 2004 and 2006, respectively. He joined the Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin, in 2009. Under the supervision of Dr. Alan C. Bovik and the co-supervision of Dr. Lawrence K. Cormack, he received the Doctor of Philosophy (Ph.D.) degree in electrical and computer engineering from The University of Texas at Austin, Austin, Texas, USA, in 2014. His research interests include image/video processing and quality assessment, human vision perception, and 3D natural scene statistics.

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.