2012

NOVEL DENSE STEREO ALGORITHMS FOR HIGH-QUALITY DEPTH ESTIMATION FROM IMAGES

Liang Wang
University of Kentucky, liangwan@microsoft.com

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained and attached hereto needed written permission statements(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine).

I hereby grant to The University of Kentucky and its agents the non-exclusive license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless a preapproved embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's dissertation including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Liang Wang, Student

Dr. Ruigang Yang, Major Professor

Dr. Raphael A. Finkel, Director of Graduate Studies

NOVEL DENSE STEREO ALGORITHMS FOR HIGH-QUALITY DEPTH ESTIMATION FROM IMAGES

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy in
the Department of Computer Science
at the University of Kentucky

By
Liang Wang
Lexington, Kentucky

Director: Dr. Ruigang Yang, Associate Professor of Computer Science
Lexington, Kentucky 2012

ABSTRACT OF DISSERTATION

NOVEL DENSE STEREO ALGORITHMS FOR HIGH-QUALITY DEPTH ESTIMATION FROM IMAGES

This dissertation addresses the problem of inferring scene depth information from a collection of calibrated images taken from different viewpoints via stereo matching. Although it has been heavily investigated for decades, depth from stereo remains a long-standing challenge and popular research topic for several reasons. First of all, in order to be of practical use for many real-time applications such as autonomous driving, accurate depth estimation in real-time is of great importance and one of the core challenges in stereo. Second, for applications such as 3D reconstruction and view synthesis, high-quality depth estimation is crucial to achieve photo realistic results. However, due to the matching ambiguities, accurate dense depth estimates are difficult to achieve. Last but not least, most stereo algorithms rely on identification of corresponding points among images and only work effectively when scenes are Lambertian. For non-Lambertian surfaces, the "brightness constancy" assumption is no longer valid. This dissertation contributes three novel stereo algorithms that are motivated by the specific requirements and limitations imposed by different applications.

In addressing high speed depth estimation from images, we present a stereo algorithm that achieves high quality results while maintaining real-time performance. We introduce an adaptive aggregation step in a dynamic-programming framework. Matching costs are aggregated in the vertical direction using a computationally expensive weighting scheme based on color and distance proximity. We utilize the vector processing capability and parallelism in commodity graphics hardware to speed up this process over two orders of magnitude.

In addressing high accuracy depth estimation, we present a stereo model that makes use of constraints from points with known depths - the Ground Control Points (GCPs) as referred to in stereo literature. Our formulation explicitly models the influences of GCPs in a Markov Random Field. A novel regularization prior is naturally integrated into a global inference framework in a principled way using the Bayes rule. Our probabilistic framework allows GCPs to be obtained from various modalities and provides a natural way to integrate information from various sensors.

In addressing non-Lambertian reflectance, we introduce a new invariant for stereo correspondence which allows completely arbitrary scene reflectance (bidirectional reflectance distribution functions - BRDFs). This invariant can be used to formulate a rank constraint on stereo matching when the scene is observed by several lighting configurations in which only the lighting intensity varies.

KEYWORDS: Stereo Matching, Bilateral Filtering, Dynamic Programming, Global Optimization, Light Transport Constancy

Author's signature: Liang Wang

Date: March 21, 2012

NOVEL DENSE STEREO ALGORITHMS FOR HIGH-QUALITY DEPTH ESTIMATION FROM IMAGES

By
Liang Wang

Director of Dissertation: Ruigang Yang

Director of Graduate Studies: Raphael A Finkel

Date: March 21, 2012

To my beloved wife, Jin.

ACKNOWLEDGMENTS

My thanks go first to my advisor, Dr. Ruigang Yang. I sincerely appreciate him for his insightful guidance, creative inspiration, and endless encouragement throughout my graduate studies. I have learned a lot from him about doing research, presenting results, and interacting with people. In retrospect, I am very grateful to Dr. Yang for giving me the freedom to pursue various research topics and raising my interest in the field of computer vision. His support made substantial impacts in my dissertation work. I consider myself truly fortunate to have him being my advisor.

This dissertation is built upon part of the research projects I have done with several collaborators. Here I would like to thank my co-authors, who all have provided valuable efforts to different parts of the dissertation. In particular, I wish to express my deep gratitude to Minglun Gong, Hailin Jin, and David Nister. It is extremely enlightening to discuss with them and our collaboration on a variety of projects is highly enjoyable.

Next I would like to express my gratitude to my committees, including Professors Brent Seales, Fuhua Cheng, and Laurence Hassebrook. I appreciate them for reading my dissertation and offering thoughtful comments. I also owe sincere thanks to the staff of the Center for Visualization and Virtual Environments and the Director of Graduate Studies of the Computer Science Department, Professor Raphael A Finkel who provided administrative supports.

Many thanks to my colleagues in the GRAVITY lab: Xinyu Huang, Miao Liao, Xianwang Wang, Jizhou Gao, Qing Zhang, Chenxi Zhang, Mao Ye, Bo Fu, Jiejie Zhu, Qingxiong Yang, and Yongwook Song. Besides from the discussion we had and the invaluable support they provided, I appreciate all the wonderful moments we spent together. They made these last several years truly memorable.

Of course, I could never finish this dissertation without the support from my family. I would like to take this opportunity to mention my parents Tieliang and Jian, who have had the most significant impact on my life, and my new born daughter, Judy, who has been and continues to be my source of joys and distractions.

Finally, but most importantly, thanks to my wife Jin for her unconditional love. I cannot thank her enough for her care, support, and patience in my pursuit of the Ph.D. degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

Recovering the 3D shape of a scene from one or multiple images has long been a topic of research in computer vision and photogrammetry. This problem is known as shape-from-X, where X can be shading, motion, texture, silhouettes, and focus/defocus etc. Solving this problem opens many applications, ranging from CAD-based industrial manufacturing, scene understanding to 3D modeling. The methodology addressed in this dissertation belongs to a discipline that is called *Stereo Matching*. With the assumption of scene rigidity[1] and known camera geometry, a stereo matching algorithm aims at estimating three-dimensional scene structure from a collection of images taken from distinct viewpoints.

Stereo algorithms rely on the ability to establish correspondences of points of the scene across different images. Two image points match if they result from the projection of the same 3D point in the scene. Correspondences are usually obtained by putting assumptions on the reflectance properties of the scene. The most common assumption is that the scene is Lambertian, without specularities, reflective surfaces, or transparency. Under this Lambertian or *brightness constancy* assumption, locations in the scene will appear equally bright from any viewing direction, and therefore correspondences can be established via feature- or area-based matching. Equally important is the knowledge about the camera positions and orientations in 3D. The known camera configuration provides a powerful *epipolar geometry* constraint for matching.

[1] scene rigidity means that either the images have to be taken at the same time instant from multiple cameras or the objects in the scene are stationary.

Once a correspondence is established, one can apply the well-studied theory of multi-view geometry [7] to reconstruct a point's location in 3D. The desired output of a stereo algorithm is a dense disparity map[2], specifying the relative displacement of matching points between images. By dense, we mean a disparity estimate is assigned for every pixel of a *reference frame* chosen from the multiple input images.

Depth from stereo has traditionally been, and continues to be one of the most actively researched topics in in the computer vision community. Although multiple methods exist for acquiring 3D information, stereo is becoming the technology of choice for range sensing by a wide variety of applications because by using passive cameras stereo systems are economic in size, weight and cost. Additional advantages of using stereo to infer scene depth include that the setup can be adapted to work in both indoor and outdoor environments and the process can be easily automated. Stereo vision is therefore highly important in various fields. Traditional applications of stereo include industrial inspection, people tracking, aerial surveys, cartography, mobile robotics navigation, etc. More recently, the advances in stereo algorithms allow stereo to be applied to many new areas such as detailed 3D urban modeling [8], scene parsing and segmentation [9], teleconferencing [10] and image-based rendering [11].

1.1 Motivation and Contributions

The research presented in this dissertation aims to make depth from stereo more accurate and feasible for demanding applications that require precise, reliable, and dense

[2]Disparity refers to the difference in image location between corresponding pixels in the two images, which is projectively related to the depth of the feature in the scene.

depth estimates. Towards this goal, we address three key challenges for estimating dense scene structure using stereo matching and contribute several novel algorithms that are motivated by the specific requirements and limitations imposed by different types of application.

First, we address the difficulty of acquiring high-quality depth estimates in real-time. As a result of the public available Middlebury benchmark [1], recent stereo research has significantly advanced the state-of-the-art in terms of depth quality. However, in terms of speed, top algorithms typically take several seconds or minutes to compute a disparity map [12, 13]. Excessively long computation time needed to match stereo images is one of the obstacles on the way to the practical application of stereo techniques. There are demanding applications, such as automotive driver assistance and augmented reality, in which reliable dense depth estimates at video frame rate is crucial. For real-time stereo, the options are rather limited that in general only correlation based [14] and scanline optimizations based approaches [15] are feasible. Most local approaches, although being fast, are quite fragile and prone to have difficulties within textureless regions or near occlusion boundaries. Scanline optimization utilizes dynamic programming (DP) to produce better quality results. However, as each scanline is optimized independently, erroneous horizontal strokes, i.e. the "streaking" artifacts, often arise in the disparity maps. In this dissertation, we present a novel algorithm that achieves high quality depth estimation while maintaining real-time processing power. The proposed algorithm is simple yet effective. The key idea is to employ an adaptive cost-volume filtering stage in a DP framework. The per-pixel matching costs are aggregated via a separable implementation

of the bilateral filtering technique. The separable approximation leads to a significant reduction in computational complexity compared to the traditional 2D filter but offers comparable edge-preserving smoothing capability. The cost aggregation step alleviates the depth inconsistency between image scanlines, which is the typical problem for conventional DP-based stereo approaches. For computational efficiency, we utilize the vector processing capability and parallelism in commodity graphics hardware to speed up the aggregation process over two orders of magnitude. Our current implementation can achieve over 50 million disparity evaluations per second (MDE/s)[3].

The second challenge that this dissertation addresses is how to resolve the matching ambiguities for applications that require high-accuracy depth estimation. The stereo correspondence problem is inherently under-constrained. A practical stereo algorithm has to deal with the problem of matching ambiguity results from sensor noise in image formation, homogeneous texture regions, delineation of object boundaries, and unmatched pixels due to occlusions. Prior constraints are typically needed to regularize the ill-posed correspondence problem. Two most popular priors are the spatial smoothness [16] and the segment-based priors [17]. The former encourages neighboring pixels to have similar depth values based on the assumption that the scene is locally smooth. The segment-based stereo model encodes the assumption that homogeneously textured image regions correspond to planar surfaces in 3D. Nowadays, nearly all competitive stereo methods use these constraints to decrease

[3]The number of disparity evaluations per seconds (MDE/s) corresponds to the product of the number of pixels times the disparity range times the obtained frame-rate and therefore captures the performance of a stereo algorithm in a single number.

the ambiguities in the matching process. Nevertheless, it is well-known that segmentation is a double-edged sword. Despite the fact that segment-based methods usually improve results in large textureless regions, they inevitably introduce errors in textured areas and do not handle well the situation that the scene contains non-planar surfaces. Toward this end, we present a novel global stereo formulation that makes use of constraints from points with known depths, i.e., the *Ground Control Points* (GCPs) as referred to in stereo literature [4]. Our formulation explicitly models the influences of GCPs in a Markov Random Field (MRF). A GCPs-based regularization prior is naturally integrated into a global optimization framework in a principled way using the Bayes rule. Quantitative evaluations demonstrate the effectiveness of our stereo model for improving reconstruction accuracy. The probabilistic inference framework makes no specific restriction on the GCP's acquisition strategy. This nice property allows the GCPs to be obtained from different sources, e.g., reliably matched pixels, low resolution range data, user interaction or any combination of these modalities. Therefore our method provides a natural way to integrate the information from multiple sensors. In this dissertation, we demonstrate that it can be utilized to fuse measurements from sparse laser scanning and high resolution image data for urban 3D reconstruction.

The third contribution of this dissertation is a new matching invariant for reconstructing a large class of non-Lambertian surfaces. As mentioned above, nearly all existing stereo methods rely on the assumption that objects in the scene reflect light equally in all directions (Lambertian reflectance) and use brightness constancy as a matching invariant to establish correspondences. Unfortunately, this assumption is

violated for objects with non-Lambertian (specular reflectance) surfaces because the appearances of such objects in images can change drastically from one view to another, leading to incorrect matching. In the past, a considerable amount of methods for overcoming this limitation have been developed, but all require some combination of calibrated light sources, calibration objects in the scene, or smoothness assumptions on the surface reflectance. In this dissertation, we present a new constraint for stereo, namely, *light transport constancy* (LTC), which allows completely arbitrary scene reflectance (BRDFs). Different from the brightness constancy, LTC is based on the observation that the percentage of light reflected by a particular surface patch remains constant for a given viewing direction. We show that this invariant can be used to formulate a rank constraint on multi-view stereo when the scene is observed in several lighting configurations. In addition, we demonstrate that this multi-view constraint can be used with as few as two cameras and two lighting configurations. Compared to previous solutions, LTC does not require precisely configured/calibrated light source, nor calibration objects in the scene. Importantly, this constraint can be used to provide BRDF invariance to any existing stereo methods whenever appropriate lighting variations are available.

1.2 Guideline for Reading

This dissertation is divided into two parts. The next two chapters (Chapters 2 and 3) contain background materials and related work and can be used as a reference on stereo matching. In particular, Chapter 2 starts with an introduction of existing image-based depth estimation approaches. Then we provide preliminaries for stereo

and revisit the taxonomy of stereo algorithms proposed by Scharstein and Szeliski [18] to review a set of key algorithmic building blocks of stereo algorithms. This chapter ends with a description of the quality metrics we use in this dissertation for quantitatively evaluating the performance of stereo correspondence algorithms. In Chapter 3, we review existing stereo methods that are most relevant to the stereo algorithms proposed in this dissertation.

Chapters 4, 5, and 6 contain the core material of this dissertation. Chapter 4 addresses the challenge of inferring dense scene geometry in real-time. Since our algorithm is inspired by the idea of edge-preserving filters, we first review the bilateral filtering technique and its application in stereo correspondence. We then introduce a fast separable approximation of the bilateral filtering based cost aggregation approach that significantly reduces the computational complexity. In addition, we show that our aggregation scheme can be incorporated into a DP scanline optimization framework for improved reconstruction accuracy. To further improve speed performance, we utilize the graphics hardware to perform cost aggregation in massive parallelism and report implementation details. The leverage of GPUs allows depth estimation in video frame rate. This chapter ends with with experimental results from various data sets, including static benchmark images and live stereo videos with dynamic scenes.

In Chapter 5 we switch our attention from real-time depth estimation to offline but high accuracy stereo algorithms. The main contribution of Chapter 5 is a new regularization prior for stereo correspondence. We start with the definition our basic stereo matching model in Section 5.1. We explain in detail the GCPs-based regularization prior in Section 5.2 and propose an adaptive propagation algorithm

for modeling the prior likelihood from sparse GCPs. The experimental section covers two interesting scenarios. First we assume that there is no additional sensors other than cameras available to provide GCPs and show that GCPs can be computed from images themselves via stable matching. Furthermore we apply our stereo model to outdoor 3D reconstruction. In this scenario, low resolution laser range scans serve as GCPs and are fused with high resolution image data via our stereo model.

Chapter 6 discusses stereo for non-Lambertian scenes. We start with a local analysis on the scene radiance and arrive at a ratio constraint for BRDF invariant stereo. This simple constraint can be adopted to design a practical stereo system using two cameras and a single uncalibrated light source. We later extend our formulation and derive a series of linear equations that can accommodate an arbitrary number of cameras and light sources. Based on these equations we introduce a general rank constraint on multi-view stereo matching regardless of the surface BRDF complexity. In the experiment section we validate and evaluate our method using an extensive set of stereo images captured under varying illumination conditions.

Finally, in Chapter 7 we conclude the dissertation with discussions on possible directions for future developments. In this dissertation, Chapter 4 extends the joint work with Miao Liao, Minglun Gong, Ruigang Yang, and David Nister, first presented in 3DPVT 2006[4]. Chapter 5 is an extension of a joint work with Ruigang Yang, first presented in IEEE CVPR 2011[5] . Chapter 6 describes a joint work with Ruigang Yang and James Davis, first presented in IEEE PAMI 2007[6].

[4]Third International Symposium on 3D Data Processing, Visualization and Transmission.
[5]IEEE Computer Vision and Pattern Recognition Conference.
[6]IEEE Transactions on Pattern Analysis and Machine Intelligence.

Chapter 2 Background

The purpose of this chapter is to provide background material about stereo matching. We first discuss methods relating to the problem of depth estimation from digital images. We then give a brief review of stereo and outline a framework for stereo from which most of the stereo algorithms are constructed. This chapter ends with a description of the quality metrics we use in this dissertation for evaluating the performance of stereo algorithms. Much of the discussion in this chapter is at a general level and may safely be skipped for readers who are familiar with stereo.

2.1 Depth Estimation from Images

Over the last century, a vast number of depth acquisition methods have been developed. These methods vary significantly in terms of their specialties, capabilities and hardware requirements. In this section, we briefly review existing methods that attempt to infer 3D structure from photographs taken by one or multiple cameras. These methods can be further divided into passive and active methods, depending on whether the images are captured under natural or controlled lighting environments.

2.1.1 Passive Methods

Passive methods recovery 3D shape from images taken under natural lighting conditions and do not interfere with the reconstructed object. In other words, no other device besides camera(s) is required. The majority of these methods are based on the

principle of multi-view triangulation. Based on this principle, a point's 3D position can be reconstructed by intersecting the lines of sight of the corresponding pixels in multiple images. Two fundamental 3D reconstruction approaches in computer vision, structure from motion and multi-view stereo, belong to this class.

Structure from Motion. Given a set of image features together with their correspondences across views, structure from motion (SFM) aims at recovering both camera motion and the 3D positions of these feature points. Hartley and Zisserman [7] provide a comprehensive overview of existing methods and explain how to and implement the SFM algorithms. Nister in [19] describes a complete SFM system and applies it to automatic 3D reconstruction with a hand-held video camera. Recent work in SFM [20, 21] addresses the problem of handling non-rigid scenes, which gives a high degree of flexibility and allows an extended range of applications to be fulfilled. A typical difficulty in SFM is that pixel correspondences can only be established stably for salient image features [22]. Therefore, SFM often produce sparse 3D estimates only.

Passive Stereo. Passive stereo assumes the camera configurations are known and seeks to compute pixel correspondence for dense 3D reconstruction. Several excellent surveys of recent advances in this field can be found in [18, 23, 24]. The limitation of stereo comes from the fact that they rely on image-to-image correspondence. Correspondence-based stereo methods perform well when the scene is Lambertian and contains rich texture in the albedos. But they usually fail for scenes that are non-Lambertian or Lambertian with little texture. Modern stereo methods resolve matching ambiguities by assuming smoothness or planar prior model [17] for the

underlying 3D shape. Nevertheless, obtaining precise and robust depth estimates remains a very active and challenging area of research. Since stereo is the main focus of this dissertation, Sections 2.2 and 2.3 will provide more detailed background materials.

In addition to multi-view triangulation based methods, there are passive methods that attempt to infer depth from a single image. Single view depth estimation is difficult without prior constraints because depth typically remains ambiguous given only image features. There are semi-automatic methods leveraged by user interactions [25–27]. This class of methods reconstruct a 3D surface that satisfies a sparse set of user-specified constraints, e.g. surface normals, silhouettes and depth. An example of automatic methods is shape from texture (SFT), which reconstructs depth via monocular cues such as texture variations and gradients [28–30]. The limitation of SFT is that it generally assumes uniform texture distribution and would perform poorly on unconstrained or highly textured scenes. Recently, data driven and machine learning based single view reconstruction has been successfully demonstrated for outdoor scenes [31–33]. The performance of these methods depends largely on the training data. For instance, they would fail on unseen objects or environments that do not belong to any of the training images.

2.1.2 Active Methods

Active methods reconstruct 3D shapes by emitting radiance towards the object and then measure its reflected part. A large body of literature in this field uses lights as energy waves. These methods differ in the way they control the lights and the way

they reconstruct shapes from the returned signals.

Shape-from-Shading. Shape-from-Shading (SFS) deals with the recovery of shape from a gradual variation of shading in the image [34–36]. Given one gray level image and known light direction, the surface shape at each pixel can be recovered by studying the image formation process [37]. Since developed, most work in SFS makes simplified assumption, that is the reconstructed surface is Lambertian and with constant or known albedo [38–41]. Several extensions have been proposed to address this limitation [42–45]. Nevertheless, satisfactory results are still hard to achieve on real images with arbitrary surface BRDFs [46].

Photometric Stereo. Photometric stereo recovers the shape and albedo of an object using multiple images among which camera position is fixed, and only the light directions vary [47]. It then computes the surface orientation for each pixel based on its shading variation under different lighting conditions. The surface shape can be generated by integrating over the estimated surface normal. Although multiple lighting variations lead to accurate results, traditional methods follow the same image formation assumption as made in SFS [47]. As a result, most work cannot well handle non-Lambertian surfaces. There exist methods to address this limitation. Some approaches require extra constraints on the number and positions of light sources and allow only a class of diffuse non-Lambertian surfaces to be handled [48, 49]. Some require a calibration object with BRDF similar to the unknown scene as a prior knowledge [50].

Active Stereo. Active stereo addresses the difficulty of matching low texture areas in passive stereo by projecting a high contrast pattern onto the scene [51]. This idea in-

troduces synthetic textures over the surface without physically touching it. Recently, Spacetime stereo [52, 53] formulates stereo matching in the presence of illumination variation and achieves excellent results. There are also active stereo methods proposed to handle non-Lambertian objects. For example Helmholtz stereopsis allows matching of arbitrary BRDFs and uses reflectance function reciprocity as an invariant [54, 55]. By collocating point light sources with each camera, it is possible to record reciprocal pairs using two different lighting conditions. However, this method assumes the light sources to be collocated with the optical center of each camera and requires an extra calibration procedure.

Active Shape-from-Focus/Defocus. Traditional Shape-from-Focus /Defocus algorithms collect images at multiple lens settings and define metrics that evaluate sharpness or the amount of blurring over a small spatial area surrounding the pixel [56–59]. Most of these methods follow the equalfocal assumption, i.e., the surface depth is constant within that area, therefore suffer from poor performance near depth discontinuities. To address this issue, Hasinoff et al. [60] and Zhang et al. [61] have presented new methods which allow per-pixel focus/defocus analysis to be applied and can achieve sharper, more accurate geometric details. Since nearly all Shape-from-Focus/Defocus methods have difficulty dealing with textureless regions due to focus ambiguity, illumination patterns are usually projected to provide synthetic scene textures at the expenses of light source calibration [61, 62].

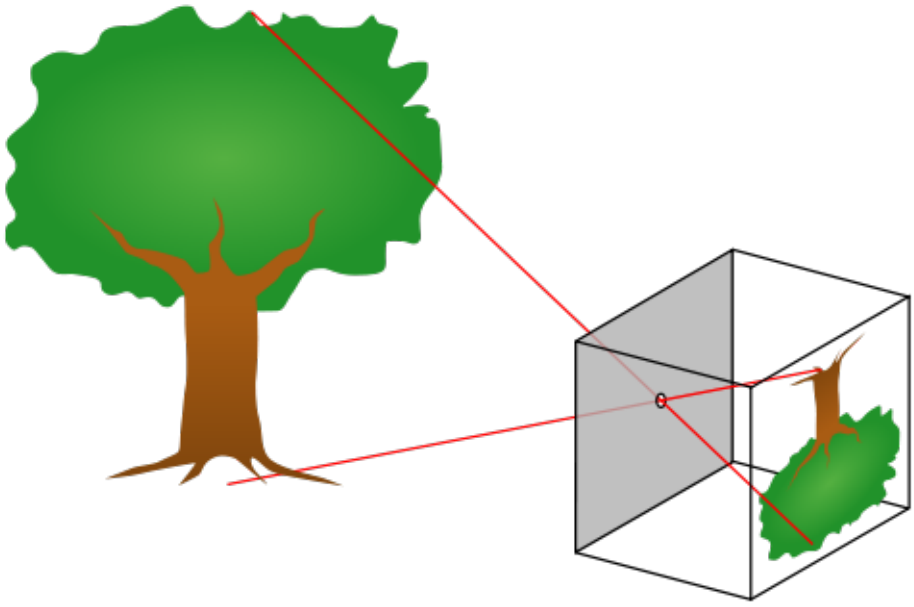


Figure 2.1: The pinhole camera model. An image of a 3D object is formed by perspective projection: each ray of light passes through a common center of projection and intersects the image plane.

2.2 Preliminaries

For readers not familiar with computer vision, we now provide a brief overview of stereo. For interested readers who wish to learn more about the field, a number of books on computer vision are available for a more detailed discussion [7, 11, 19, 63].

2.2.1 Image Formation

Throughout this dissertation, we use perspective projection as our geometric model of image formation. In detail, an image is formed by projecting each 3D scene point along a straight line through the center of projection onto a 2D image plane. This model is commonly referred to as the pinhole camera model (see Figure 2.1): light from a scene passes through a pinhole and projects an inverted image of the scene

on the opposite side of an opaque box. The pinhole camera model describes the mathematical relationship between the coordinates of a 3D point and its projection onto the 2D image plane of a pinhole camera, where the camera aperture is a point and there are no lenses used to focus light. In the computer vision community, the pinhole model is a widely adopted camera model because it resembles closely the image formation process of a real camera. The principal difference is that real cameras have a lens instead of a point. Therefore, radial distortions introduced by the lens are not accounted for by the simple pinhole model. Fortunately, lens distortion can be corrected by an image transformation process as described in [7].It also does not take into account blurring of unfocused objects caused by lenses and finite sized apertures.

When working with perspective projection for computer vision it is customary and convenient to use homogeneous coordinates. Mathematically, each point in homogeneous coordinates is extended by a dummy coordinate $w \neq 0$ that maps the point to a line through the origin in a space whose dimension is one higher than that of the original space [64]. For example, a 2D image point (x, y) and a 3D scene point (X, Y, Z) are represented by the set of vectors $[wx\ wy\ w]^T$ and $[wX\ wY\ wZ\ w]^T$, respectively. Homogeneous coordinates allow us to express perspective projection of a 3D scene point onto a 2D image plane using the following linear equation: $[u\ v\ w]^T = \mathbb{P}[X\ Y\ Z\ 1]^T$. In this equation, (X, Y, Z) is the coordinate of a scene point in an arbitrary 3D coordinate system, and $(x, y) = (u/w, v/w)$ is the coordinate of its projection in an image coordinate system. $\mathbb{P}$ is a 3×4 projective matrix that encodes both the intrinsic and extrinsic camera parameters. The intrinsic parameters encom-

pass the position of the origin of the image plane (principal point), focal lengths and the skew coefficient between the two axis, while the extrinsic parameters denote the coordinate system transformations from world to camera coordinates, i.e., specifying the position of the center of projection and the camera's orientation in world coordinates. The discussion in this dissertation assumes *calibrated* cameras, i.e., both the intrinsic and extrinsic parameters are known a priori. Automatic camera calibration is a mature topic in the literature of computer vision. We refer interested readers to [7, 19, 65] for a comprehensive treatment of auto-calibration.

2.2.2 The Correspondence Problem

Solving the correspondence problem, i.e., for each point in a reference frame locating its corresponding matching points in other images, is the core problem of multi-view stereo. Most researchers implicitly assume the scene is composed of Lambertian objects and rely on the brightness constancy (corresponding points have the same intensity observed from different viewpoints) as the matching criteria to establish correspondences. Obviously, this Lambertian or intensity-consistent assumption does not hold for real-world scenes, and specularities, reflections, or transparency typically yield problems to stereo algorithms. Even when the Lambertian assumption holds, stereo correspondence remains a difficult vision problem for the following reasons.

- **Noise.** There are always uncertain intensity values due to light variations, image blurring, and sensor noise introduced by the image formation process.

- **Repetitive patterns and textureless regions.** The intensity-consistency constraint is no longer valid for scenes that contain repetitive patterns or textureless

regions.

• **Depth discontinuities.** Preserving sharp depth discontinuities along object boundaries is especially important for some applications such as view synthesis and 3D reconstruction.

• **Occlusions.** Partially occluded pixels (i.e., points visible from only one camera) should not be matched with pixels in the other view. Correctly identifying and handling occluded points is important for high-quality depth estimation.

Traditionally, there are two common approaches, namely feature- and area-based stereo algorithms, to alleviate the matching ambiguities. Feature-based approaches only attempt to establish correspondences for distinct feature points that can be matched unambiguously [22, 66, 67]. While salient features can be matched stably, these approaches have the drawback of yielding only sparse or semi-dense depth estimates. Area-based approaches consider larger image regions that contain richer information than individual pixels to yield more stable matches. As for the matching function employed, this is typically based on the dissimilarity between the two vectors representing the support regions (typically a squared window) in the stereo images, e.g., the *Sum of Absolute Differences* (SAD) or *Sum of Squared Differences* (SSD). The major problem of area-based approaches is they commonly assume that pixels within the support region have the same disparity, which is not necessarily valid for pixels near depth discontinuities or non-frontal-parallel surfaces. Therefore, improper choice of the size and shape of the matching window leads to poor depth estimates. More detailed discussion about area-based approaches will be later given in Section 2.3.

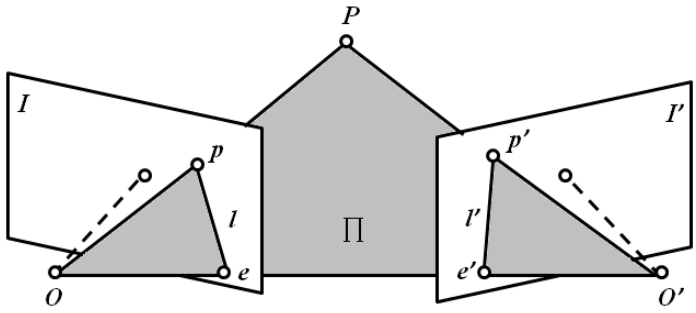


Figure 2.2: Epipolar geometry: The 3D point P, the optical centers O and O' of the two cameras, and the two images p and p' of P all lie in the same plane.

2.2.3 Binocular Stereo Geometry

So far we have discussed how image correspondences can be established. We now turn to the question of where to search for potential matches. Consider the images p and p' of a 3D point P observed by two cameras with optical centers O and O', respectively. As illustrated in Figure 2.2, these five points belong to the *epipolar plane* Π defined by the two intersecting rays OP and $O'P$. In particular, the point p' lies on the line l' where Π and the image plane I' of the second camera intersect. The line l' is the *epipolar line* associated with the point p, and it passes through the point e' where the baseline OO' intersects I'. Likewise, the point p lies on the epipolar line l associated with the point p', and the line l passes through the intersection e of the baseline with the image plane I.

The points e and e' are called the *epipoles* of the two cameras. The epipole e' is the projection of the center of projection O of the first camera in the image observed by the second camera and vice versa. As can be seen, if p and p' are images of the same point, then p' must lie on the epipolar line associated with p. This *epipolar constraint* plays a fundamental role in stereo matching because the search of correspondences

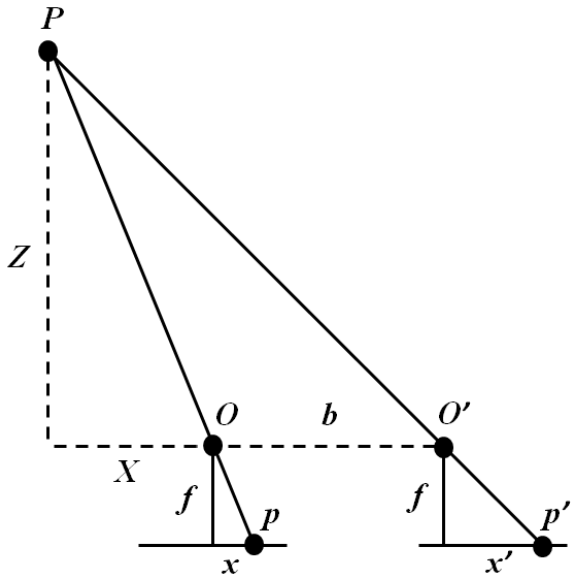


Figure 2.3: Stereo geometry: The figure shows a top-down view of two identical parallel cameras with focal length f and camera baseline b. The disparity of a scene point P of depth Z is $d = x - x' = -fb/Z$.

can be restricted to a line instead of the whole image space, greatly limiting the search range. Given a pair of calibrated cameras, the epipolar geometry can easily be computed from the explicit camera configurations [7, 63]. P's position can also be reconstructed from p and p' via triangulation.

A simple epipolar geometry as depicted in Figure 2.3 results from two identical, parallel cameras whose image planes coincide and whose x-axes are parallel to the camera baseline. In this scenario, corresponding epipolar lines are image scanlines and matched pixels p and p' must have identical y-coordinates. This special camera configuration greatly simplifies the correspondence problem since the explicit computation of epipolar lines is no longer required. In addition, for area-based stereo matching approaches two rectangular image regions can be compared directly without the need of image warping or interpolation. Due to these advantages, most stereo

Figure 2.4: A stereo rig with two parallel cameras that satisfy the simple epipolar geometry.

systems adopt this camera configuration. One way of manually achieving the simple stereo geometry is to carefully mount and adjust the cameras so that they are perfectly parallel. An example stereo rig with parallel camera setup is shown in Figure 2.4. For cameras that are not perfectly aligned, fortunately, there is a process called *rectification* that can transform the two input images so that their epipolar lines are aligned horizontally. Rectification of stereo images can be achieved by applying image warping using two 3×3 homographies computed from the camera parameters [68–70]. A pair of stereo images before and after rectification is shown in Figure 2.5.

Given two rectified images and a pair of corresponding points $p(x, y)$ and $p'(x', y')$, the correspondence can be expressed as a disparity value d. The disparity between points p and p' is defined as the difference of their horizontal image coordinates as $d = x - x'$. Note that $y \equiv y'$ since corresponding pixels must be on the same scanline for rectified images. Throughout this dissertation, unless specifically stated, we define the output of a stereo algorithm to be a dense disparity map that records the disparity value for every pixel in the reference image. In the following, when there

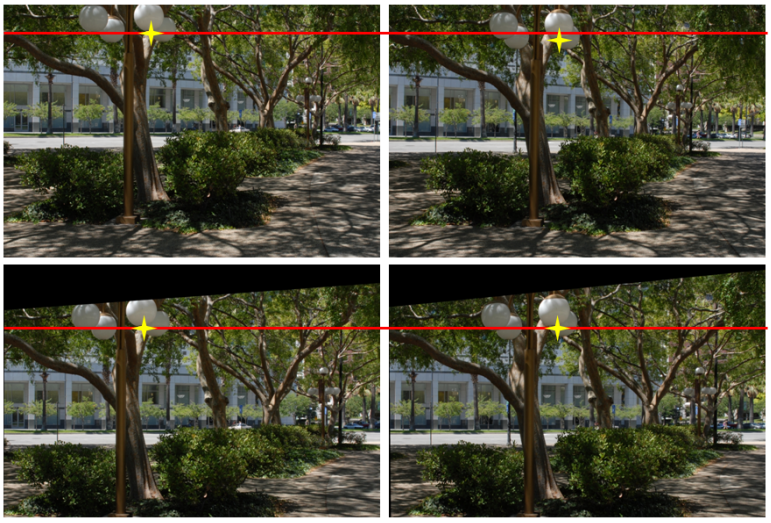


Figure 2.5: A pair of stereo images before and after rectification. The top two are the original images, while the bottom two are the rectified images. Note that the corresponding features are on the same scanline after rectification.

is no confusion we will omit (x, y) and write $d = p - p'$ for conciseness and notation clarity.

In Figure 2.3, we illustrate how the disparity of a pixel is related to its scene depth for two parallel cameras with the simple epipolar geometry. Given the 3D point $P(X, Y, Z)$ and its 2D projections $p(x, y)$ and $p'(x', y')$, we can derive equations (2.1) from the relationship of similar triangles as

$$\frac{x}{f} = \frac{X}{Z} \quad \text{and} \quad \frac{x'}{f} = \frac{X + b}{Z}, \tag{2.1}$$

where constants f and b denote the camera focal length and baseline, respectively. The disparity $d = x - x'$ of p is therefore $d = -\frac{fb}{Z}$ and is proportional to focal length and baseline, and inversely proportional to its depth Z.

2.3 A Framework for Stereo Algorithms

Following the taxonomy and evaluation of dense stereo matching algorithms presented by Scharstein and Szeliski [18], stereo algorithms generally perform (subsets of) the following four steps:

1. Matching cost computation;
2. Cost aggregation;
3. Disparity computation and optimization; and,
4. Disparity refinement.

In this section, we briefly review these key building blocks from which most existing stereo methods are constructed.

2.3.1 Matching cost computation

To establish pixel correspondences, all stereo algorithms must have a cost criteria to measure the similarity between pixels. A matching cost is therefore a value indicating how likely two pixels correspond to the same scene point. Popular pixel-based matching costs include absolute differences (AD), squared differences (SD), sampling-insensitive calculation of Birchfield and Tomasi (BT) [71], and their truncated variants, both on gray and color images.

Besides the above methods, there are filter based cost functions that are designed to compensate global intensity changes (e.g., due to gain and exposure differences, image noise, different camera settings, etc.). Typically the input images are filtered with

certain types of filter and then the transformed images are matched using common criteria, e.g., AD and SD. Popular filters include *Laplacian of Gaussian* (LoG) [72], *rank filter* [73] and *mean filter*. *Normalized cross correlation* (NCC) is another method for measuring the pixel dissimilarity. The normalization within a rectangular area effectively compensates variations in gain and bias. The main drawback of NCC is that it tends to blur depth discontinuities more than many other matching costs. A comprehensive evaluation of several matching costs can be found in [74].

2.3.2 Cost Aggregation

The pixel-based matching costs are usually ambiguous because the information available at a single pixel is not enough for finding an unambiguous match. To reduce matching ambiguities, local area-based methods aggregate the matching cost by summing over a support region. A support region is typically a squared window centered on the current pixel of interest. Conventional 2D aggregation methods smooth the cost volume by computing the weighted average using the box or Gaussian filters [75]. An advantage of using linear filters for cost aggregation is that the 2D convolution process is separable and very fast implementations can be achieved. In terms of drawback, these methods tend to blur object boundaries with the fixed size window. To avoid the *"fattening artifacts"* near depth discontinuities, shiftable windows [4, 76], windows with adaptive sizes [77–79] or adaptive weights [3, 80, 81] have been developed. We refer readers to two recent survey papers [14, 82] for the state-of-the-art cost aggregation methods and their performances .

2.3.3 Disparity Computation and Optimization

In general, stereo algorithms can be categorized into two major classes: *local* methods and *global* methods. In local methods, the per-pixel disparity is simply selected by a local "winner-takes-all" (WTA) strategy, i.e., the disparity associated with the minimum aggregated cost at each pixel is chosen. Therefore, a local method's accuracy performance depends largely on the matching cost computation and the cost aggregation stages. Local methods can be every efficient (computationally feasible for real-time implementations), but accuracy-wise they are sensitive to sensor noise and locally ambiguous regions (textureless areas, occlusion boundaries, etc.) in images because only local information from a small number of pixels surrounding the pixel of interest is utilized to make the decision.

In contrast, global methods make explicit assumptions about the scene depth field and are usually formulated in an energy-minimization framework. The most widely used assumption is that the scene is locally smooth (except for object boundaries) and neighboring pixels should have very similar disparities. This constraint is referred to as the "*smoothness constraint*" in the stereo literature. The standard and classical global stereo formulation aims to find an optimal disparity assignment function $f(p)$ that minimizes the following cost function

$$E(f) = E_{data}(f) + E_{smooth}(f), \tag{2.2}$$

where the first term, the data energy, $E_{data}(f)$ comes from the matching costs and penalizes disparity assignments that are inconsistent with the observed image data, whereas the second term, the smoothness energy $E_{smooth}(f)$, encourages neighboring

pixels to have similar disparities based on the assumption that the scene is piecewise smooth. To make the optimization computationally tractable, the smoothness energy is often defined to measure the differences only between neighboring pixels disparities, e.g., using the common *Potts model* [83] or the *truncated linear model* [84]. Once the global energy has been defined, the lowest energy corresponding to the optimal disparity assignment can be approximately achieved using the methods surveyed by Szeliski et al. [85]. Among these energy minimization approaches, Belief Propagation (BP) [84,86] and Graph Cuts (GC) [16,87] are particular favored by stereo researchers. Recent literature shows that BP- and GC-based stereo methods can produce the state-of-the-art results in terms of reconstruction accuracy [2, 12, 13, 88]. Global methods are less sensitive to the problems suffered by local methods since prior constraints provide regularization for regions difficult to match. However, global methods are usually more computationally expensive than local methods.

A different class of global optimization algorithms are those based on dynamic programming (DP) [4, 89, 90]. Unlike DP or GC which approximate the global minimum of the cost function defined in 2.2 over the 2D pixel grid, DP finds the global minimum of 2.2 for each image scanline independently in polynomial time. The main problem with DP is the difficulty of enforcing disparity consistency between scanlines.

2.3.4 Disparity Refinement

Most stereo algorithms estimate disparities in the discretized integer space. While integer disparities may be sufficient for applications such as segmentation and tracking, for view synthesis or 3D reconstruction, such quantized disparity maps usually result

in unappealing visual artifacts. To overcome this limitation, many stereo algorithms utilize a sub-pixel refinement stage to refine the initial integer disparities. A simple yet effective solution to increase depth resolution is to fit a parabolic curve to the matching costs at discrete disparity levels [91,92]. In addition to sub-pixel refinement, there are other disparity post-processing schemes available. For example, occlusion regions are usually detected using left-right consistency check [92,93] and unmatched pixels can be filled via interpolation or depth completion algorithms [64,94]. Median filer can also be used to remove small isolated mismatches. Due to the low computational complexity and the edge-preserving property, median filer based refinement is particularly favored by real-time stereo algorithms [14,80].

2.4 Stereo Quality Measures

Because there are so many algorithms for stereo correspondences, stereo images with "ground truth" disparity maps and meaningful quantitative evaluation are critical to assess the performances of existing methods and gauge the progress of stereo matching. Scharstein and Szeliski [18] provided a very comprehensive quantitative evaluation of binocular stereo algorithms and publicized several benchmark sequences with ground truth [5,18,74,95]. They also set up a web site to allow researchers to run algorithms on the benchmark data and report their comparative results online at [1].

Figure 2.4 shows the reference image and the ground truth disparities for each of the four benchmark stereo pairs used for quantitative evaluation. Of the four benchmark sequences, "Tsukuba" was originally from the University of Tsukuba [96]. The

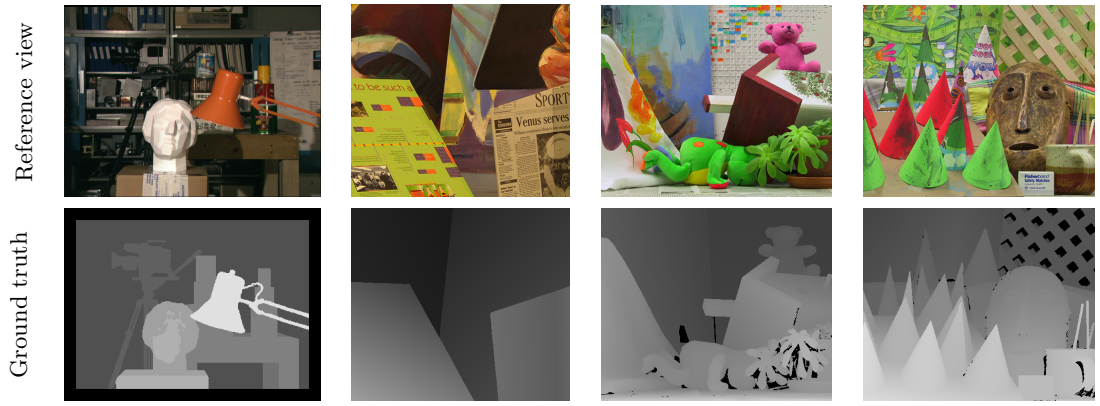


Figure 2.6: Reference images of "Tsukuba", "Venus", "Teddy" and "Cones" stereo pairs and their ground truth disparity maps.

scene contains several frontal-parallel planar surfaces and the ground truth disparities were manually labeled by hand. "Venus" was first introduced by [18] and the scene consists of piecewise (slanted) planar surfaces. The ground truth disparity map of "Venus" was computed using the affine motion estimation technique [97] designed for piecewise planar scenes. "Teddy" and "Cones" are image pairs with more complicated scene structures. Difficulties posed to the stereo algorithms include a large disparity range, complex surface shapes, textureless areas, narrow occluding objects, etc. The ground truth disparity measurements of these two sequences were acquired by coded structured light technique described in [5].

The reconstruction accuracy is measured by the percentages of bad matching (where the absolute disparity error is greater than 1 pixel). The error statistics accounts for three pixel categories:

- **nonocc** - Pixels in non-occluded areas.
- **all** - Pixels in non-occluded and half-occluded areas.

- **disc** - Pixels that near the occluded regions.

Throughout this dissertation, we mainly use these four benchmark data sets together with the online evaluation system [1] to quantitatively gauge the quality of our proposed algorithms.

Chapter 3 Related Work

This dissertation is related to a sizable body of literature on stereo vision and an exhaustive discussion of related work in stereo is beyond the scope of this dissertation. The interested readers are referred to two excellent surveys by Scharstein and Szeliski [5], and Brown et al. [23]. For general yet slightly dated surveys of the stereo literature, see Dhond and Aggarwal [98], and Barnard and Fischler [99]. In the rest of this chapter we discuss stereo methods that are relevant to the methods presented later in this dissertation (Chapters 4, 5, and 6).

3.1 Real-Time and Near Real-Time Stereo

This section focus on reviewing the progression of real-time stereo implementations over the past decade. A summary of earlier real-time stereo systems and their comparative performances can be found in [23].

Local methods establish pixel correspondences by measuring the similarity between image regions and usually have very efficient implementations [23, 82]. Representative early real-time local methods include [100,101]. A plane-sweep approach [102] is adopted to effectively use the graphics hardware to warp and process images. The central problem of local correlation-based algorithms is how to determine the size and shape of the aggregation window. For accurate depth estimation, a window must be large enough to cover sufficient intensity variations while small enough to avoid crossing depth discontinuities. This inherent ambiguity causes problems such as incorrect

disparities in textureless regions and blurred occlusion boundaries. In order to resolve these dilemmas, there has been work on varying window size and shape [79,103–106]. The basic idea is to evaluate a variety of windows with distinct sizes or shapes and select the one with the optimal matching cost. Although performing better than fixed window methods, in terms of accuracy variable window approaches are in general less powerful than global methods.

Recently, Yoon and Kweon present a new correspondence search algorithm [3] that yields high-quality results that are comparable to those obtained by global methods. The success of [3] lies in the use of joint bilateral filter for cost-volume filtering. The most attractive property of [3] is that a large window can be used to aggregate information without over-blurring occlusion boundaries. On the other hand, [3]'s bilateral filtering technique is computationally very demanding. Its execution time is comparable to that required by global methods, diminishing the efficiency advantage of local approaches. For this reason, several "*adaptive weights*" based approaches have been proposed, aiming at improving [3]'s speed performance. Mattoccia et al. [81] suggest a block-based aggregation strategy that can obtain a disparity map at a few seconds. Gupta and Cho introduce an adaptive binary window approach [107]. While strong results are demonstrated, their algorithm takes 0.46 second for a 384×288 image with 16 disparity candidates. Yu et al. develop a high performance stereo system using "*exponential step size adaptive weight* (ESAW)" technique [108]. Their approach demonstrates high data parallelism and can be efficiently mapped to GPU platform. Richardt *et al.* [109] present an approximate but real-time implementation of the bilateral filtering aggregation method. However, due to the large amount of

memory required for processing full-color images, the support weights are computed using grayscale intensities rather than three-channel color vectors, giving poor results near object boundaries. Rhemann et al. [110] present a filter-framework which achieves high-quality disparity maps efficiently. Their approach is based on the recently proposed guided filter [111], which has the edge-preserving property and a runtime independent of the filter size.

Besides from local methods, efficient global stereo algorithms have also been developed. Among the various energy minimization techniques [85], dynamic programming (DP) is of particular interest for real-time systems due to its low computational complexity. Sun [91] proposes an early DP-based stereo algorithm that executes near real-time. The image is divided into nonuniform rectangular subregions to reduce disparity search range. Gong and Yang present a stereo algorithm based on reliability-based DP [112]. Their algorithm can be implemented on the GPU and yields near real-time performance. By using a coarse-to-fine scheme and MMX instructions, Forstmann et al. [15] present an accelerated DP algorithm whose implementation achieves about 100 MDE/s on an AMD AthlonXP 2800+ 2.2G processor. Traditional DP algorithms optimize the disparity assignments on a scanline by scanline basis. The inter-scanline consistency is not enforced. A number of approaches have been proposed to address this limitation [113–115]. For example, Kim et al. [113] introduce a two-pass DP scheme that performs optimization both along and across the scanline; Lei et al. [115] optimize a global energy function defined on a 2D tree structure whose nodes are over-segmented image regions. Unfortunately, these advanced approaches in general involve more computational cost and are typically too

slow for real-time applications. In addition to DP, Yang et al. [116] propose a near real-time GPU implementation of the hierarchical belief propagation algorithm [84]. It produces better accuracy than fast DP-based algorithms but runs slower at about 17 MDE/s. Yu et al. [108] further invent a real-time "*exponential step size message propagation* (ESMP)" algorithm. As an extension of the aforementioned ESAW technique, by incorporating the smoothness prior commonly used in global stereo, ESMP improves the accuracy at the cost of lower speed in comparison with ESAW.

3.2 Regularization Priors for Global Stereo

Over the last decade, dense stereo has made considerable progress, in part because the problem can be cast in a global optimization framework for which there exist powerful inference algorithms such as graph cuts and belief propagation that can efficiently find good minima of the cost function. According to the widely used Middlebury benchmark [1], almost all top-performing stereo methods are formulated as an energy minimization problem and rely on belief propagation or graph cuts. These global methods give substantially more accurate results than were previously possible. In contrast to local methods, global methods allow us to utilize the prior constraints that encode the assumptions on scene structures to regularize the ill-posed matching problem. In this section we review several most widely used prior constraints used by global stereo methods for high-accuracy depth estimation.

The most conventional regularization prior used by early global methods [117,118] to produce piecewise smooth disparity maps is the first-order smoothness prior, e.g., the Potts model [83] or the truncated linear model [84], which encourages neighbor-

ing pixels to have similar disparities and thus favors low-curvature fronto-parallel surfaces. However, even in man-made scenes, the fronto-parallel assumption is not always valid. To overcome this limitation, the usage of second-order (penalizing large second derivatives of depth or disparity) smoothness priors were proposed. Ogale and Aloimonos [119] propose a "slanted scanline" algorithm, in which straight, 3D line segments are fitted to image scanlines using an optimization method. Their method, similar to dynamic programming, does not enforce inter-scanline consistency. Recently, Woodford et al. [88] show that second-order smoothness priors can be incorporated into graph cuts based stereo reconstruction. The authors introduce an effective optimization energy functions with triple cliques for second-order prior terms.

Methods based on second-order priors produce excellent results. However, they favor piecewise planar surfaces, which are not ideal for dealing with curved surfaces. Li and Zucker [120] introduce priors on slanted and curved surfaces, encouraging the second and third derivatives of disparity to be zero. This allows for curved surfaces in the solution and significantly improves on the piecewise planar assumption . However, the algorithm requires the surface normals to be pre-computed and in fact optimizes a first-order prior on the normals, rather than a second-order prior on the disparities.

The other popular regularization prior that used to improve the stereo reconstruction accuracy is the segment-based prior proposed by Birchfield and Tomasi [121] and Tao et al. [17]. Tao et al. in [17] assume that each color segment corresponds to a planar surface in 3D, and this key idea has inspired many stereo researchers and form the basis for most top-performing stereo algorithms. For instance, as of November

2012, almost all of the top performing stereo methods listed by the Middlebury benchmark [1] use color segmentation, either explicitly or indirectly, to estimate disparity maps [12, 13, 122, 123]. Although segment-based priors usually improve disparity estimates in textureless areas, they inevitably introduce errors in heavily textured regions and do not handle well the situation that the scene contains high-curvature details.

In order to overcome the limitations of segment-based approaches, Smith et al. in [2] propose a nonparametric smoothness prior that correlates pixels with similar features in a large neighborhood. Gallup et al. [124] introduce a binary classification procedure to classify the scene into planar and non-planar and then employ different algorithms for different image regions. The new regularization prior we propose in Chapter 5 is derived from measured or reliably matched control points. Our regularization term models the influence from control points in the global inference stage and allows stereo algorithm to better handle problematic regions such as textureless areas and occlusion boundaries. A nice feature of our method is that it does not require hard color segmentation, plane fitting, or local surface normals to be pre-computed. Experiments show that our method is comparable to [2] in terms of accuracy while superior to [2] for efficiency. And unlike [124], our method is not limited to handling urban scenes and does not need training images.

By using ground control points (GCPs) as constraints, our work also relates to semi-dense stereo methods. In the stereo literature, GCPs are referred to as the high confidence matches, started by the work of [4]. Due to the inherent ambiguities of stereo correspondence, instead of computing dense disparity maps, there are techniques invented to find unambiguous components to generate semi-dense disparity

maps [66, 67, 125, 126]. Many of these approaches require GCPs obtained via feature correspondences as seeds to start the matching process. New matches are added in a progressive manner until certain termination criteria is met. Wei et al. combined the progressive scheme with region-based approaches to produce dense matching [127]. In contrast, our method is both dense and global. The Bayesian inference framework does not require iterative region growing or hard image segmentation.

GCPs have also been used by dynamic programming based stereo methods as hard constraints. Bobick et al. incorporated GCPs into a DP framework by forcing DP to choose a path through the GCPs [4]. Different from their one dimensional scanline optimization model, by incorporating GCPs into a global inference framework, our method is able achieve full frame optimization. Furthermore, instead of treating GCPs as hard constraints, our formulation models GCPs as soft constraints and does not require all GCPs to be perfect. In [76, 128], GCPs are used in preprocessing stages to restrict the disparity search ranges. In contrast, our method makes no hard constraint on the disparity search ranges.

3.3 Stereo Beyond Lambert

All stereo depth recovery methods make explicit or implicit assumptions about which image features are held constant. The primary differences arise from the choice of invariant. A number of possible invariants that allow stereo matching for Lambertian scenes have been explored [74]. In this section, we review constraints employed by stereo matching techniques for non-Lambertian surfaces.

Stereo matching of specular surfaces has most commonly been approached by

treating specularities as outliers to the brightness constancy invariant, which should be detected and either removed or avoided [129–133]. An alternate approach treats surfaces as diffuse-plus-specular and formulates a mult-iview constraint that all observations must lie on a line in color space [134]. Unfortunately, all of these methods limit the range of surface BRDFs to those which can be represented as a simple combination of diffuse and specular terms. The light transport constancy invariant presented in this work allows stereo matching of surfaces with completely arbitrary BRDF.

Jin et al. show that a multi-view rank constraint on reflectance complexity is implied by a diffuse-plus-specular surface model and use this constraint to reconstruct non-Lambertian surfaces [135]. Although the method proposed in this‘dissertation also formulates a rank constraint, we rely on a different matching invariant and allow for truly arbitrary surface BRDFs at each scene point.

Helmholtz stereopsis [136–139] allows matching of arbitrary BRDFs using reciprocity. That is, $R(x_i, \theta_A, \theta_B) = R(x_i, \theta_B, \theta_A)$. By collocating point light sources with each camera it is possible to record reciprocal pairs using two different lighting conditions, such that image A is illuminated by light B, and image B is illuminated by light A. Due to reciprocity, the reflected light to cameras A and B will be equal. Unfortunately, this method requires the light sources be collocated with the optical center of each camera. Although acceptable results are possible by simply placing the light nearby, a proper implementation requires calibrated optics to ensure collocation. The method presented in this dissertation makes use of a different property and does not require the position of light sources to be precisely calibrated or even known.

Orientation constancy has been used to allow reconstruction of scenes with arbitrary BRDF in both photometric stereo and multi-view stereo configurations [140, 141]. Although very accurate results are possible, these methods require a known calibration object with BRDF similar to the unknown scene, as well as distant cameras and light sources. In contrast, this work does not require a known object and allows for arbitrarily located light and camera positions.

Unlike many previous approaches that make use of geometric illumination changes, our formulation requires *radiometric* illumination variations, i.e., rather than changing position, the light sources in our work change only their intensities. Prior approaches using radiometric variations include structured light (e.g. [5, 142]), and the more general space-time stereo framework [143, 144]. Image intensity ratios are also a well studied method for recovering depth which are often formulated as using radiometric variation [145, 146].

It is argued in [147] that image ratios are only applicable to diffuse surfaces. Our method is fundamentally different from that work in that we assume radiometric variations only while the derivation in [147] is based on an illumination distribution which includes geometric variation. Experimental results demonstrate that using radiometric variation, scenes with arbitrary surface BRDFs can be effectively reconstructed using image ratios.

The invariant proposed by this dissertation, named light transport constancy, has not previously been explored for stereo matching. However, in the case of laser scanning, it was explicitly identified and articulated by Curless and Levoy [148]. In addition, it has implicitly been used in other domains. Magda et al. capture hundreds

of images illuminated by precisely calibrated light source positions on two concentric spheres surrounding an object. The two sampled representations of the incoming illumination field can then be aligned to find the depth of a given scene point [136].

Chapter 4 Real-Time Stereo Using Approximated Joint Bilateral Filtering and Dynamic Programming

In this chapter, we present a stereo algorithm that is capable of estimating scene depth information with high accuracy and in real-time. The rest of this chapter is organized as follows: Section 4.1 gives a high level overview of our approach. Section 4.2 contains background material about bilateral filter and its application in stereo correspondence problem. In Section 4.3, we present a precise description of our proposed algorithm followed by Section 4.4, which is about specific GPU implementation issues. We evaluate our algorithm with experiments in Section 4.5 and summarize this chapter in Section 4.6.

4.1 Algorithm Overview

Our algorithm is inspired by the idea of cost-volume filtering via edge-preserving filters, started from [3, 149], which introduce cost aggregation schemes that use a fix-sized window with per-pixel varying support weight. The support weights are computed based on the color similarity and geometric distance to the center pixel of interest. In fact, taking both geometric distance and photometric similarity of neighboring pixels into account to construct the filter kernel is the key idea behind bilateral filtering [150]. Although bilateral filter based aggregation methods have proven to be effective, their applications in real-time stereo are limited by their speed. It is nonlinear and its computational complexity grows quadratically with the kernel

size. Brute-force implementations are on the order of minutes for generating a small depth map [3].

In this chapter, we first attempt to reformulate [3]'s aggregation algorithm using a fast separable implementation of the bilateral filter. In the first pass the raw cost-volume is bilaterally filtered in the horizontal direction using a 1D kernel and the intermediate matching costs are bilaterally filtered subsequently in the vertical direction. The separable approximation reduces the complexity of the aggregation approach from $O(|I|N\ell^2)$ to $O(|I|N\ell)$, where $|I|$ and N are the image size and disparity search range respectively and ℓ is the kernel width of a square window. Our approximation, which is an effective trade-off between speed and accuracy, leads to fast cost-volume filtering and satisfactory results. Motivated by its suitability for hardware implementation, we propose a GPU implementation which further improves the speed by one or two orders of magnitude.

In addition to the GPU-based local WTA solution, we further incorporate our two-pass aggregation scheme into a DP scanline optimization framework for improved reconstruction accuracy. We found that changing the window shapes from conventional squares to vertical rectangles allows robust performance near depth discontinuities and effectively alleviates DP's scanline inconsistency artifacts. The aggregated cost-volume is transferred back from the GPU to CPU memory for DP optimization. Thus our approach not only makes use of both CPU and GPU in parallel, but also makes each part do what it is best for: the graphics hardware performs cost aggregation in massive parallelism, and the CPU carries out DP that requires more flexible looping and branching capability. The current implementation is capable of running

at video frame rate. In terms of accuracy, quantitative evaluation using data sets with ground truth disparities shows that our approach is among the state-of-the-art real-time stereo algorithms. Combined with its high speed capability, our algorithm is suitable for many real-time applications that require high quality depth data. This stereo formulation that built on fast approximate bilateral cost-volume smoothing and dynamic programming optimization is the main contribution of this chapter.

4.2 Bilateral Filter and Its Application in Cost Aggregation

Before describing our proposed stereo algorithm, we start with a brief description of bilateral filtering and Yoon and Kweon's adaptive weights stereo algorithm [3].

The bilateral filter is a filtering technique to smooth an image while preserving edges [150]. One of its variants, the joint bilateral filter [151], smoothes an image with respect to edges in a different image. Its basic formulation is very similar to Gaussian convolution: each pixel is replaced by a weighted average of its neighbors. The core difference is that the bilateral filter takes into account the dissimilarity in pixel values with the neighbors while constructing the blur kernel. More formally, given an image I and a central pixel $p \in I$ (we use the notation I_p for the pixel value at position p), the support weight $w(p, q)$ of p's neighbor q is written as:

$$w(p,q) = \exp(-\frac{\|I_p - I_q\|}{\sigma_c} - \frac{\|p - q\|}{\sigma_g}), \tag{4.1}$$

where $\|I_p - I_q\|$ and $\|p - q\|$ represent the color dissimilarity (Euclidean distance between pixel values) and the spatial distance between p and q, respectively. The bilateral filter is controlled by two parameters σ_c and σ_g. These two values respectively

control the influence from intensity/color similarity and spatial proximity. An image filtered by a bilateral filter $BF(\cdot)$ is defined by

$$BF(I)_p = \frac{\sum_{q \in \Omega_p} w(p,q) \cdot I_q}{\sum_{q \in \Omega_p} w(p,q)}, \tag{4.2}$$

where Ω_p denotes the set of all pixels in the support region and the normalization factor $\sum_{q \in \Omega_p} w(p,q)$ ensures support weights sum to one. More interesting properties, implementation details, and applications of bilateral filtering can be found in [152].

Yoon and Kweon [3] utilize the bilateral filtering as an aid in local WTA stereo. Given a pair of stereo images $\{I, I'\}$, the raw matching cost between pixels is written as $\widetilde{C}(p,d)$ where p represents the pixel location in the reference view I and d is a disparity hypothesis. In [3] the final cost-volume is computed as a weighted average of raw matching costs

$$C(p,d) = \frac{\sum_{q \in \Omega_p, q' \in \Omega'_p} w(p,q) w(p',q') \widetilde{C}(q,d)}{\sum_{q \in \Omega_p, q' \in \Omega'_p} w(p,q) w(p',q')}, \tag{4.3}$$

where $p' = p - d$ represents $p's$ corresponding pixel in I' given disparity d.

Note that unlike conventional bilateral filtering, equation (4.3) takes into account the support weights in both stereo images. According to the authors' explanation and our experimental observations, combining the support weights in both windows helps to improve correspondence search, especially for pixels near occlusion boundaries. In terms of speed, the proposed method however is computationally more expensive than other window-based local stereo algorithms. The reported running time for the benchmark "Tsukuba" image with a 35×35 support window is about one minute on an AMD AthlonXP 2700+ 2.17G processor.

4.3 Algorithm Description

In this section, we present the proposed stereo formulation. Given multiple images taken from different viewpoints, the goal of a stereo algorithm is to establish pixel correspondences across images. For the scope of this chapter, we focus on dense two-frame stereo and assume the input stereo images $\{I, I'\}$ are rectified, i.e., the epipolar lines are aligned with corresponding scanlines.

Following the taxonomy in [18], our algorithm consists of three major steps: 1) matching cost computation; 2) adaptive cost-volume filtering; and 3) disparity optimization via DP. Details about each module are presented below. Besides from these key components, in all experiments, a 3×3 median filter based disparity refinement step is employed to remove isolated noises from disparity maps.

4.3.1 Matching Cost Computation

The matching cost computation step initializes the cost-volume $\widetilde{C}(p, d)$ by computing raw pixel-wise matching costs. Using the brightness constancy constraint, pixels that correspond between the left and right images should have similar intensities. Thus we adopt the widely used absolute difference (AD) dissimilarity function to measure the difference between two corresponding pixels:

$$\widetilde{C}(p, d) = \min(\frac{\sum_{c \in \{R,G,B\}} |I_p^c - I_{p-d}^c|}{3}, C_{max}), \tag{4.4}$$

where the parameter C_{max} ($0 < C_{max} \leq 255$) is a truncation value. The truncation is necessary to make the matching costs robust to occlusion and non-lambertian

objects that violate the brightness constancy assumption. For every pixel $p(x, y) \in I$, we loop through all disparity hypotheses to calculate their matching costs using equation (4.4). In the end, we obtain the initial cost volume $\widetilde{C}$, which is a three-dimensional array that can be indexed by x, y, and d.

4.3.2 Fast Adaptive Cost-Volume Filtering

We modify [3]'s approach as our baseline cost aggregation algorithm. Two major changes are: 1) In Yoon and Kweon's work, the similarity between two pixels within the support window is measured in the CIELab color space. Our approach however measures the color proximity in the RGB color space for simplicity and efficiency; 2) Inspired by [109], we reformulate equation (4.1) as

$$w(p, q) = \exp(-\frac{\|I_p - I_q\|}{\sigma_c})\sqrt{\exp(-\frac{\|p - q\|}{\sigma_g})}, \tag{4.5}$$

where the square root is applied to the geometric proximity weight, so that $w(p, q) \cdot w(p', q')$ in equation (4.3) involves the proximity weight only once. In our baseline implementation we employ a 35×35 support window. The running time for the "Tsukuba" sequence is about 25 seconds on an Intel Xeon 2.66GHz processor with our not fully optimized implementation.

As can be seen, the full-kernel implementation of the bilateral cost-volume filtering is computationally expensive because the pixel-wise support weights need to be recomputed for every pixel. Unfortunately, unlike separable box and gaussian filters which have very fast implementations, bilateral filter is not separable in theory due to the color dependent term in equation (4.5). Nevertheless, in order to address the

crucial runtime issue, Pham and van Vliet [153] attempt to approximate the full-kernel bilateral filter using two separate 1D kernels. Their separable implementation is applied to video enhancement and compression. Ansar et al. [149] first apply bilateral filtering to stereo and conclude that a separable approximation is adequate. However, no thorough analysis or comparison is proposed and the performance of this acceleration in stereo correspondence remains unclear.

In this section we revisit the separable approximation and attempt to speedup the bilateral aggregation using a two-pass implementation: a 1D bilateral filter is applied to smooth the cost-volume along the first dimension (either horizontal or vertical) and the intermediate results are filtered in the subsequent dimension. In essence, this simplified approach reformulates equation (4.3) as

$$C^{tmp}(p,d) = \frac{\sum_{u=x-\frac{\ell}{2}}^{u=x+\frac{\ell}{2}} w(p(x,y),q(u,y))w(p',q')\widetilde{C}(q,d)}{\sum_{u=x-\frac{\ell}{2}}^{u=x+\frac{\ell}{2}} w(p(x,y),q(u,y))w(p',q')} \tag{4.6}$$

$$C(p,d) = \frac{\sum_{v=y-\frac{\ell}{2}}^{v=y+\frac{\ell}{2}} w(p(x,y),q(x,v))w(p',q')C^{tmp}(q,d)}{\sum_{v=y-\frac{\ell}{2}}^{v=y+\frac{\ell}{2}} w(p(x,y),q(x,v))w(p',q')}, \tag{4.7}$$

where C^{tmp} is a temporary buffer to store the matching costs obtained from the first pass.

As aforementioned, this separable implementation does not produce exactly the same results as the full-kernel filtering because of the non-separability of $w(\cdot,\cdot)$ in equation (4.3). Figure 4.1 shows both the original and approximated support weights for several selected pixels in the "Tsukuba" image. In most cases, especially for uniform areas and axis-aligned edges, the original support weights are very similar to their approximated counterparts. For the rightmost patch which contains two

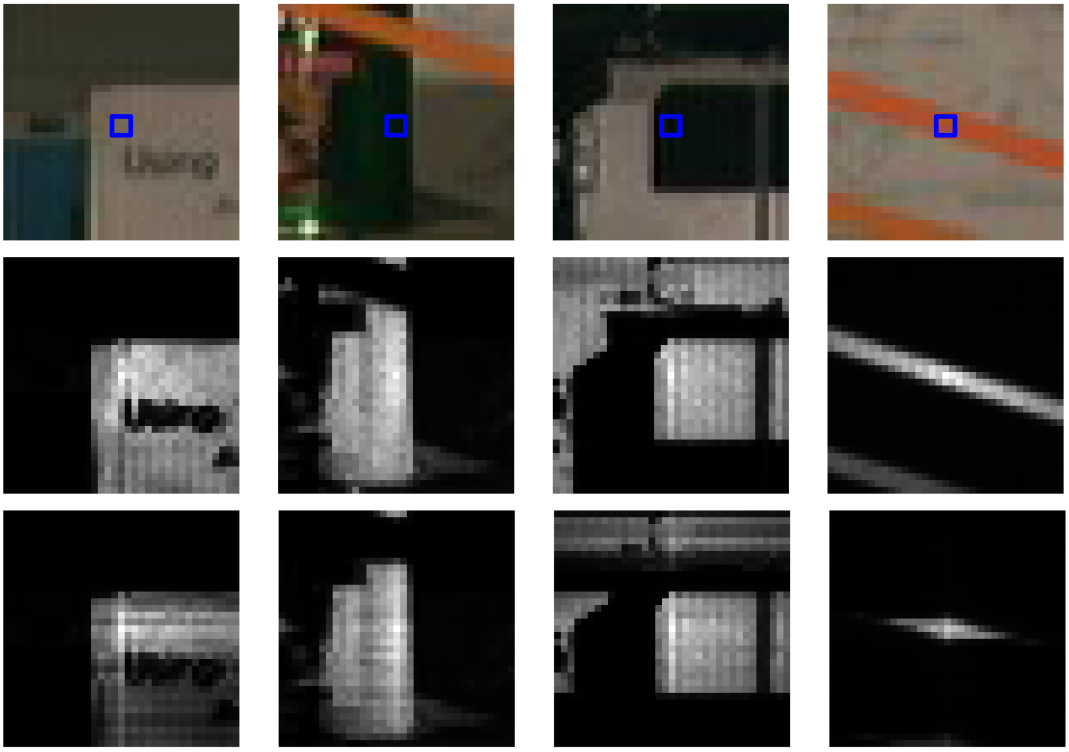


Figure 4.1: A comparison of full-kernel with approximated support weights. (top row) close-up views at several pixel locations in the "Tsukuba" image. The blue square marks the center pixel of interest. (second row) the original 35×35 support weights. (third row) the corresponding support weights computed using our two-pass approximation.

diagonal line structures, our approach still tends to assign higher weights to pixels that are closer or with similar color but spatially the support weights attenuate much faster compared to the original 2D kernel. In this scenario where there are thin diagonal structures, our approximation is similar to a full-kernel with a smaller support region.

In Figure 4.2 we further provide visual and quantitative comparisons of the achieved disparity maps. Compared to the full-kernel filtering, the separable bilateral smoothing still performs edge-preserving cost aggregation effectively. While visually similar disparity maps can be obtained, as expected, quantitative evaluation with ground truth data confirms that the two-pass approximation yields slightly less accurate re-

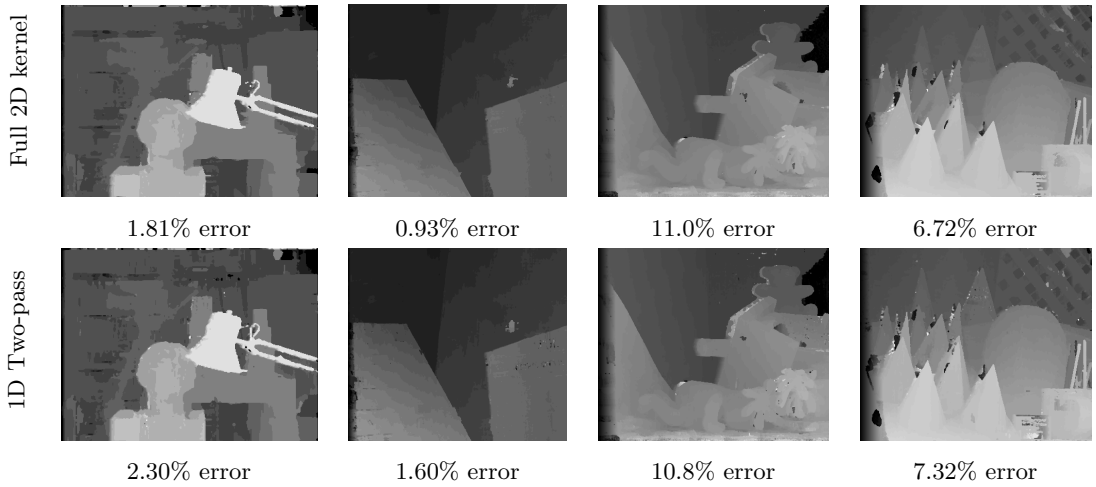


Figure 4.2: Disparity maps for the Middlebury benchmark data generated from (top row) full-kernel (35×35) bilateral cost aggregation and (bottom row) the separable two-pass approximation, respectively. Identical parameter settings are used to generate these results. Error disparity percentages are measured in non-occluded areas.

sults, especially for textured regions. It is worth noting that neither implementation achieves the accuracy numbers reported in [3]. We believe this is mainly due to the left-right consistency check and occlusion filling post-processing steps adopted by [3]. Similar observation and conclusion can also be found in [109]. In terms of speed, this two-pass acceleration dramatically speeds up the computation, reducing the complexity per disparity estimation from $O(\ell^2)$ to $O(\ell)$. For instance, for the "Tsukuba" image our result is generated in 1.9 seconds while the full-kernel approach takes about 32 seconds (kernel width $\ell = 35$). On the other hand, the downside of this approximation is that its resultant disparity maps are less smooth than the brute-force implementation and there is no formal characterization of their differences in accuracy. As a consequence, this two-pass aggregation scheme produces an interesting trade-off between accuracy and speed.

4.3.3 Disparity Optimization via DP

In this section, DP is performed for disparity optimization. As an early framework introduced for the stereo correspondence problem, DP is still one of the most popular techniques for its 1D optimization capability and high efficiency.

DP-based algorithms formulate stereo correspondence as a least-cost path finding problem. Given an image scanline $S_y = \{p(\cdot, y)\}$, DP finds an optimal path through a 2D slice $C(\cdot, y, \cdot)$ of the 3D cost-volume. The optimal path is equivalent to a disparity assignment function $f(p)$ that minimizes the global cost function defined in 2.2. In this chapter, $E_{data}(f) = \sum_{p \in S_y} C(p, f(p))$ comes directly from the aggregated matching costs. $E_{smooth}(f)$ is defined as

$$E_{smooth}(f) = \lambda_s \cdot \sum_{p \in S_y} \sum_{q \in \xi_p} \max(\exp(-\frac{|I_p - I_q|^2}{\sigma_s}), \epsilon) \cdot \min(|f(p) - f(q)|, \tau), \qquad (4.8)$$

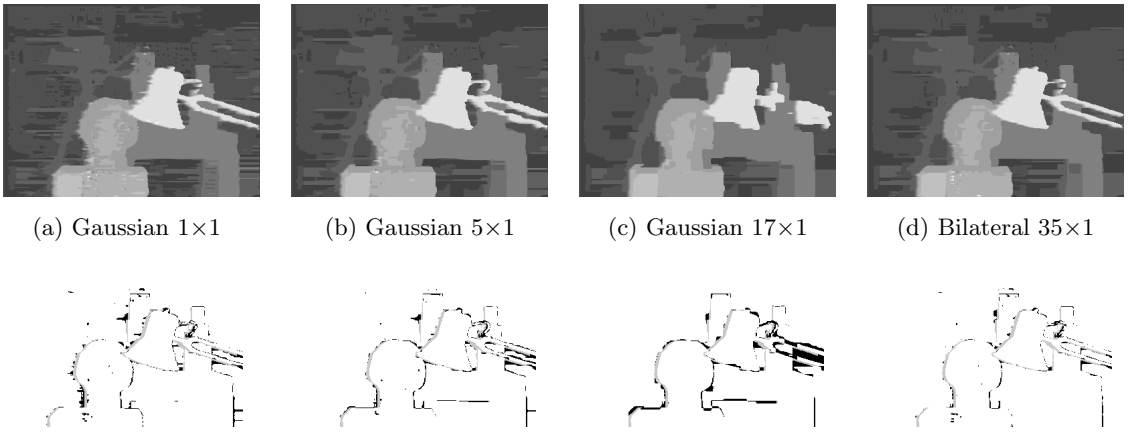
where λ_s is the rate of increase in the smoothness cost; $\xi_{p(x,y)}$={p(x-1,y),p(x+1,y)} and $\exp(-|I_p - I_q|^2/\sigma_s)$ is a monotonically decreasing function of intensity differences that lowers smoothness penalty costs at high intensity gradients; parameters σ_s and ϵ control the sharpness and lower bound of the exponential function, respectively. In order to allow for sharp depth edges, the smoothness cost stops growing after the disparity difference becomes large. Parameter τ controls the upper bound of discontinuity penalty between neighboring pixels.

Energy functions with the form defined in equation (2.2) can be minimized by DP. For each scanline S_y in the reference view we construct a cost matrix M and an ancestor matrix A. Both M and A have $N \times W$ entries, where N and W represent the disparity range and image width, respectively. Each entry is a potential place

Algorithm 1 Three-state DP for optimal path extraction

for $d = 0$ to $N - 1$ **do**
  $M(d, 0) = C(0, y, d)$;
end for
for $x = 1$ to $W - 1$ **do**
  compute the smoothness cost λ between $(x - 1, y)$ and (x, y) based on equation (4.8);
  $nvocc = 0$;
  for $d = N - 1$ to 0 **do**
    $cmin0 = C(x, y, d) + M(d, x - 1)$; //match state
    $cmin1 = C(x, y, d) + M(d - 1, x - 1) + \lambda$; //diagonal occlusion
    $cmin2 = M(d + 1, x) + (nvocc < \tau \ ? \ \lambda : 0)$; //vertical occlusion
    $M(d, x) = \min(cmin0, cmin1, cmin2)$;
    $if(M(d, x) == cmin0)\ A(d, x) = (d, x - 1)$; $nvocc = 0$;
    $if(M(d, x) == cmin1)\ A(d, x) = (d - 1, x - 1)$; $nvocc = 0$;
    $if(M(d, x) == cmin2)\ A(d, x) = A(d + 1, x)$; $nvocc = nvocc + 1$;
  end for
end for

along the path. We traverse M from left-to-right updating the entries in M and A. The complexity of the brute-force implementation is $O(WN^2)$ per-scanline since updating $M(d, x)$ requires considering N previous entries $M(0, x - 1) \dots M(N - 1, x - 1)$. Inspired by [4,90], we impose the common occlusion and monotonic ordering constraints [154] and employ the *three-state* (horizontal match, diagonal occlusion and vertical occlusion states) scanline optimization algorithm as outlined in Algorithm 1 to construct the optimum path. By assuming the ordering rule, three instead of N potential moves need to be considered, which greatly reduces the complexity of the pathfinding problem. After the rightmost column is filled, the optimum path can be extracted via back-tracking [90]. This DP process is repeated over all the scanlines to generate a dense disparity map.



(a) Gaussian 1×1 (b) Gaussian 5×1 (c) Gaussian 17×1 (d) Bilateral 35×1

2.09% error 2.17% error 3.65% error 1.41% error

Figure 4.3: Comparison of cost-volume smoothing with Gaussian and bilateral filtering. Disparity maps are computed using DP after aggregation. Top row (a)-(c): disparity maps from $\ell \times 1$ support window with Gaussian weights, where (a) $\ell = 1$, (b) $\ell = 5$, and (c) $\ell = 17$, respectively. Disparity (d) is obtained from 35×1 bilateral filtering aggregation. Quantitative error rates in non-occluded regions (bad pixels labeled in black) are given in the bottom row.

Vertical aggregation Global approaches usually use the raw pixel-wise matching costs and skip the aggregation step [18]. In this chapter, we present a novel stereo formulation that combines the strengths of the edge-preserving cost-volume smoothing and the DP optimization framework to achieve high accuracy depth estimation. Motivated by DP's well-known difficulty of enforcing inter-scanline consistency (resulting in horizontal "streaks" in the estimated disparity maps), we enforce vertical smoothness by constructing the data term with an approximated $\ell_y \times \ell_x$ rectangular aggregation window, where $\ell_y \geq \ell_x$ guarantees the dominant aggregation direction is orthogonal to image scanlines.

Figure 4.3.3 illustrates the effects of combining vertical smoothing and DP. With a 5×1 gaussian filter, noise and "streaking" artifacts are somewhat reduced compared to performing DP optimization alone (no aggregation applied). However, pixels near

occlusion boundaries tend to be blurred and thin structures are not very well preserved (lamp bar in Figure 4.3.3 (b)). Similar observation is reported in [15] in which the costs from the previous scanline is aggregated. With a large 17×1 gaussian kernel, the overall disparity map is smoothed at the cost of occlusion boundaries being heavily blurred. In contrast, using a large 1D vertical bilateral filter can preserve sharp occlusion boundaries, suppress noise and enforce scanline consistency in the disparity map.

4.4 Acceleration using Graphics Hardware

To achieve real-time performance, we take advantage of GPU's massively data parallel architectures and implement the matching cost computation and cost-volume smoothing steps on graphics hardware to enhance computational speed of our algorithm.

In the matching cost computation stage, the input stereo images are stored as two textures. For each disparity hypothesis d, we draw a 2D rectangle aligned with two input textures, one of them being shifted horizontally by d pixels. We use the pixel shader, a programmable unit in the graphics hardware [155], to compute the per-pixel absolute difference and the results are written to an output texture. Since the graphics hardware is most efficient at processing four-channel (RGB + alpha) color images, we compute four disparity hypotheses at a time and store the absolute-difference images in different channels. To search over N disparity hypothesis, $\lceil N/4 \rceil$ rendering passes are needed.

Similar to existing real-time stereo GPU implementations [14, 112], the matching

costs obtained are stored as 8-bit integers in GPU memory instead of floating points for lower computational overhead. Representing matching costs with 8 bits makes accurate disparity estimation more challenging since small cost differences cannot be presented due to the limited precision. In our GPU implementation the matching costs in (4.4) are truncated and scaled to make better use of the range of a single byte as

$$\widetilde{C}(p,d) = \min(\frac{\sum_{c\in\{R,G,B\}} |I_p^c - I_{p-d}^c|}{3}, C_{max}) \times \frac{255}{C_{max}}. \tag{4.9}$$

After truncating and scaling, The resultant 3D cost-volume is stored as a stack of 2D images. Four adjacent disparity hypotheses are packed into one color image to utilize the vector processing capacity of GPU. The color images are tiled together to form a large matching cost texture. An example is shown in figure 4.4.

For the cost aggregation step, we first compute the per-pixel adaptive weights for both images. Similar to the cost computation process, we shift the image over itself to compute the pixel-wise weights according to equation (4.1) and store them in textures. The 1D kernel width is always set to a multiple of four to facilitate the four-vector processing capability on GPU. After computing the weights for bilateral filters, we can step through the cost-volume to compute the weighted average. A fairly complex pixel shader program is implemented to index into both the matching cost textures and weighting textures to calculate the final cost. Aggregating over N disparity hypotheses with an approximated $\ell_y \times \ell_x$ bilateral filtering kernel requires $\lceil N \cdot (\ell_y + \ell_x)/16 \rceil$ rendering passes in our implementation.

The advantage of using graphics hardware mainly comes from the parallelism

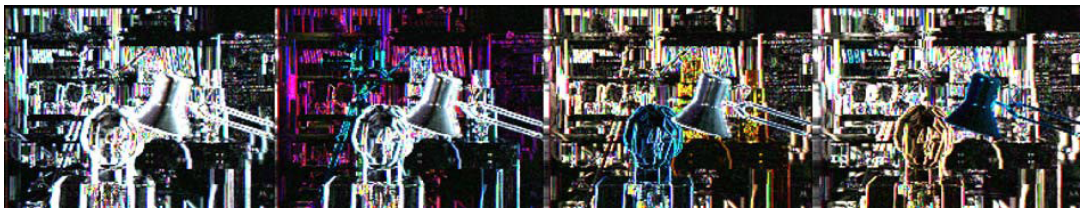


Figure 4.4: The texture used to store matching costs. The four color channels of a single pixel in the texture store the matching costs of a pixel under four different disparity hypotheses.

inherent in today's GPU. The latest generation has up to 24 pixel shader units. Both cost computation and aggregation are regular per-pixel operations that can benefit most from GPU's parallel architecture. The smoothed cost-volume can be used by a WTA selection scheme on GPU (as in [14]), or it can be read back to CPU memory for CPU processing using DP. It should be noted that it is possible to implement the entire DP optimization process on GPU. However, as reported in [112], a GPU-based DP implementation is actually slower than its CPU counterpart. This is mainly due to the significant number of rendering passes needed and the lack of true branching capability on GPU [112]. Therefore we adopted a co-operative approach, using the GPU to compute the cost volume and the CPU to carry out DP. With the new PCI-Express interface between CPU and GPU, the communication bandwidth is huge (full-duplex at 4GB/second), removing a long-existing bottleneck between GPU and CPU.

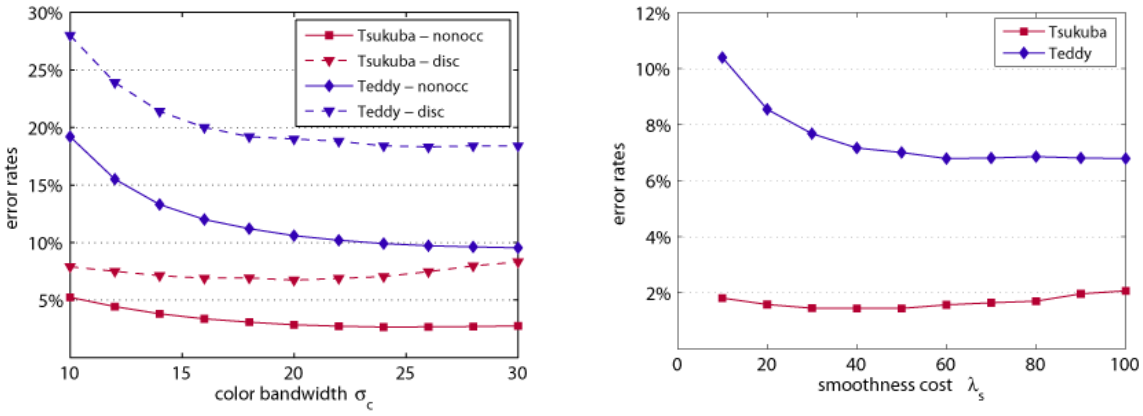
4.5 Experiments

4.5.1 Static Images

The main parameters in our algorithm can be divided into three sets: 1) truncation value $\{C_{max}\}$ for matching cost computation; 2) four parameters $\{\sigma_c, \sigma_g, \ell_x, \ell_y\}$ for cost aggregation; and 3) $\{\sigma_s, \epsilon, \lambda_s, \tau\}$ for disparity selection using DP. Following the experimental observations in [14], C_{max} is set to 25 throughout. Parameters σ_c and σ_g are color and spatial bandwidths for the bilateral filtering, respectively. Figure 4.5 (a) shows the performance of two-pass aggregation for the "Tsukuba" and "Teddy" images as a function of σ_c. In this experiment, we keep the width of the support window and σ_g constant, $\ell_x = \ell_y = 35$, $\sigma_g = 17.5$ (radius of the support window), and use WTA to select the disparities. Note that besides from error rates in non-occluded areas, we also plot error percentages for pixels near depth discontinuities to assess the parameter's edge-preserving performance. In our experiments, we set σ_c to 20 for all test images according to the results learned from this plot.

Among the four DP parameters, σ_s, ϵ and τ are less sensitive and we empirically set $\sigma_s = 400$, $\epsilon = 0.4$ and $\tau = 2$. To determine λ_s, namely the rate of increase in the smoothness cost, we set $\ell_y = 35$, $\ell_x = 1$ and plot the error rates with respect to λ_s in figure 4.5 (b). Note that we use truncated and scaled matching costs in equation (4.9) for these experiments. As can been seen, the optimal λ_s varies for different images. Fortunately, $\lambda_s \in [40, 60]$ typically generates good results. For Middlebury quantitative evaluation we fix $\lambda_s = 60$.

Finally, we evaluate the effects of cost aggregation using windows with different



(a) (b)

Figure 4.5: (a) Error rate with respect to the color bandwidth σ_c for bilateral filtering (equation (4.1)). Statistics in non-occluded regions (nonocc) and areas near depth discontinuity boundaries (disc) are both reported. Disparity maps are generated using "winner-takes-all" and two-pass (35×35) bilateral aggregation; (b) Error rate as a function of the smoothness penalty cost λ_s (equation (4.8)). Disparity maps are generated using DP and vertical (35×1) bilateral aggregation.

sizes. In figure 4.6, we fix kernel height $\ell_y = 35$ and plot the error rates as a function of width ℓ_x. It is worth noticing that since DP performs horizontal optimization, we let $\ell_y \geq \ell_x$ to ensure the dominant aggregation direction is orthogonal to image scanlines. Figure 4.6 suggests that increasing the width of the support window in general tends to marginally improve the accuracy. When $1 < \ell_x \leq 35$, in three of the four data sets approximated bilateral filtering achieves better (or comparable) results compared to the 1D vertical smoothing. And for the "Venus" sequence, the increase in error is mainly caused by the constant parameter setting $\lambda_s = 60$ adopted in our experiments, which is considered to be too large for "Venus". On the other hand, it also reveals the risk of over-smoothing the results when performing both 2D aggregation and global optimization.

Using the online system at [1], we compare our method against other relevant

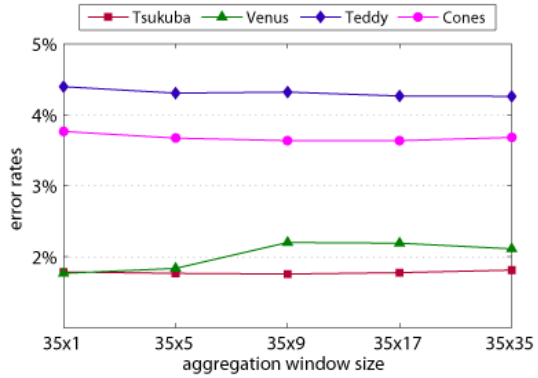


Figure 4.6: Error rate with respect to different aggregation window sizes. Disparity maps are generated using DP.

stereo algorithms listed in the Middlebury evaluation table and summarize the results in table 4.1. With DP optimization, the vertical aggregation window is set to 35×1 for the CPU implementation (VAggCPU+DP) or 32×1 for the GPU counterpart (VAggGPU+DP). For two-pass bilateral aggregation with WTA disparity selection, 35×35 and 32×32 windows are used by CPU (2PassAggCPU) and GPU (2PassAggGPU) implementations, respectively. The average percent of bad pixels in non-occluded regions in the second column is used as the metric by which the table is sorted. Corresponding disparity maps from our approach are shown in figure 4.7. In addition to quantitative error percentages, run time comparisons in MDE/s (last column) are also reported to provide readers with a more clear picture of the compared algorithms. More detailed runtime analysis is given in Section 4.5.2.

The VAggGPU+DP algorithm outperforms other DP-based real-time or near real-time solutions [80,112,114,156] in terms of both matching accuracy and speed. There are two DP-based approaches [115,157] (not listed in table 4.1) that yield better accuracy than ours. However, they both require color segmentation and are typically

slow for real-time applications. In comparison with [116] which performs full-frame optimization via BP, our proposed algorithm can achieve much higher throughput at comparable accuracy. Another near real-time BP-based algorithm [158] relies on color segmentation and plane fitting. Even through with segmentation and BP components implemented on a GPU, it is much slower than our approach. Compared to most edge-preserving filter based local methods [81, 107–109, 159], our proposed algorithm achieves better trade-off between accuracy and efficiency. Our algorithm falls behind a GPU-based local method [110]. Note that [110] refines the final disparity maps by employing advanced post-processing steps such as mutual consistency check [93] (required to compute both left and right disparity maps) and hole filling. For results reported in this chapter, only a 3×3 median filtering is applied to refine the disparity maps. Incorporating effective and efficient disparity refinement step into our existing stereo framework is a future research direction. The approximated 2PassAggGPU approach can produce reasonably accurate disparity maps in real-time. Compared to VAggGPU+DP, although being less accurate, it has the advantage that the computations are completely carried out by the GPU, leaving the CPU free to handle other tasks.

Our GPU implementations (cost aggregation only) attain an average speedup factor of 245 compared to their CPU counterparts, with some sacrifice in accuracy. The principal source of accuracy loss is our choice of GPU precision. Although the pixel shader performs computation in 32-bit floating point numbers, we store the aggregated matching costs in 8-bit textures for lower computational and read-back overhead. Although this precision problem can be addressed by truncating-then-scaling

Table 4.1: Accuracy and speed comparison of related stereo algorithms in the Middlebury online evaluation system [1]. VAggCPU+DP: dynamic programming with CPU-based vertical bilateral aggregation (35×1); VAggGPU+DP: dynamic programming with GPU-based vertical bilateral aggregation (32×1). 2PassAggCPU: two pass CPU-based approximated bilateral aggregation (35×35); 2PassAggGPU: two pass GPU-based approximated bilateral aggregation (32×32).

Algorithm	Non-occ error %				Avg. error %	MDE/s
	Tsukuba	Venus	Teddy	Cones		
CostFilter [110]	1.51	0.20	6.16	2.71	2.65	145.7
PlaneFitBP [158]	0.97	0.17	6.65	4.17	2.99	9.4
VAggCPU+DP	1.57	1.53	6.79	5.53	3.86	2.62
RealtimeBP [116]	1.49	0.77	8.72	4.61	3.90	20.9
FastBilateral [81]	2.38	0.34	9.83	3.10	3.91	0.3
VAggGPU+DP	1.57	1.47	6.93	6.07	4.01	91.7
OptimizedDP [156]	1.97	3.33	6.53	5.17	4.25	19.0
RealtimeABW [107]	1.26	0.33	10.7	4.81	4.28	3.9
RealtimeGPU [80]	2.05	1.92	7.23	6.41	4.40	52.8
2PassAggCPU	1.47	1.40	9.48	5.27	4.41	1.43
ESAW [108]	1.92	1.03	8.48	6.56	4.50	194.8
RealtimeBFV [159]	1.71	0.55	9.90	6.66	4.71	106.9
2PassAggGPU	1.66	1.86	10.3	5.47	4.82	350.1
DCBGrid [109]	5.90	1.35	10.5	5.34	5.77	133.6
ReliabilityDP [112]	1.36	2.35	9.82	12.9	6.61	20.0

the original matching costs obtained, the resulting algorithms are more sensitive to the selection of the truncation value than corresponding CPU implementations. And also note that the stereo parameters are tuned for the CPU implementations and may not be optimal for their GPU counterparts.

4.5.2 Video Sequences of Dynamic Scenes

In addition to performing well on static stereo images, we have applied our method to stereo videos of dynamic scenes. Even though the videos are processed on a frame by frame basis without incorporating temporal smoothness constraints, figure 4.8

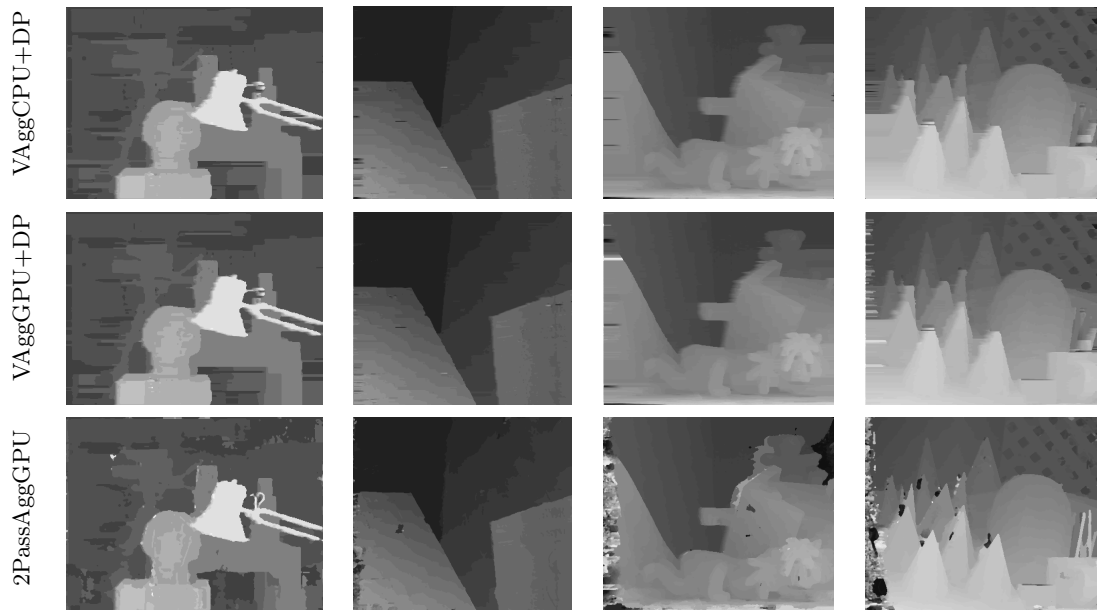


Figure 4.7: Disparity maps for the Middlebury benchmark data generated from our proposed approaches.

shows that combining vertical bilateral aggregation and DP yields more temporally coherent depth estimation than using either edge-preserving cost-volume filtering [3] or DP optimization [4].

We also integrated our algorithm into a stereo system with live video input. The input images are rectified with lens distortion removed. This preprocessing is implemented on the graphics hardware using texture mapping functions. Figure 4.9 shows some live images from our system. Notice the fine structures and clean object boundaries our approach is able to produce. The speed performance with respect to different image resolutions and disparity ranges is summarized in table 4.2. Our GPU-accelerated version is two orders of magnitude faster than its CPU counterpart. It should be noted that our CPU implementation is not yet optimized at the assembly level, which could lead to 2-3 times speedup.

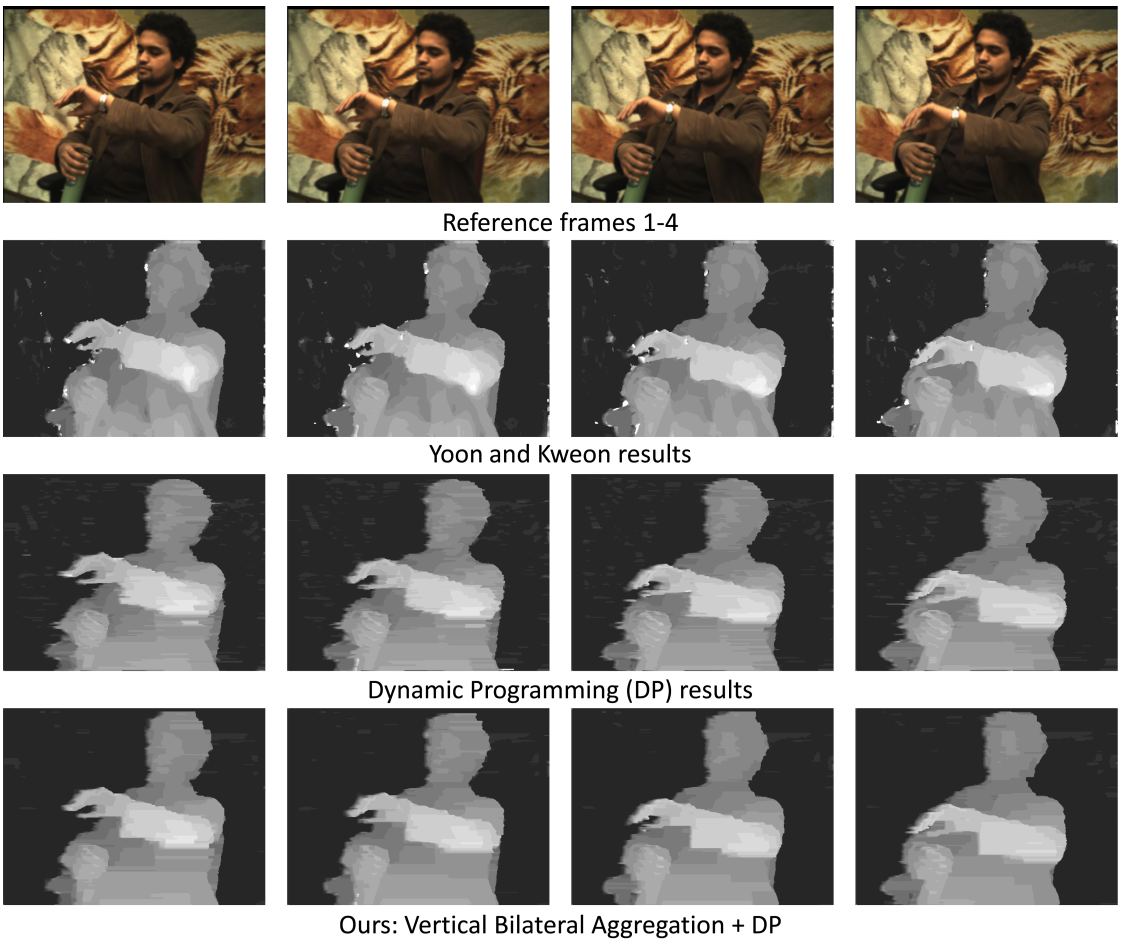


Figure 4.8: Selected disparity maps for a stereo video of dynamic scene (this data set was publicized by [2]). First row: reference images form frames 1-4 of the scene. Second row: results obtained using our implementation of Yoon and Kweon's algorithm [3]. Third row: results from the three-state DP algorithm similar to [4]. Last row: results from vertical bilateral aggregation (32×1) and DP optimization. A 3×3 median filter is applied to refine the disparity maps for all three approaches. Note the improved spatial and temporal consistency from our algorithm.

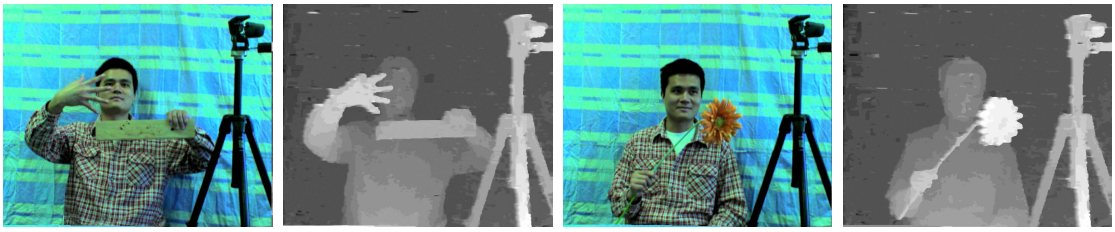


Figure 4.9: Two sample images and their depth maps from our live system on a 2.66GHz PC with a NVIDIA's GeForce GTX 580 graphics card. We can achieve 71 fps with 320×240 input images and 16 disparity levels.

Table 4.2: Real-time Performance. The test system is a 2.66Ghz PC with a GeForce GTX 580 graphics card from NVIDIA.

Image Size	Disp. Range	Runtime MDE/s			
		CPU Only		GPU Accelerated	
		2PassAggCPU	VAggCPU+DP	2PassAggGPU	VAggGPU+DP
320×240	16	1.38	2.59	292.1	87.2
	32	1.56	2.90	326.9	91.2
640×480	16	1.28	2.42	353.9	92.1
	32	1.48	2.55	427.6	96.1

4.6 Summary

In this chapter, we present a stereo framework that operates at real-time while still estimating high quality depth information for live stereo video sequences. Our proposed algorithm combines edge-preserving cost-volume filtering and DP optimization. The use of a color and distance weighted cost aggregation window in the vertical direction reduces DP's "streaking" artifacts. Experimental results have shown that it is among the best performing real-time stereo algorithms in terms of both reconstruction quality and efficiency. In addition, an approximation for the 2D bilateral aggregation is developed, which leads to a fully GPU-accelerated implementation to achieve two orders of speed-up compared to the original approach in [3]. This simplified approach can produce reasonably accurate disparity maps in real-time.

Looking into the future, optimizing DP using MMX (as in [15]) is likely to further improve the speed performance. We would also like to investigate the precision issue on the graphics hardware. Current graphics hardware does provide limited support for high-precision texture maps, at the cost of performance degradation (the hardware is optimized to work with 8-bit textures). From an algorithmic standpoint, our DP implementation enforces the ordering constraint for speed consideration. Thin 3D

structures (such as the flower stem in Figure 4.9) may disappear if it is far away from the background. We plan to investigate the use of scanline optimization [18], which enforces the smoothness constraint directly without employing the ordering constraint. Another interesting venue to explore is to enforce the temporal consistency in the video to reduce the flickering artifacts.

Chapter 5 Global Stereo Matching Leveraged by Sparse Ground Control Points

In this chapter, we present a novel global stereo model that makes use of constraints from points with known depths. The rest of this chapter is organized as follows: Section 5.1 introduces our basic stereo formulation. In Section 5.2, we present our regularization prior and explain how to compute the prior likelihood given known control points via an adaptive propagation algorithm. We valid our stereo model with experiments in Section 5.3. Section 5.4 presents conclusions and planned future work.

The main contribution of this chapter lies in the use of a new regularization prior for global stereo matching. We make an assumption that there exists a sparse set of scene points whose 3D positions are given and propose a Markov Random Field (MRF) stereo formulation that incorporates priors from such Ground Control Points (GCPs) [4]. Our motivation comes from the observation that the scene depth field is piecewise smooth and even a small amount of GCPs can encode rich information on scene structure. Under this view, we model stereo matching as a maximum a posterior MRF (MAP-MRF) problem. The GCPs-based constraints are naturally integrated into an MRF as *soft* constraints using the Bayes rule. Although the concept of GCPs has been introduced in early stereo methods, to the best of our knowledge, the use of punctual depth priors has never been explored in global stereo frameworks. Quantitative evaluations with ground truth data demonstrate the effectiveness of

using sparse GCPs to leverage the ill-posed stereo matching problem. Experimental results show that our formulation clearly improves upon existing methods on the Middlebury benchmark data set.

5.1 Problem Formulation

In this section, we present the stereo formulation proposed in this chapter. Note that for notation clarity, our derivation will focus on rectified two-frame stereo. However, it is relatively easy to generalize our method to handle multi-view stereo, for instance, computing matching costs via plane-sweep based approach [160], as later shown in Section 5.3.2.

Given a stereo image pair $I = \{I_L, I_R\}$, where I_L, I_R are the left and right images, the goal of stereo matching is to compute the dense disparity map D of one reference view, say I_L. In our stereo model, in addition to the input images, we assume there exists a sparse set of GCPs, denoted G, on the reference view whose disparities are known with high confidence. For each pixel $p \in I_L$, $p \in G$ implies p is a control point and we use D_p and G_p to denote p's disparity value from D and G, respectively.

We formulate our stereo model as a MAP-MRF problem. We assume the GCP acquisition is independent of the image formation process of the stereo pair I. Under this assumption and using the Bayes' rule, the posterior probability over D given I and G can be written as $P(D|I,G) \propto P(I|D)P(G|D)P(D)$. As maximizing this posterior is equivalent to minimize its negative log likelihood, our objective is to find a disparity map D that minimizes a global energy function

$$
\begin{aligned}
E(D) &= -\ln(P(I|D)) - \ln(P(D)) - \ln(P(G|D)) \\
&= E_{data}(D) + E_{smooth}(D) + E_{gcp}(D).
\end{aligned}
\tag{5.1}
$$

The first term, the data energy E_{data}, comes from the negative log likelihood of the probability of disparity assignment given the observed image pair, whereas the second term, the smoothness energy E_{smooth}, encourages neighboring pixels to have similar disparities based on the assumption that the scene is locally smooth. The last term E_{gcp}, which we refer to as the GCP energy, encodes the constraints from sparse GCPs. In our stereo formulation, $E_{data}(D)+E_{smooth}(D)$ is equivalent to the standard cost function used by existing global stereo methods [18]. The GCP energy, which plays a regularization role, is the key contribution of this work.

5.1.1 Basic Stereo Model

In the MRF stereo framework, the data energy comes from the negative log likelihood of the matching costs and measures how well the disparity map D agrees with the input images. We define the data term as the sum of per-pixel color difference as

$$
E_{data}(D) = \sum_{p \in I_L} \Phi(p, D_p), \tag{5.2}
$$

where $\Phi(,)$ is a pixel-wise color dissimilarity function between corresponding pixels given certain disparity value. For rectified two-frame stereo, we use the sampling insensitive calculation of Birchfield and Tomasi [71] for increased robustness to image sampling.

The smoothness energy, under the MRF-based formulation, comes from the negative log likelihood of the smoothness-based prior. In this chapter, we assume that pixels form a 2D grid and employ the widely used truncated linear model defined upon a standard 4-connected neighborhood system N_4 as

$$E_{smooth}(D) = \lambda_s \sum_{p \in I_L} \sum_{q \in N_4(p)} w_{pq} \cdot \min(\Delta d_{pq}, T), \tag{5.3}$$

where $\Delta d_{pq} = |D_p - D_q|$ is the disparity difference between pixels p and q. λ_s is the rate of increase in the discontinuity cost and T controls the limit of the cost. The spatially varying per-pairing weights $\{w_{pq}\}$ are computed based on the color differences between neighboring pixels on the reference view as

$$w_{pq} = \max(\exp(\frac{-\Delta c_{pq}}{\gamma_c}), \epsilon). \tag{5.4}$$

where Δc_{pq} is the Euclidean distance between pixels in the RGB color space. Parameters γ_c and ϵ control the sharpness and lower bound of the exponential function, respectively.

5.2 Regularization using GCPs

The energy terms in Section 5.1.1 forms the basis of many standard MRF stereo models [86, 117, 118], just to name a few. Although the global formulation can substantially improve the reconstruction quality over local correlation-based methods, the capability of standard MRF stereo model is still limited. State-of-the-art stereo models require additional regularization terms, such as the segmentation-based con-

straints employed in [12]. What differentiates our formulation from existing MRF stereo models is the leverage of prior knowledge about the scene structure encoded in the GCP energy E_{gcp}. In this section, we provide detailed descriptions of the GCP regularization term proposed in this work.

5.2.1 Adaptive Propagation via Optimization

In order to model the likelihood $P(G|D)$, our basic idea is to predict the disparity values for non-GCP pixels from sparse control points. In other words, the objective is to interpolate a dense disparity map from the GCP set G. Without prior assumption this problem is clearly ill-posed. Inspired by the fact that the scene depth field is always piecewise smooth, we present an adaptive disparity propagation algorithm that is built upon a premise that neighboring pixels with similar color should have similar disparities. Our adaptive propagation algorithm is given as input the reference image together with the GCP set G and automatically propagate the disparity values of GCPs to the rest pixels whose disparity values serve as the unknown.

We impose the constraint that two neighboring pixels p and q should have similar disparity values if their colors are similar by trying to minimize the difference between the disparity of pixel p and the weighted average of the disparities at p's neighbors. A global cost function can be defined as

$$\begin{aligned} J(\widetilde{D}) &= \sum_{p \in I_L} (\widetilde{D}_p - \frac{\sum_{q \in N_8(p)} w_{pq} \widetilde{D}_q}{\sum_{q \in N_8(p)} w_{pq}})^2 \\ &= \sum_{p \in I_L} (\widetilde{D}_p - \sum_{q \in N_8(p)} \alpha_{pq} \widetilde{D}_q)^2, \end{aligned} \tag{5.5}$$

where N_8 is the 8-connected neighborhood system and the function w_{pq} as defined in

equation (5.4) correlates pixels based on their color similarities. α_{pq} is the pairwise weighting function that sums to one. Note that similar weighting functions have been previously employed in [161, 162], where they are usually referred to as affinity functions.

Note that if we consider $\widetilde{D}$ as an one dimensional vector, the quadratic form $J(\widetilde{D}) = \widetilde{D}^T(L - W)\widetilde{D}$ is exactly the cost function we wish to minimize. Here W is a $N \times N$ (N is the total number of pixels) matrix whose elements are the pairwise affinities and L is a diagonal matrix whose diagonal elements are the sum of the affinities $\{\alpha_{pq}\}$. In our case L is simply an identity matrix.

Given that the cost function is quadratic, we minimize $J(\widetilde{D})$ by solving $\nabla J(\widetilde{D}) = 0$, which leads to the unconstrained system $(L - W)\widetilde{D} = 0$ (any constant vector $\widetilde{D}$ is a trivial solution). Now given a set of GCPs with known disparities G, we minimize $J(\widetilde{D})$ subject to these additional constraints. The optimal $\widetilde{D}$ can be efficiently computed by solving a system of sparse linear equations $Ax = b$. Here A is a $N \times N$ sparse matrix with diagonal entries equal to one, and $A(p, q) = (L-W)(p, q)$ if $p \notin G$ and $A(p, q) = 0$ otherwise. Likewise, $b_p = G_p$ if $p \in G$ and $b_p = 0$ otherwise. In this work we solve the sparse linear systems using the UMFPACK library [163].

5.2.2 Likelihood from Disparity Propagation

After the optimal $\widetilde{D}$ is computed, we model the likelihood $P(G|D)$ as

$$P(G|D) \propto \prod_{p \in I_L} \exp(-\Psi(D_p, \widetilde{D}_p)). \tag{5.6}$$

where function $\Psi(D_p, \widetilde{D}_p)$ penalizes disparity assignment that diverges from the interpolated disparities. In this work, our robust penalty function $\Psi(x, y)$ is derived from the Total Variance model [164] as

$$\Psi(x, y) = -\ln((1-\eta)\exp(\frac{-|x-y|}{\gamma_d}) + \eta). \tag{5.7}$$

Parameters γ_d and η, respectively, control the sharpness and upper-bound of the robust function. The GCP regularization term, E_{gcp}, is then modeled as

$$\begin{aligned} E_{gcp}(D) &= -\ln(P(G|D)) \\ &\propto \lambda_r \sum_{p \in I_L} \Psi(D_p, \widetilde{D}_p) \end{aligned} \tag{5.8}$$

where λ_r is a regularization coefficient that controls the strength of the GCP energy.

After modeling the GCP regularization term, optimal disparity assignment that minimizes equation (5.1) can be obtained using existing energy minimization techniques surveyed in [85]. In this work, we use graph cuts [165] method to compute the dense disparity map D.

5.3 Experimental Results

We have evaluated our stereo framework on different benchmark stereo images with known ground truth data [5, 6, 95], and we show that our formulation quantitatively improves the reconstruction accuracy. In the experiments we demonstrate that GCPs used in our algorithm can be obtained in different ways. We first show that by using consistency check among several local stereo algorithms, GCPs can be reliably

Table 5.1: GCP densities and outlier percentage for the Middlebury stereo data. Outlier (%) is the percentage of GCPs whose absolute disparity error is larger than 1 pixel.

	Tsukuba	Venus	Teddy	Cones	Avg.
density (%)	19.1	15.8	12.1	9.34	14.1
outlier (%)	0.30	0.36	1.00	1.31	0.74

extracted from the stereo images themselves without resorting to additional sensors. In addition, when sparse GCPs are provided from external sensors, our formulation is capable of incorporating them to improve passive stereo vision.

5.3.1 Improving Passive Stereo: Computing GCPs from Stereo Images

For many vision applications people intend to recover scene depth based solely on images without relying on exterior sensors or structured light patterns. In order to obtain GCPs in this scenario, a straightforward way is to extract GCPs from the stereo images via feature matching.

In this chapter, we adopt a simple approach to obtain sparse GCPs without resorting to complicate algorithms. Our method is based on a voting strategy and simply requires a few disparity maps from local methods. In detail, for the reference image I_L we compute three disparity maps via WTA. These disparity maps are: 1) D_{BT} from the *Birchfield and Tomasi* matching costs without aggregation; 2) D_{NCC} from *normalized cross correlation* using a 5×5 image patch; and 3) D_{AW}, which is computed using the *adaptive weight aggregation* method [3]. The window size for D_{AW} is set to 39×39. A pixel $p \in I_L$ is selected as a GCP candidate if p's WTA disparities in different disparity maps are consistent (variance smaller than 1 dispar-

ity) and p is not near any intensity edge (edges detected by Canny edge detector). To further remove matching outliers, we apply the same procedure to compute a set of GCP candidates for I_R. Finally, GCP candidates that survive the left-right consistency check [93] are retained as the GCPs. We demonstrate in Figure 5.1 the resultant GCPs for the Middlebury benchmark data. The GCP densities and outlier percentages are provided in Table 5.1. As can be seen, the GCPs obtained from our method are fairly sparse and contain very few outliers.

We experimentally validate the effectiveness of our algorithm using reliably matched pixels as GCPs. Before reporting our results, we first present parameter settings used in our experiments. The threshold ϵ in equation (5.4) is set to 0.3 to prevent the weight from being too small. When computing w_{pq} in equation (5.5), we instead set $\epsilon = 0$ for edge-preserving interpolation. The associated color bandwidths γ_c is set to 3.6 for equation (5.4) and 1.25 for equation (5.5), respectively. The parameters that control the shape of the robust function (5.7) are chosen as $\gamma_d = 2$ and $\eta = 0.005$. The truncation parameter T used for the discontinuity penalty is set to 2. The regularization coefficients λ_s and λ_r are set to 20 and 8 throughout this experiment.

To evaluate the performance of our approach, we follow the methodology proposed by Scharstein and Szeliski in [18]. The three disparity maps included in our first comparison are: 1) D^*, which is the disparity map computed by minimizing the energy term $E_{data} + E_{smooth}$, without considering the GCP energy; 2) $\widetilde{D}$, the interpolated disparity map using the adaptive propagation method as described in Section 5.2.1; and 3) the result from our formulation, i.e., the disparity map D that minimizes the cost function (5.1). Table 5.2 summarizes the percentages of error disparities (where

the absolute disparity error is greater than 1 pixel). The error statistics accounts for three pixel categories, classified as non-occluded (nonocc), near discontinuous (disc), and the entire image (all). Note that we use constant parameters as reported for the four evaluation stereo pairs. Associated disparity maps are demonstrated in Figure 5.1.

As can be seen, disparity maps from our formulation outperform others for almost all categories. In comparison with D^*, the reconstruction accuracy has been significantly increased in D. The fact that our cost function (5.1) is built upon the standard global stereo model with the GCP regularization term incorporated suggests that the performance gain comes from the GCP constraints/priors. Our evaluation also shows that even through the GCPs are sparse, the interpolated disparities $\widetilde{D}$ from our disparity propagation scheme are quite satisfactory in general. For "Venus", "Teddy", and "Cones" sequences the results are better than the standard graph cuts stereo. Although $\widetilde{D}$ is less accurate for "Tsukuba", by formulating E_{gcp} as a soft constraint, the additional GCP energy plays an effective role for regularizing stereo matching for all data set.

We have also compared our results with those produced by competitive stereo algorithms. Among the five compared methods, "Klaus" [12] is the current top performer in the Middlebury evaluation table [1]; "Woodford" [88] and "Smith" [2] are recently invented algorithms that also use novel regularization priors; "SymBP" [122] is one of the state-of-the-art algorithms that uses segmentation-based constraint as regularization prior. "GC" [117] is the pioneer in the use of energy minimization framework and graph cuts to solve stereo correspondence. The energy function defined in [117]

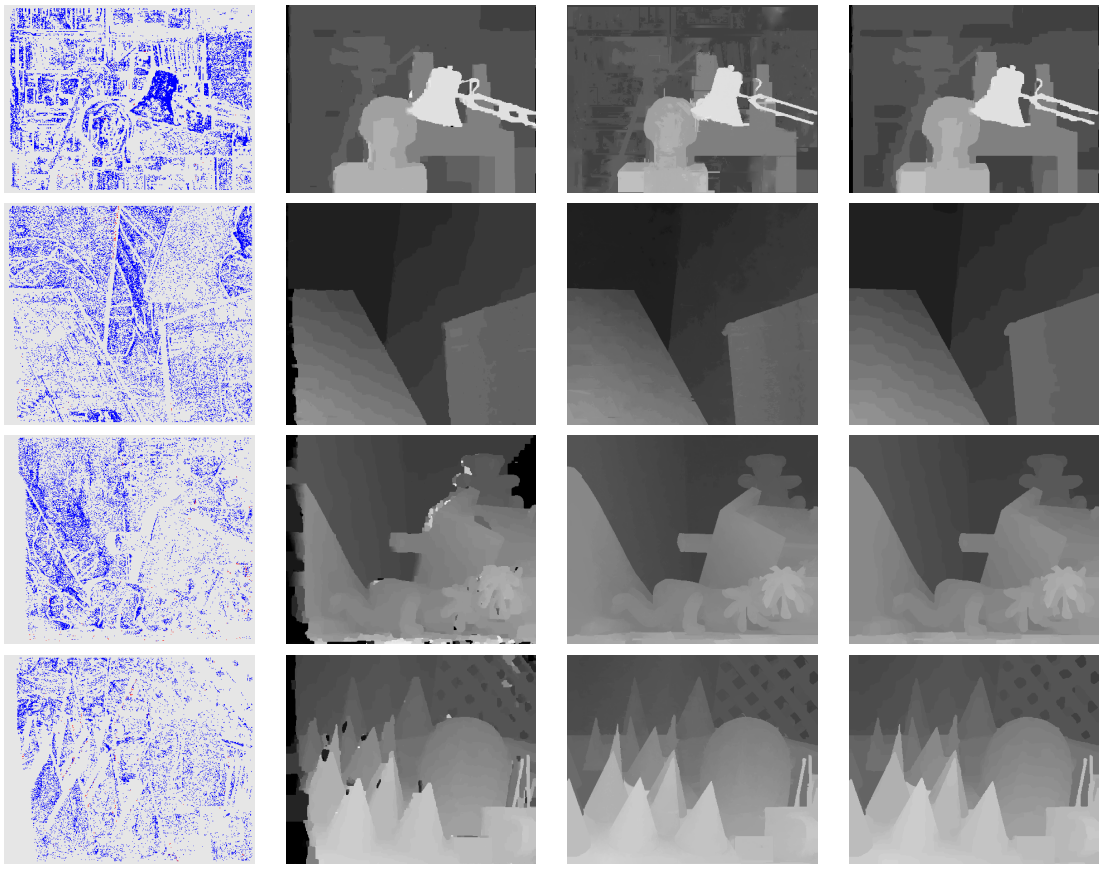
Table 5.2: Comparison of the results on the Middlebury data sets.

	Tsukuba			Venus			Teddy			Cones		
	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc
D^*	1.71	3.87	9.02	0.97	2.57	12.0	11.4	20.5	23.0	5.51	15.9	13.0
$\widetilde{D}$	1.92	2.41	9.71	0.46	0.74	3.99	6.58	11.7	16.4	4.87	10.5	12.0
$\boldsymbol{D}$	0.87	2.54	4.69	0.16	0.53	2.22	6.44	11.5	16.2	3.59	9.49	8.95

is similar to our basic stereo model defined in Section 5.1.1. Quantitative error percentages in non-occluded areas for the four evaluation images are shown in Table 5.3. As can be seen, results from our algorithm are comparable to other state-of-the-art results. As of April, 2011, our method ranked 8th out of more than 100 published stereo algorithms listed by the Middlebury evaluation table. It is also worth noting that among the top ten approaches, only ours does not rely on color segmentation.

In addition to the standard data sets, we have applied our algorithm on a scene that contains highly curved surfaces. As shown in Figures 5.2 and 5.3, our method performs equally well on high-curvature areas, in which [122]'s segmentation prior and [88]'s second order smoothness prior show their limit.

In terms of run time, for the "Teddy" sequence (450×375 with 60 disparity levels) on a 2.83GHz dual core CPU, detecting GCPs takes about 1.6 minutes (parallel processing for left/right images using multi-thread technique); solving the system for the interpolation takes about 10 seconds and the graph-cut optimization takes 24 seconds. The total runtime for "Teddy" is about 130 seconds. The most time consuming part is the adaptive weight aggregation [3]. However that step can be implemented on GPU for better speed performance and much faster CPU implementations are also available.



$(GCPs)$ (D^*) $(\widetilde{D})$ (D)

Figure 5.1: Our results for Middlebury benchmark data. The first column shows GCPs. Inliers and outliers are shown in blue and red, respectively. D^* is from minimizing $E_{data} + E_{smooth}$ without incorporating the regularization term E_{gcp}; $\widetilde{D}$ is the disparity map from disparity propagation as defined in Section 5.2.1; Our resultant disparity maps D are shown in the last column.

Table 5.3: Middlebury evaluation of our results compared with those produced by competitive stereo algorithms. The numbers are the percentage of error disparities in *non-occluded* areas.

	Tsukuba	Venus	Teddy	Cones	Avg. Error	Rank
Klaus [12]	1.11	0.10	4.22	2.48	1.98	1
Woodford [88]	2.91	0.24	10.9	5.42	4.88	59
Smith [2]	1.12	2.23	7.25	4.46	3.77	20
GC [117]	1.19	1.64	11.2	5.36	4.84	58
SymBP [122]	0.97	0.16	6.47	4.79	3.09	12
Ours	0.87	0.16	6.44	3.59	2.77	8

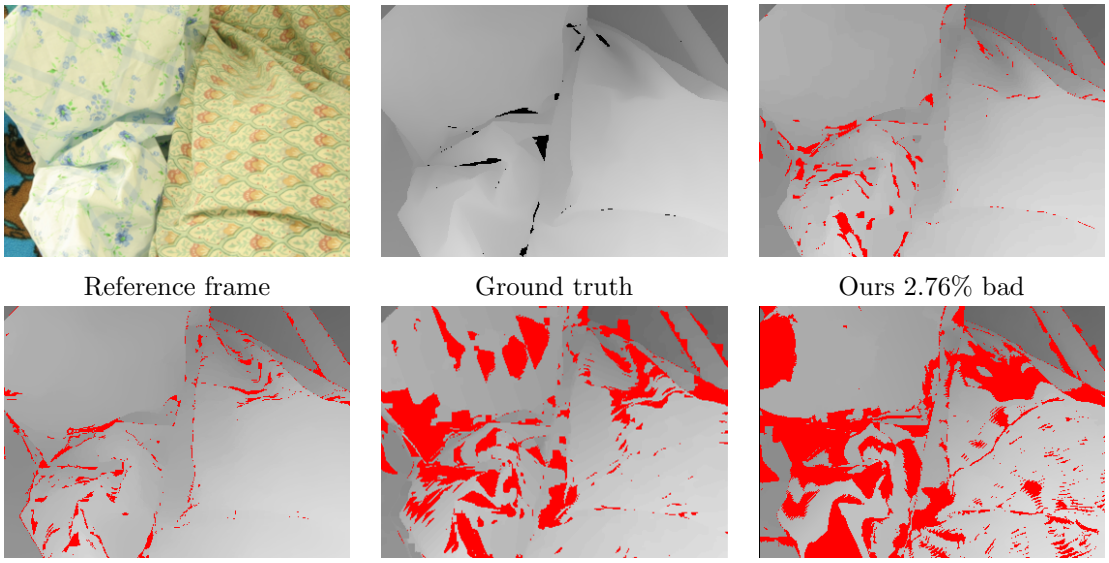
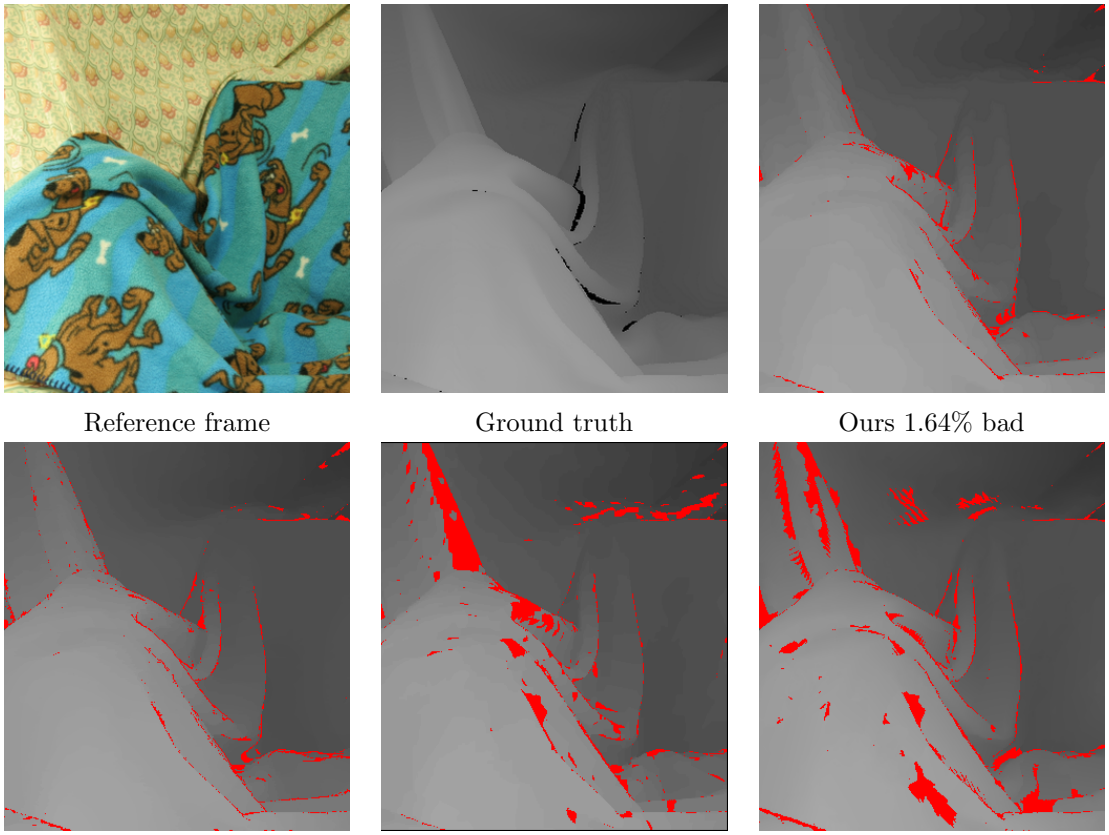


Figure 5.2: Results demonstrating the effectiveness of our method on the Middlebury "Cloth2" data set [5] with curved surfaces. Red pixels are bad disparities in non-occluded areas.

5.3.2 Active and Passive Sensing Fusion: Incorporating GCPs from Laser Scanning

In this experiment we investigate our algorithm using GCPs obtained from laser range scanning. While active range sensing techniques have been widely used in construction, survey, and military, the sampling density of these devices is limited since they usually take only one depth measurement at one time. This limitation is particularly problematic on mobile sensing platform. For example, typical lateral resolution from airborne LiDAR system is about 1 point per meter while a digital imagery can easily achieve 10 even 100 points per meter from the same height. In order to produce high resolution scans, multiple scans are required to handle missing data and occlusion [6], which is both laborious and time consuming. Alternatively, 2D image acquisition is a flexible, efficient, and inexpensive operation. In this experiment, we



Reference frame | Ground truth | Ours 1.64% bad

Smith [2] 2.01% bad | SymBP [122] 4.56% bad | Woodford [88] 6.33% bad

Figure 5.3: Results demonstrating the effectiveness of our method on the Middlebury "Cloth3" data set [5] with curved surfaces. Red pixels are bad disparities in non-occluded areas.

seek to incorporate low resolution LiDAR data into our stereo framework to improve the reconstruction accuracy.

For quantitative evaluation on ground truth data, we employ the "Fountain-P11" multi-view stereo sequence publicized by Strecha et al. [6] as our test data. The range data is acquired from a time-of-flight laser scanner. In addition to the 3D point clouds, there are 11 high resolution color images whose camera parameters are provided in the coordinate system defined by the LiDAR system. In order to simulate sparse range scans, we downsample the original point clouds with a scale factor 16 using the nearest-neighbor interpolation and treat resultant sparse 3D points as GCPs for our

stereo matching algorithm. This implies that only about 1/256 ($\approx 0.4\%$) 3D points from the original dense scan are preserved.

We select the center frame (frame 5) as our reference view and use 6 of its neighboring frames to compute the matching costs (the Euclidean norm of color differences in RGB space) using the standard multi-view plane-sweep approach [160]. The color images are downsampled to $1,536 \times 1,024$, which is half of their original size. Given the scene depth range, we equally quantize the depth space into 360 levels. Due to memory and speed consideration, we divide the reference image into 4 rectangular tiles (784×528). The width of the overlapping areas between two tiles is 16 pixels. We perform graph cuts optimization for each tile independently and merge the 4 depth maps to form a high resolution depth map. For pixels within overlapping areas their depths are set to the mean of multiple measurements to preserve the global smoothness. As shown in Figure 5.4, the depth map (c) produced from our formulation compares favorably to the one from basic MRF stereo model, i.e., using $E_{data} + E_{smooth}$ as cost function. The depth map estimated using GCPs priors preserves both fine structure details and sharp discontinuities near object boundaries. Quantitative evaluation for this data set can be performed via the online system maintained by Strecha et al. [6], given a triangle mesh. As depth maps fusion and model building are not the focuses of this work, instead of fusing multiple depth maps to form a complete 3D model, we simply construct a triangle mesh from frame 5's depth map and upload it to the evaluation system to obtain the error histograms. *Sigma* in Figure 5.5 denotes the standard deviation of the depth returned by the laser range scanner. As can be seen, without using our GCP priors, 34.7% of the depth

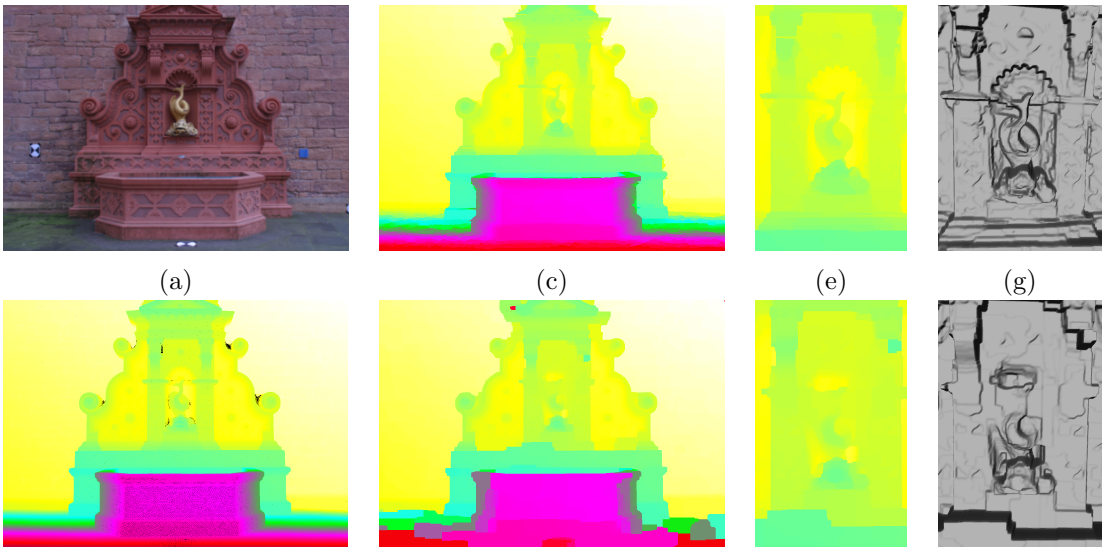


Figure 5.4: Results for Fountain-P11 data set [6]. (a) the reference view. (b) ground truth depth map from LiDAR data, black pixels are missing data. (c)-(d) depth maps computed with and without the GCP energy, respectively. (e)-(h) zoomed in views of depth maps and associated mesh rendered in 3D. Notice the fine details preserved by our algorithm in (e) and (g). This figure is best viewed in color.

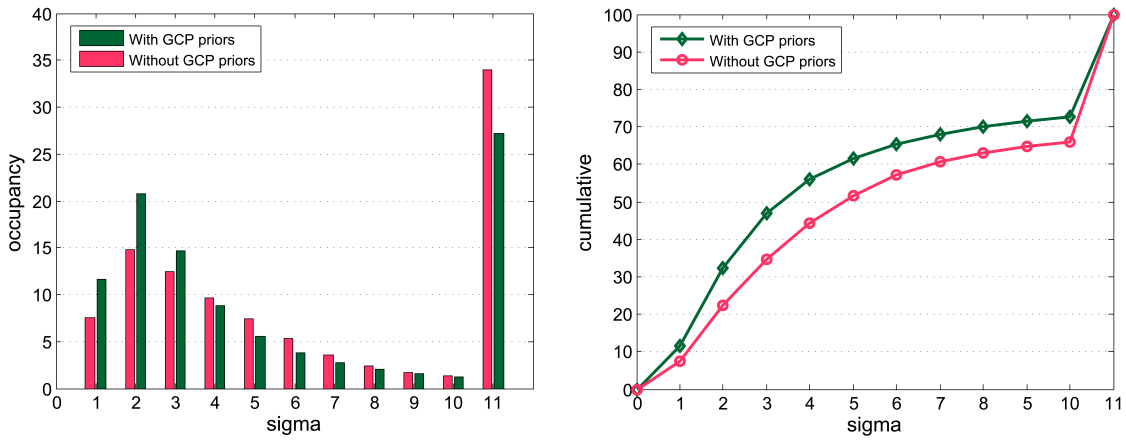


Figure 5.5: Error histograms for depth maps (c) and (d) shown in Figure 5.4.

estimates are within the $3 \times Sigma$ range of the ground truth data. Alternatively, corresponding accuracy numbers have been quantitatively improved to 47% using our algorithm. Note that for both methods we have applied moderate parameter tuning to enable a fair comparison.

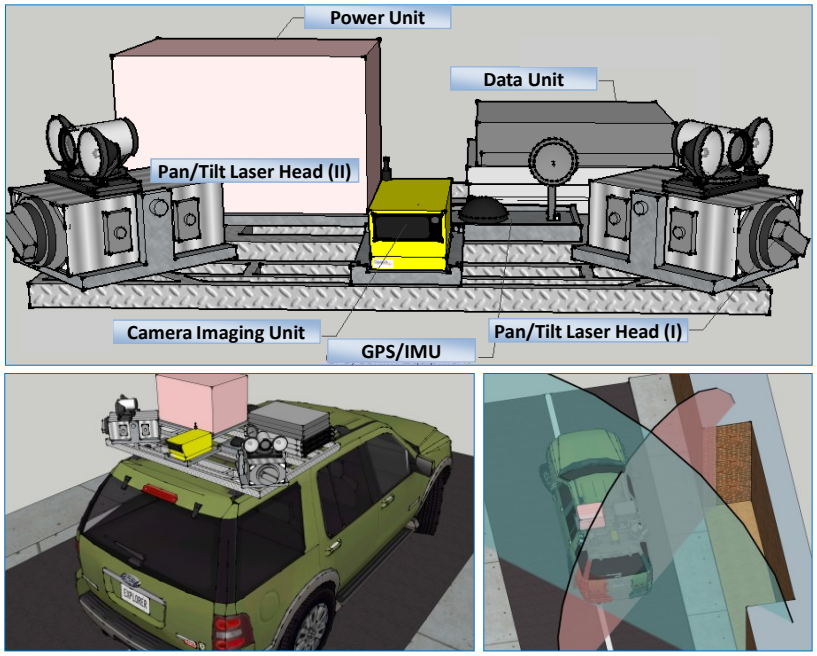


Figure 5.6: The conceptual sketch of our mobile scanning unit. It can be vehicle mounted for continuous mobile scanning. The two semi-transparent circles show the trajectory of the scanning path.

In addition to quantitative evaluation using ground truth, we also apply our formulation to active and passive sensing fusion for 3D reconstruction of urban environment. In particular, the goal of this experiment is to fuse LiDAR data from an active mobile scanner with image data from a passive video camera to generate high-quality depth maps for 3D modeling purpose. As illustrated in Figure 5.6, our system contains the following units: GPS and inertial measurement unit (IMU), laser range sensor heads, camera imaging units, an integrated power unit, a storage unit, and a control laptop (not shown in Figure 5.6). The system is designed to be portable and it can be mounted on a vehicle. As the vehicle moves, the two-side mounted laser sensors scan in a 360 degree circle. The reason we use two heads is that it can provide more coverage in a single trip by scanning two helix trajectories. A panoramic camera

Figure 5.7: The prototype scanner system. The lower images show the panoramic camera and one of the GPS receivers and the two laser scanners.

is used to allow continuously record color imagery of the scene. All the sensors share the same control and data storage unit, as well as the GPS/IMU signal so that all the data can be geo-referenced. Figure 5.7 demonstrates our system prototype.

In order to perform sensing fusion, we first estimate the intrinsic camera model (focal length, imaging center, and distortion coefficients) using the methods presented in [65]. Non-linear lens distortion is removed after intrinsic calibration. At run-time, the scene is simultaneously scanned by the LiDAR scanner and captured by the video camera. The laser sensor calibration and GPS/IMU initialization are provided by the hardware venders. After data collection, an off-line extrinsic calibration step is performed to recover the absolute camera pose of each image under the LiDAR coordinate system. We develop an interactive camera pose estimation system to

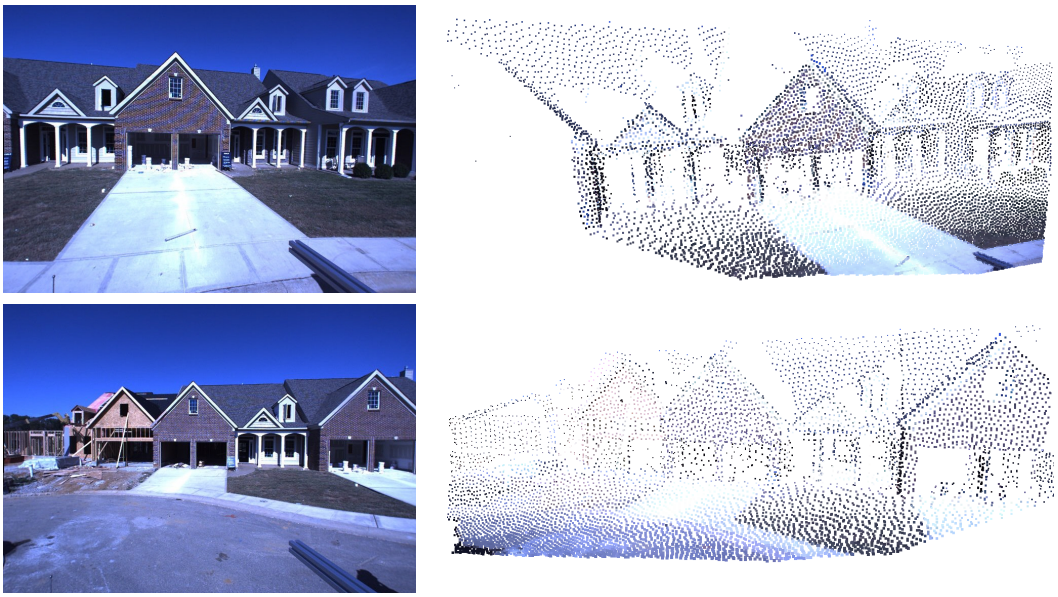


Figure 5.8: Left column: example video frames captured by the passive video camera. Right column: corresponding sparse 3D point clouds (GCPs) returned by the laser scanner.

fulfill the task. The general idea is to firstly recover the camera pose of a reference frame using a set of manually specified 2D and 3D correspondences[1] [166]. Camera poses of the rest frames can be estimated using the reference camera pose together with the IMU data [8].

In Figure 5.8 we demonstrate two color images captured by our video camera together with the corresponding 3D point clouds returned by the laser range sensor. The 3D points are colorized after estimating the camera poses in the LiDAR coordinate frame. In Figure 5.9 we provide qualitatively comparison of the depth estimation results from passive stereo and passive-active sensing fusion. In order to better highlight the differences, 3D mesh models are shown instead of depth maps. As can been seen, results leveraged by the sparse LiDAR data outperforms passive

[1] 2D points are distinct image features such as corners, and the 3D points are the corresponding 3D structures whose positions $\{X, Y, Z\}$ are defined under the LiDAR coordinate frame.

stereo algorithms, especially for textureless areas and regions that contain fine structures. Depth maps and screenshots of colorized 3D dense point clouds after sensing fusion are shown in Figure 5.10. Before sensing fusion, there are in total 20269 LiDAR points in the scenes shown in Figure 5.8 and after fusion the total number of 3D points raised to 569601, which is about 28 times denser than the raw measurements returned by the laser sensor.

5.4 Summary

In this chapter we present a global stereo matching framework that utilizes a sparse set of points with highly reliable depths, i.e., the ground control points (GCPs). While the concept of GCP has been introduced in early stereo literature, to the best of our knowledge, it is the first time that it is incorporated in a full frame global optimization framework. Using the Bayes rule, GCPs are included in an MRF in a principled way. Our generic formulation allows GCPs to be obtained from various modalities. In this chapter, we explore two interesting scenarios where 1) GCPs are obtained from stereo images themselves via stable matching; and 2) GCPs are provided from sparse laser range scans by exterior sensor. By evaluating our method with ground truth data, we demonstrate the effectiveness of our algorithm on an extensive set of experimental results.

Looking towards the future, we will continue our research on using stereo matching to enhance the resolution of laser range data from mobile scanning, which usually has a much lower resolution than image data. This chapter is primarily of interest to stereo, however, our formulation could potentially be applied to other MRF-

passive stereo

passive and active sensing fusion

Figure 5.9: 3D models of the scenes shown in Figure 5.8. For the top two models, depth maps are computed using the standard stereo model. In comparison, the bottom two are from our proposed sensing fusion framework.

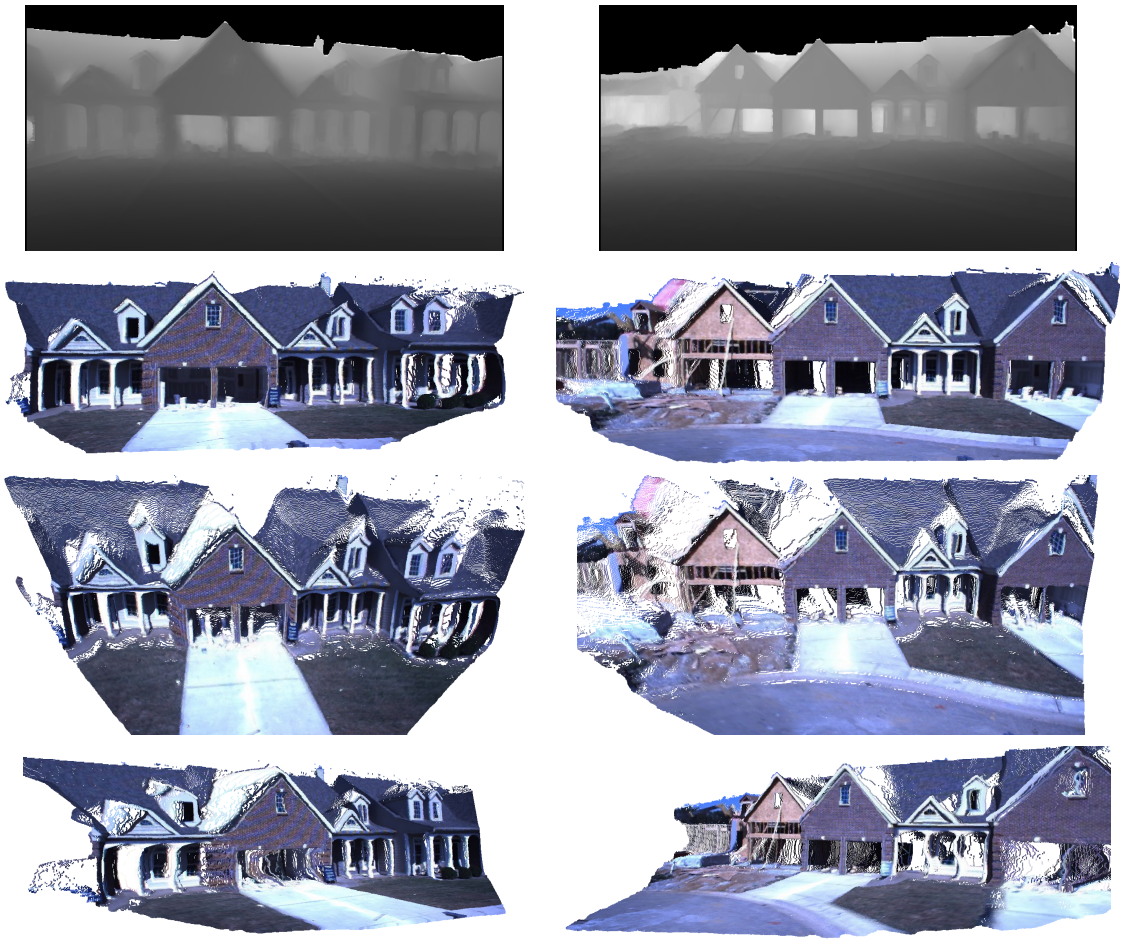


Figure 5.10: Dense depth maps and 3D point clouds of the scenes shown in Figure 5.8.

formulated labeling problems in computer vision [85].

Chapter 6 BRDF Invariant Stereo using Light Transport Constancy

In this chapter, we introduce *light transport constancy* (LTC) as a constraint on stereo matching. LTC simply asserts that the percentage of light reflected by a particular surface patch (the BRDF) remains constant for a given viewing direction. This constraint has not been previously exploited and allows stereo correspondence to be correctly determined for surfaces with an arbitrarily complex BRDF and does not require calibrated light sources or objects.

As an intuitive introduction to this constraint, consider the scene configuration in Figure 6.1. The scene is illuminated by a single point light source, L. A particular point in the scene, x_i, will reflect light to each of cameras C_1 and C_2 according to:

$$E_{C_j}(x_i) = L(x_i)R(x_i, \theta_L, \theta_{C_j}) \tag{6.1}$$

where $E_{C_j}(x_i)$ is the radiance in the direction of C_j from the point x_i, $L(x_i)$ is the observed irradiance of point x_i, and $R(x_i, \theta_L, \theta_{C_j})$ is the BRDF at point x_i, indexed by the vectors in the direction of L and C_j. Throughout the text, direction vectors are written as single variables for notational simplicity (e.g. θ_L, θ_{C_j}) despite the fact they represent 2D quantities. Also for the sake of simplicity, we do not include the dependency of wavelength in this exemplary scenario.

The traditional Lambertian assumption is that the reflectance (BRDF) is equal in the directions of C_1 and C_2, i.e.,

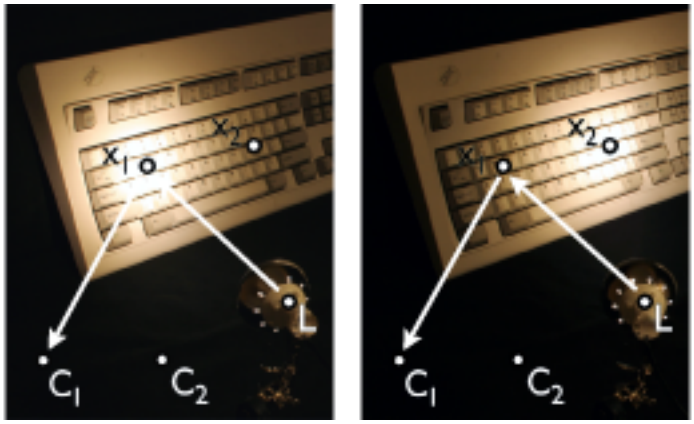


Figure 6.1: (Left) The BRDF at x_1 determines the percentage of light reflected from light source L toward each of cameras C_1 and C_2. (Right) The spatial position of all components is the same, but the light distribution has been altered by rotating the light about its light bulb (i.e., steering the light beam to a different place). Although the incident intensity at x_1 has changed, the percentage of light reflected remains constant.

$$R(x_i, \theta_L, \theta_{C_1}) = R(x_i, \theta_L, \theta_{C_2}) \tag{6.2}$$

Thus we legitimately have $E_{C_1}(x_i) = E_{C_2}(x_i)$. However, this relation will not in general hold true for arbitrary BRDFs.

Light transport constancy assumes that the surface BRDF, $R(x_i, \theta_L, \theta_{C_j})$, remains constant under variable illumination. If we vary the lighting conditions so that the irradiance varies by a factor of $k(x_i)$, then the observed reflected radiance, $E'_{C_j}(x_i)$, will also vary by a factor of $k(x_i)$.

$$E'_{C_j}(x_i) = k(x_i)L(x_i)R(x_i, \theta_L, \theta_{C_j}) \tag{6.3}$$

Note that in general neither the irradiance nor the change in irradiance will be equal at different scene points. That is, $L(x_1) \neq L(x_2)$ and $k(x_1) \neq k(x_2)$. This is in contrast to the assumption made in many vision algorithms that the light source

is a precisely isotropic emitter. Consider the two scene variants in Figure 6.1. The configuration of components is identical, but the emitted light intensity field has been changed by rotating the flashlight. The emitted light is not uniform in all directions, and thus $L(x_1) \neq L(x_2)$ and $k(x_1) \neq k(x_2)$.

One thing distinctly worth noticing is that light sources mentioned in this chapter are geometrically static, i.e. stationary during image acquisition. Illumination variations simply come from variable radiant intensity distributions, instead of any spatial position variation of light sources.

Redefining our observation, $E''_{C_j}(x_i)$, as the ratio of two different lighting conditions, gives:

$$E''_{C_j}(x_i) = \frac{E'_{C_j}(x_i)}{E_{C_j}(x_i)} = \frac{k(x_i) \cdot L(x_i) \cdot R(x_i, \theta_L, \theta_{C_j})}{L(x_i) \cdot R(x_i, \theta_L, \theta_{C_j})} = k(x_i) \tag{6.4}$$

Note that the observations are invariant to camera viewpoint and $E''_{C_1}(x_i) = E''_{C_2}(x_i)$ regardless of the surface BRDF.

The simplified formulation just given is sufficient to design a practical stereo system which uses two cameras and a single uncalibrated light source. Practically, this design is easier to implement than existing methods for BRDF invariant stereo, because it requires fewer known or precisely calibrated scene components.

More important from a theoretical standpoint, the introductory formulation can be extended to handle incident lighting for which a single constant k_i can not explain the lighting variation. By factoring the incident light field into a number of basis functions which vary independently, a series of linear equations which relate obser-

vations to lighting and reflectance can be derived. We can then use light transport constancy to formulate a rank constraint on multi-view stereo matching, providing a relation between observations, lighting complexity, and BRDF complexity. One implication of this relation is that stereo matching can be performed precisely even when scenes contain arbitrary BRDFs.

This chapter makes several contributions: the derivation of a rank constraint for stereo using light transport constancy which allows correspondence of arbitrary surface BRDFs, a practical implementation which is easier to reproduce than existing methods for BRDF invariant stereo, and an evaluation of our method on several real scenes to show that it is both practical and effective.

The rest of the chapter is organized as follows. We first develop our light transport constancy and discuss its variations with different lighting and BRDFs in Section 6.1. Experimental results are presented in Section 6.2, using several images captured from scenes with arbitrary BRDFs. Finally we summarize in Section 6.3.

6.1 Light Transport Constancy

Light transport constancy (LTC) can be used to formulate a general constraint on multi-baseline stereo matching regardless of the surface BRDF complexity, provided that sufficient illumination variations and viewpoints are available. A point to emphasize here is that when we discuss illumination variation in the scope of LTC, we mean *radiometric* variations of the light source, i.e., changes in the radiant intensity. This is fundamentally different from *geometric* lighting variations, i.e., moving the light source around, as required in many photometric stereo methods.

This section first presents the rank constraint in the context of multiple point light sources, each of which varies independently. We then show how this can be applied to arbitrary lighting by replacing point lights with arbitrary lighting basis functions. Finally, we expand the formulation to include the concept of BRDF complexity and show that simple BRDFs also provide a rank constraint.

6.1.1 LTC as a rank constraint

The simplified derivation in equation (6.4) assumes that the irradiance is due to a single light source and varies by a single multiplier, k_i. We now formally introduce our radiometric model.

For a single point x_i on the display surface, it is illuminated by a point light source and observed by several cameras. For the sake of simplicity, let us for now assume the camera have just one channel (e.g., a gray-scale camera).

The irradiance at x_i is denoted as $D(x_i, \lambda)$ where λ is the wavelength. Let $R(x_i, \lambda, \theta_L, \theta_{C_j})$ be the spectral reflectance (i.e., BRDF) of x_i indexed by the incident direction θ_L and viewing direction θ_{C_j}. If $t(\lambda)$ is the spectral response for the camera, then the irradiance detected by the camera sensor is:

$$I_{C_j}(x_i) \quad = \quad \int_{\Lambda} D(x_i, \lambda) \cdot R(x_i, \lambda, \theta_L, \theta_{C_j}) \cdot t(\lambda) \, d\lambda, \tag{6.5}$$

where Λ is the camera's spectrum. Note that strictly speaking, the integration should include a cosine term to account for the fore-shortening effect. Since we are dealing with a static scene, it is a per-point scale factor and we consolidate it in the BRDF. Finally, the measured irradiance $I_{C_j}(x_i)$ is converted to a pixel value via a camera

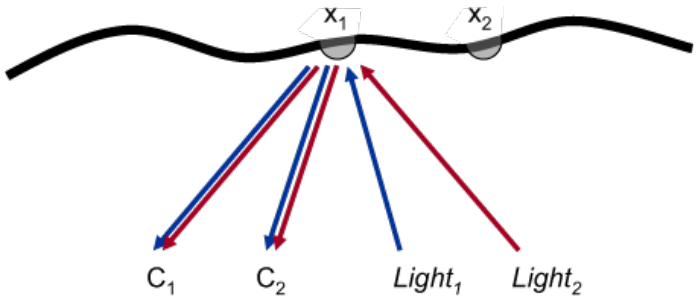


Figure 6.2: Light reflected toward camera C_1 can be explained as a combination of reflected light from each of $Light_1$ and $Light_2$.

response function. For the scope of this chapter, we assume that the camera has a linear response, in other words, the camera is measuring relative irradiance directly. To deal with cameras with non-linear responses, standard radiometric calibration procedures (e.g. [167, 168]) should be applied to correct the pixel values.

If we change only the intensity of $D(x_i, \lambda)$ by a scale factor $k(x_i)$ and keep everything else fixed, $I_{C_j}(x_i)$ will be modulated by $k(x_i)$ according to equation (6.5) and our assumption of linear camera response. This concurs with our intuitive introduction in equation (6.4). We now expand to derive a series of linear equations that can accommodate an arbitrary number of light sources. These equations are the basis for a rank constraint on stereo matching.

Figure 6.2 shows a scene observed from multiple cameras and illuminated by multiple light sources. We can explain the perceived irradiance from a particular scene point, x_i, in the direction of a particular camera, C_j, as a combination of the reflected light from each individual source, $Light_1..Light_M$.

$$I_{C_j}(x_i) = \int_\Lambda D_1(x_i, \lambda) \cdot R(x_i, \lambda, \theta_{L_1}, \theta_{C_j}) \cdot t(\lambda) d\lambda + \int_\Lambda D_2(x_i, \lambda) \cdot R(x_i, \lambda, \theta_{L_2}, \theta_{C_j}) \cdot t(\lambda) d\lambda + \ldots \quad (6.6)$$

For notational convenience we will hereafter drop the indexing for scene location, x_i, since it is understood that each scene location is considered separately. Further, we denote integration constants for particular pairs of light-camera directions as

$$R_{C_1 L_1} = \int_\Lambda D_1(\lambda) \cdot R(\lambda, \theta_{L_1}, \theta_{C_1}) \cdot t(\lambda) d\lambda \quad (6.7)$$

Equation (6.6) can be rewritten using the new notation as:

$$I_{C_j} = R_{C_j L_1} + R_{C_j L_2} + R_{C_j L_3} + \ldots \quad (6.8)$$

We can include the notion of lighting variation in which $D_i(\lambda)$ is modulated by a scalar L_{iV_j}. Let $I_{C_1 V_1}$ be the observed irradiance at camera C_1 under the illumination variation V_1, we can write a sequence of bilinear equations relating the observations from each camera, $C_1..C_J$, under illumination conditions, $V_1..V_N$:

$$\begin{array}{c} I_{C_1 V_1} = L_{1V_1} R_{C_1 L_1} + L_{2V_1} R_{C_1 L_2} + \ldots \\ I_{C_2 V_1} = L_{1V_1} R_{C_2 L_1} + L_{2V_1} R_{C_2 L_2} + \ldots \\ \ldots \\ I_{C_1 V_2} = L_{1V_2} R_{C_1 L_1} + L_{2V_2} R_{C_1 L_2} + \ldots \\ I_{C_2 V_2} = L_{1V_2} R_{C_2 L_1} + L_{2V_2} R_{C_2 L_2} + \ldots \\ \ldots \end{array} \quad (6.9)$$

Note that light transport constancy holds that $R_{C_j L_m}$ is constant for a given pair of light source and camera position regardless of how we vary the illumination conditions. In addition, the illumination variation for a given light source, L_{mV_n}, does not depend on either the BRDF or the camera viewpoint.

This set of linear equations can be rewritten in matrix form as:

$$\underset{\#\text{ of variations}}{\overset{\#\text{ of cameras}}{\begin{bmatrix} I_{C_1V_1} & I_{C_2V_1} & \cdots \\ I_{C_1V_2} & I_{C_2V_2} & \\ I_{C_1V_3} & I_{C_2V_3} & \\ \vdots & & \end{bmatrix}}} = \underset{\#\text{ of variations}}{\overset{\#\text{ of lights}}{\begin{bmatrix} L_{1V_1} & L_{2V_1} & \cdots \\ L_{1V_2} & L_{2V_2} & \\ L_{1V_3} & L_{2V_3} & \\ \vdots & & \end{bmatrix}}} \underset{\#\text{ of lights}}{\overset{\#\text{ of cameras}}{\begin{bmatrix} R_{C_1L_1} & R_{C_2L_1} & \cdots \\ R_{C_1L_2} & R_{C_2L_2} & \\ \vdots & & \end{bmatrix}}} \tag{6.10}$$

Let us denote the matrix on the left side I, and the two matrices on the right L (lighting modulation matrix) and R (reflectance matrix). From the factorization, we can see that there is a rank constraint on matrix I. When the number of light sources, M, is less than both the number of lighting variations and the number of cameras, matrix I has rank of at most M. This constraint allows stereo correspondence to be determined.

6.1.2 Rank constraint with multiple color channels

In the case of color cameras, irradiance I_{C_j} is typically represented as a triple of three intensity values, each representing a distinct color channel in red, green or blue. Let us denote them as $\{I^r_{C_j}, I^g_{C_j}, \text{and } I^b_{C_j}\}$. Similarly we further decompose the spectral response of the light source $Light_i$ into three separate channels: $\{D^r_i(\lambda), D^g_i(\lambda), D^b_i(\lambda)\}$ (imagine that we have a 3-color light projector).

By plugging in different camera/light spectral responses $t(\lambda)$ and $D(\lambda)$ for each

color channel in equation (6.5), $\{I^r_{C_j}, I^g_{C_j}$, and $I^b_{C_j}\}$ can be obtained as

$$\begin{aligned} I^r_{C_j} &= R^r_{C_jL^r_1} + R^r_{C_jL^g_1} + R^r_{C_jL^b_1} \\ I^g_{C_j} &= R^g_{C_jL^r_1} + R^g_{C_jL^g_1} + R^g_{C_jL^b_1} \\ I^b_{C_j} &= R^b_{C_jL^r_1} + R^b_{C_jL^g_1} + R^b_{C_jL^b_1} \end{aligned} \tag{6.11}$$

where $R^l_{C_jL^m_1} = \int D^m_1(\lambda) \cdot R(\lambda, \theta_{L_1}, \theta_{C_1}) \cdot t^l(\lambda) d\lambda$ and $l, m \in \{r, g, b\}$.

With multiple views and multiple lights, we can rewrite the matrix in equation (6.10) for color input as:

$$\underset{\#\text{ of variations}}{\overset{(3\times)\,\#\text{ of cameras}}{\begin{bmatrix} I^r_{C_1V_1} & I^g_{C_1V_1} & I^b_{C_1V_1} & \cdots \\ I^r_{C_1V_2} & I^g_{C_1V_2} & I^b_{C_1V_2} & \\ I^r_{C_1V_3} & I^g_{C_1V_3} & I^b_{C_1V_3} & \\ \vdots & & & \end{bmatrix}}} = \underset{\#\text{ of variations}}{\overset{(3\times)\,\#\text{ of lights}}{\begin{bmatrix} L^r_{1V_1} & L^g_{1V_1} & L^b_{1V_1} & \cdots \\ L^r_{1V_2} & L^g_{1V_2} & L^b_{1V_2} & \\ L^r_{1V_3} & L^g_{1V_3} & L^b_{1V_3} & \\ \vdots & & & \end{bmatrix}}} \underset{(3\times)\,\#\text{ of lights}}{\overset{(3\times)\,\#\text{ of cameras}}{\begin{bmatrix} R^r_{C_1L^r_1} & R^g_{C_1L^r_1} & R^b_{C_1L^r_1} & \cdots \\ R^r_{C_1L^g_1} & R^g_{C_1L^g_1} & R^b_{C_1L^g_1} & \\ R^r_{C_1L^b_1} & R^g_{C_1L^b_1} & R^b_{C_1L^b_1} & \\ \vdots & & & \end{bmatrix}}} \tag{6.12}$$

The reflectance matrix R consists of many 3×3 sub-matrices. Each matrix, typically called as a *color mixing* matrix, records the interaction of the spectral responses of the light source and camera. Typically, the responses of cameras and projectors are wide band and have large overlaps [169]. Thus each sub-matrix has a general form shown above. Nevertheless, as far as stereo matching is concerned, we are only interested in the rank of matrix I on the left side, not the actual decomposition. Therefore the rank constraint we have developed for gray-scale images can be simply extended. That is, matrix I has rank of at most $3 \times M$. Note that although I now has a higher rank, it has $3\times$ the columns as well, so that on balance we expect little change in the outcome.

There are two special cases we can consider. First, with a white light captured by color cameras, the number of columns in the lighting matrix L reduces by a

factor of three. Therefore the rank constraint on I is at most M, i.e., same as using gray-scale cameras. Since we are measuring $3\times$ the data but have only $1\times$ the rank constraint, we expect white lighting and color cameras to be a desirable measurement configuration. Second, with color lighting captured by gray-scale cameras, we have two subcases. If the three color channels scale independently, I has only J columns where J is the number of cameras, but its rank constraint remains $3 \times M$. Naturally this is undesirable since more cameras will be required to ensure that I has a sufficient number of columns. On the other hand, if the three channels scale in the same way, I's rank remains M, which is the same as the gray-scale case.

Dealing with color images is a direct extension from the gray-scale case. Because the notation is cumbersome, we will resume the assumption of a gray-scale world in our remaining discussion.

6.1.3 Arbitrary lighting basis functions

Light transport constancy applies even when light sources are not simple point light sources. Each light in the preceding analysis can be replaced with a lighting basis function, each of which might have broad spatial support.

In general, the irradiance value from a scene point, x_i, in the direction of camera C_j can be written as an integral over all incoming light directions. Therefore equation (6.5) can be modified as the following for a more general lighting setup.

$$I_{C_j} = \iint_{\Phi\ \Lambda} D(\lambda, \phi) \cdot R(\lambda, \phi, \theta_{C_j}) \cdot t(\lambda) d\lambda d\phi \tag{6.13}$$

where $D(\lambda, \phi)$ is the incident light irradiance function indexed by incoming angle ϕ,

and Φ ranges over a hemisphere.

The irradiance field D can be decomposed into a linear combination of basis vectors:

$$D(\lambda, \phi) = k_{L_1} D_1(\lambda, \phi) + k_{L_2} D_2(\lambda, \phi) + \ldots + k_{L_M} D_M(\lambda, \phi) \tag{6.14}$$

It is conceptually helpful to think of each basis as a separate light source. We previously discussed individual point lights as the basis, however area lights represented as a piecewise constant basis, or a wavelet decomposition of the incident illumination field would work equally well. By truncating the wavelet expansion after a sufficient amount of variation has been accounted for, very general lighting can be modeled using a finite set of coefficients. The graphics community has in fact used such an expansion to represent incident illumination fields [170].

We can now rewrite equation (6.13), taking into account the lighting bases and indexed by illumination condition.

$$\begin{aligned} I_{C_j V_n} = k_{L_1 V_n} \iint_{\Phi\,\Lambda} D_1(\lambda, \phi) \cdot R(\lambda, \phi, \theta_{C_j}) \cdot t(\lambda) d\lambda d\phi \\ + k_{L_2 V_n} \iint_{\Phi\,\Lambda} D_2(\lambda, \phi) \cdot R(\lambda, \phi, \theta_{C_j}) \cdot t(\lambda) d\lambda d\phi \end{aligned} \tag{6.15}$$

That is, the observation from camera C_j under illumination condition V_n, is a summation over the individual lighting bases, each modified by their own variation multiplier, $k_{L_m V_n}$.

Notice that each integral term is constant because it relies only on the lighting basis and the surface BRDF. Just as is true in the case of discrete point light sources, lighting variation will induce a set of bilinear equations. These equations can be

written identically to equation (6.10) by redefining variables in terms of the new continuous formulation.

$$L_{mV_n} = k_{L_m V_n}$$
$$R_{C_j L_m} = \iint_{\Phi\,\Lambda} D_m(\lambda, \phi) \cdot R(\lambda, \phi, \theta_{C_j}) \cdot t(\lambda) d\lambda d\phi \tag{6.16}$$

6.1.4 Limited BRDF complexity

So far we have formulated the problem assuming completely arbitrary surface reflectance. However, most real world BRDFs are not arbitrary, and it is unlikely that the reflectance is truly independent in every camera direction. In this case we can further factor the reflectance matrix, R, into a set of reflectance bases, B, and a mixing matrix M.

$$\underset{\text{\# of lights}}{\overset{\text{\# of cameras}}{\begin{bmatrix} R \end{bmatrix}}} = \underset{\text{\# of lights}}{\overset{\text{\# of BRDF bases}}{\begin{bmatrix} B \end{bmatrix}}} \underset{\text{BRDF bases}}{\overset{\text{\# of cameras}}{\begin{bmatrix} M \end{bmatrix}}} \tag{6.17}$$

We now have a trilinear equation **I=LBM**, which has a rank constraint on I if either I or B has a small number of columns. For example, if the surface is Lambertian, then a single BRDF basis describes the outgoing light in all camera directions, and B has a single column. Thus we have a rank constraint if either the illumination or the BRDF is sufficiently "simple". In this work we address completely arbitrary BRDFs and have not evaluated the expected complexity of real world BRDFs.

6.1.5 Stereo matching

It is not necessary to find an actual factorization of the observation matrix I in order to evaluate stereo correspondence. It is sufficient to calculate the singular values of matrix I and select the disparity which results in a matrix of minimum rank.

Because the matrix will be corrupted with noise, it is impossible to calculate rank exactly. Conceptually, we prefer matrices which have most of their energy in the first few principal components rather than those with evenly distributed energy. Thus, we use moments to approximate the notion of minimum rank and select the disparity with minimum score. If the singular values of I are encoded in $w_1..w_n$, then we choose the disparity which minimizes $\Re$.

$$\Re = \sum_i (i \cdot w_i^2) / \sum_i w_i^2 \tag{6.18}$$

When a single light source and only two cameras are used, simply minimizing the second singular value is equivalent to equation (6.18). However, in general it is impossible to use the second (or any particular) singular value as a matching metric, because the expected rank of the matrix is not known a priori.

The introductory matching metric which uses image ratios given in equation (6.4), is also equivalent to equation (6.18). A proof of this equivalence is provided in the Appendix. When only two cameras are used, this simpler matching metric is quite convenient, because it allows existing stereo implementations to be used without modification.

Scharstein and Szeliski have introduced a taxonomy of stereo algorithms which includes matching cost, aggregation, and disparity selection [18]. Light transport

constancy and the implied rank constraint are local operators and replace only the matching cost in existing stereo algorithms. Aggregation, disparity selection, and any global regularization are all orthogonal issues, and the new invariant introduced in this work can be used in conjunction with a wide variety of existing algorithms.

6.2 Experiments

In order to facilitate the evaluation of our technique, we captured several stereo data sets under varying illumination conditions. Our data acquisition setup includes up to four synchronized VGA (640×480) cameras and two light projectors, as shown in Figure 6.3. The cameras are calibrated with respect to each other, but the projectors are completely uncalibrated. Note that much simpler light sources could be substituted–for example, the flashlight shown in Figure 6.1. We use projectors only because they allow the light distribution to be controlled remotely rather than by physically manipulating the light source. The actual light output of the projector is unknown to our algorithm. We used several types of patterns for lighting variation (shown in Figure 6.4), attempting to verify that our results work for both low and high frequency variation. The first is a smooth ramp that is used in the minimum configuration of two lighting variations. The second is a randomly moving Gaussian blob that exhibits low-frequency brightness variation. The third is a pattern acquired from a real flashlight. And the last is a stripe pattern with random intensity values which exhibits high frequency variation. Unless noted otherwise, all the experiments were carried out with the low-frequency (blob) pattern since we expected this to most closely mimic a spotlight which is brighter in the center of its field, similar to the

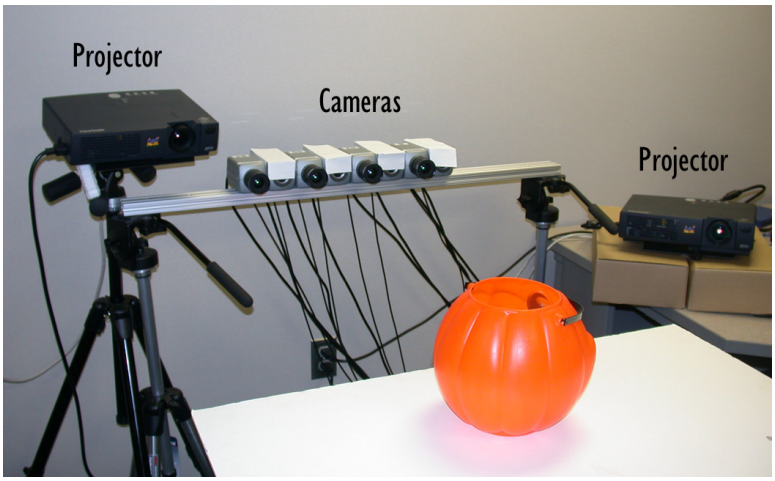


Figure 6.3: Our experimental setup with four cameras and two variable light sources.

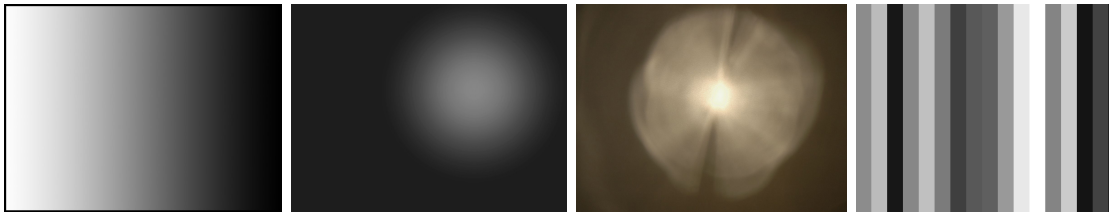


Figure 6.4: Patterns used for lighting variation. From left to right: *ramp* lighting (boxed for illustration purpose), *blob* lighting, *flashlight*, *stripe* lighting.

motivational example shown in Figure 6.1.

Another practical issue to mention is the dynamic range. Saturated pixels (e.g., from specular highlights) will violate the rank constraint we have developed. In our experiments we carefully control the exposure to avoid saturation. It is also possible to combine images taken with multiple exposures to generate a high-dynamic-range (HDR) image (e.g. [168, 171]).

Two-view stereo is the dominant method by which stereo algorithms are evaluated. Although our method is inherently multi-view, we defer to tradition and first evaluate our method in the arrangement we believe will be most commonly implemented. Following these evaluations we provide some analysis of the rank constraint

when multiple cameras and lights are present. Finally we show some quantitative evaluations with a ground-truth data set.

6.2.1 Two-view with one light source

In this setup, we used two cameras and a single light source position. We experimented with gray-scale images to evaluate our method against traditional stereo.

Minimal configuration. We captured gray-scale images from each of two cameras under two different lighting variations. Figure 6.5 shows the two lighting variations from the viewpoint of one of the cameras. The first lighting pattern is a flat gray-field and the second is the ramp in Figure 6.4. Brightness constancy (i.e., traditional intensity difference based on lambertian surfaces) is evaluated using one of the two lighting configurations. Light transport constancy is evaluated by first computing a new image as the ratio of the two illumination conditions, as given in equation (6.4). This process is mathematically equivalent to evaluating the rank constraint. The resulting ratio image is shown in Figure 6.6. Note that neither the specular highlights nor any other view-dependant effect are visible in the ratio image.

Standard stereo matching is applied to the stereo pairs arising from both brightness constancy and light transport constancy using a Sum-of-Absolute-Differences (SAD) metric. Because we are interested in the performance of a local matching operator, we use a WTA approach and simply accept the minimum SAD disparity as correct rather than applying a global regularization method.

Figure 6.7 shows the stereo results from each method. The left column is derived from brightness constancy, and the right column is from light transport constancy.

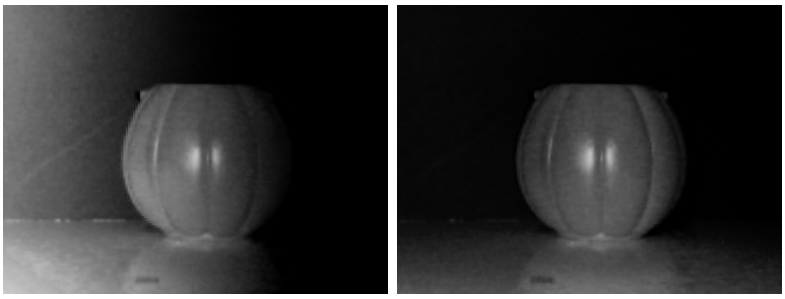


Figure 6.5: A plastic pumpkin illuminated by a single light source under two different lighting conditions.

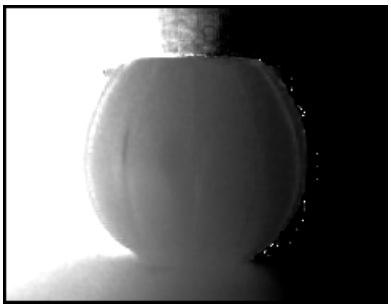


Figure 6.6: The ratio of images taken under two lighting conditions.

The first row shows the disparity map computed by each method. Depth is coded such that white pixels indicate depths closer to the camera. The second row shows the same data along a single scanline as scaled disparity values. In both visualizations, it is clear that our new method has superior results. Note the garbled depth values in the case of brightness constancy. In the third row of Figure 6.7, we investigate the reason that our method performs well by plotting the matching profile for a single pixel. Note that brightness constancy has no clear global minimum, whereas our method has a very clear minimum at the correct disparity. This presumably leads to much better depth estimates.

Together with existing stereo methods. In order to validate that existing stereo

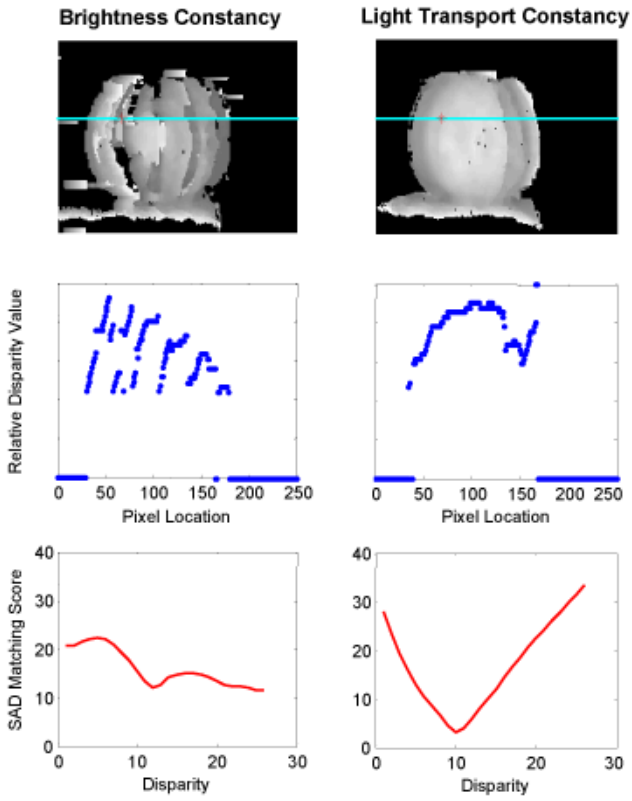


Figure 6.7: Results from using brightness constancy (left column) and light transport constancy (right column). (Row 1) Disparity maps computed by stereo matching using each invariant. (Row 2) Scaled disparity estimates along a single scan line. (Row 3) Matching profile for the pixel marked with a red cross.

methods can be adapted to handle non-Lambertian objects, we tested the same two sets of gray-scale stereo pairs with a stereo implementation available on the web [172]. This implementation happens to be based on graph cuts [117], which allowed us to further verify that no undesirable artifacts are caused by integration with a global regularization method. Since we have computed a ratio image to use for matching, absolutely no modification to the existing code is required. The computed disparity maps are shown in Figure 6.8. Similar to the WTA example above, the disparity map computed using light transport constancy shows much better results.

Increased lighting variation. It is possible that our improved results come merely because by imposing lighting variation more information is available when computing disparity, rather than because our new invariant actually performs better. To eval-

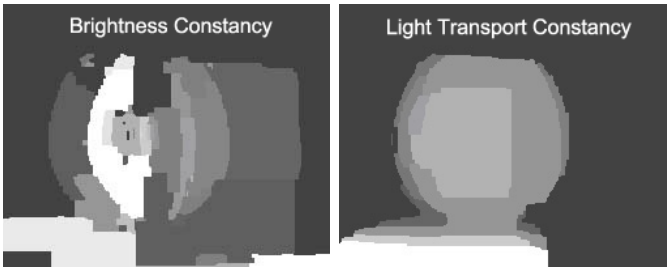


Figure 6.8: Disparity maps computed using an unmodified graph-cut stereo algorithm with brightness constancy (left) and our new invariant (right).

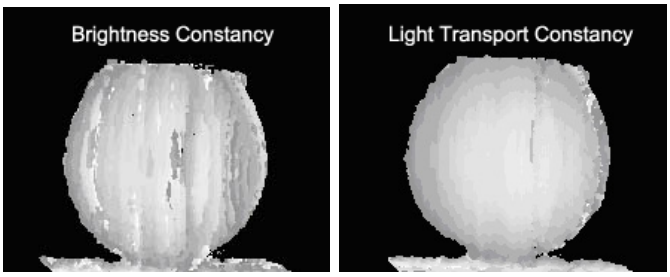


Figure 6.9: Disparity maps computed from a data set with six illumination variants. Left is from brightness constancy; right is from light transport constancy.

uate whether this is true, we computed disparity using a data set with six lighting variations, as shown in Figure 6.9. Brightness constancy is evaluated as the Sum-of-Absolute-Differences over the vector of all six image pairs. Light transport constancy is evaluated as a rank constraint over the same input images. Although it is clear that additional lighting variations improve the result from brightness constancy, the result from light transport constancy also improves. We conclude that additional lighting variations improve the results from either constraint but that our new invariant performs better on objects such as the pumpkin, which exhibit non-Lambertian effects.

Using a simulated flashlight. We captured a lighting pattern of a regular flashlight (shown as one pattern in Figure 6.4). To facilitate automatic data acquisition, we use a projector to display five variations of the flashlight pattern with shift or rotation, simulating the scenario described in Figure 6.1. Good results can be obtained as shown in Figure 6.10.

Complex reflectance. We further experimented with scenes containing more com-

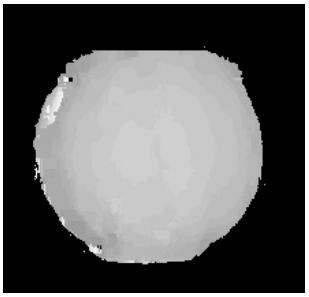


Figure 6.10: Reconstructed depth map using a simulated flashlight with five lighting variations.

plex surface material properties. We first captured a piece of silk glued onto a slightly curved surface. The view dependent reflectance of the silk is very obvious in the stereo pair, as shown in Figure 6.11. Using seven lighting variations, we evaluate brightness constancy against our new invariant and find that light transport constancy is better able to deal with this highly non-Lambertian scene. The improvement is particularly obvious in the plot of disparity along a scanline, shown in the bottom row of Figure 6.12. Brightness constancy results in many incorrect disparity estimates, whereas light transport constancy results in a smooth curve.

Multi-channel color. The advantage of light transport constancy over brightness constancy is further demonstrated in Figure 6.13. We captured full color images of a lady's purse made from materials with a complex anisotropic BRDF. Note the surface color changes in the stereo image pair: the right side of the purse appears to be blue in one image and pink in the other image. With a white light source, we captured just two lighting variations in full color. We use as few lighting variations as possible to illustrate the effectiveness of our approach. All color images were used to compute each of light transport constancy and brightness constancy. The reconstructed depth maps are shown in the bottom row of Figure 6.13. We would not expect brightness constancy to perform well under these conditions, and indeed

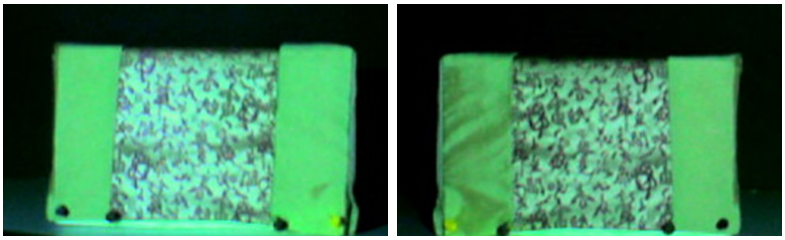


Figure 6.11: Silk cloth from two different viewpoints. Note the non-Lambertian reflectance.

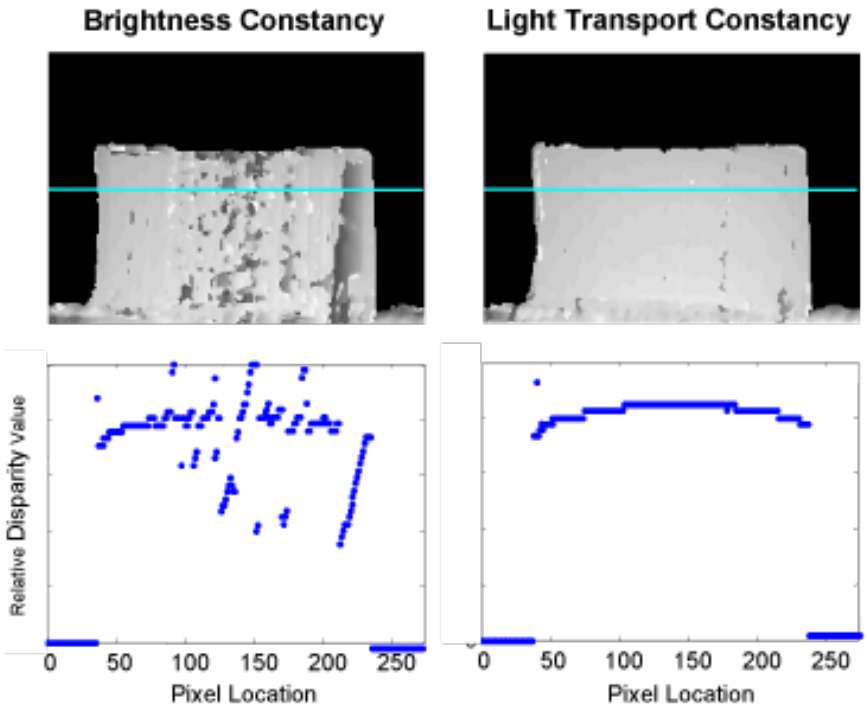


Figure 6.12: (Top) Disparity maps computed using brightness constancy and light transport constancy (LTC). (Bottom) Scaled disparity values along a single scanline. Note how much more robustly LTC estimates depth.

we see that the computed object depth is erroneous in the region exhibiting color change. In contrast, light transport constancy is able to evaluate depth accurately.

Note that it is not required to use the multichannel color formulation to compute disparity on colored objects such as this. We converted the input images to gray-scale to experiment with the formulation given in equation (6.10), and found the result to

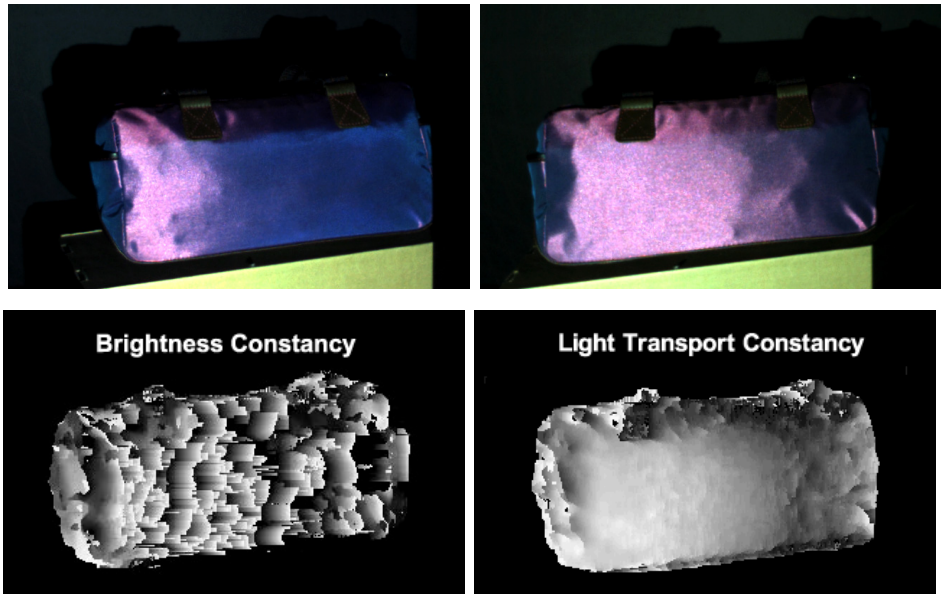


Figure 6.13: Stereo reconstruction of a lady's purse with anisotropic BRDFs. (Top row) the left and right images under one lighting condition; note the color changes in two images. (Bottom left) reconstructed depth map using brightness constancy. (Bottom right) reconstructed depth map using light transport constancy.

be qualitatively similar to that in Figure 6.13, which is computed using the full 3-channel color formulation given in equation (6.1.2), with the caveat that the rank on the matrix I is expected to be one because we use a gray-scale light source. As we have discussed in Section 6.1.2, this is a more favorable configuration for matching.

Complex geometry. Our next data set is a live tree with substantial specular highlights. This scene would be challenging for traditional stereo algorithms due to the non-Lambertian effects and because there are many depth discontinuities. For this setup, we used the high-frequency (stripe) pattern with 30 variations to calculate the disparity map shown in Figure 6.14. With such a large number of lighting conditions, we would anticipate good performance. As expected, the results are of high quality. Individual leaves are well represented by clean boundaries and smooth estimates of depth, despite the fact that no global regularization method is applied.

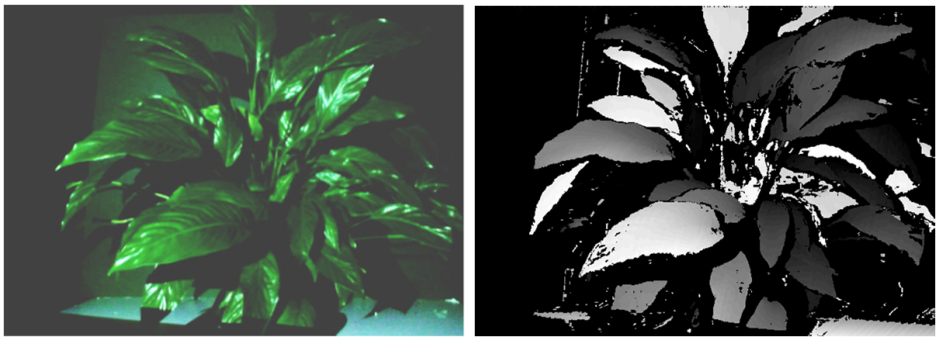


Figure 6.14: (Left) Tree with non-Lambertian reflectance properties and many depth discontinuities. (Right) Disparity map computed from thirty lighting variations.

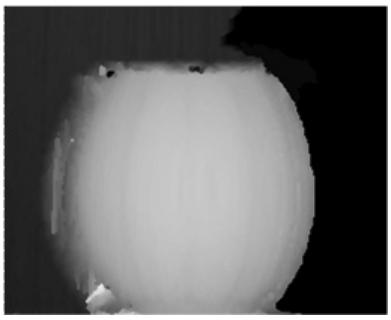


Figure 6.15: Disparity map for the pumpkin calculated from multiple cameras and multiple light sources.

6.2.2 Multi-view with two light sources

To evaluate the behavior of the rank constraint under multi-view conditions, we computed disparity on the pumpkin scene using four cameras, two light sources, and thirty lighting variations. The resulting disparity map can be seen in Figure 6.15. As a whole, the results are very good, with smooth estimates of depth across the surface of the pumpkin. There is an error in the lower left corner which we believe is caused by occlusion from some camera viewpoints. Accounting for partial occlusion is typically handled during the aggregation stage of stereo processing, and, as mentioned earlier, we focus on the matching cost in this work.

Analysis of singular values. When two light sources are used, the rank of the

observation matrix is limited to 2 for surfaces with arbitrary BRDF. In this case, we expect the third singular value to be minimized at the correct disparity. However, if the complexity of the surface reflectance is limited, the rank may be lower. This could happen either if the surface was actually Lambertian or merely because it appears Lambertian from the limited set of viewpoints available.

To provide some insight into the behavior of our rank constraint, we plotted the 2^{nd}, 3^{rd}, and 4^{th} singular values as a function of disparity for two different scene points, drawn from the multi-view example above. For the scene point in the top plot of Figure 6.16, we see that the 2^{nd} singular value has an obvious minimum and that the combined metric $\Re$ is minimized at this same disparity. However, in the case of the scene point in the bottom plot, $\Re$ is minimized at the same disparity as the 3^{rd} singular value. Although the 4^{th} singular value is not precisely zero as would be expected in an ideal environment without noise, we can see that $\Re$ has an easily locatable global minimum which confirms that our approximation of "minimum rank" is performing as expected.

6.2.3 Quantitative Evaluation

While the Middlebury stereo evaluation web site [1] has become the gold standard to evaluate performance of stereo algorithms, the datasets there do not include lighting variations therefore cannot be used for our approach. In order to generate our own "ground-truth" data, we project a single vertical strip pattern from the light projector and calculate the depth along this strip using traditional stereo. The strip is swept across the scene simulating a laser triangulation-based scanner. Since only the strip

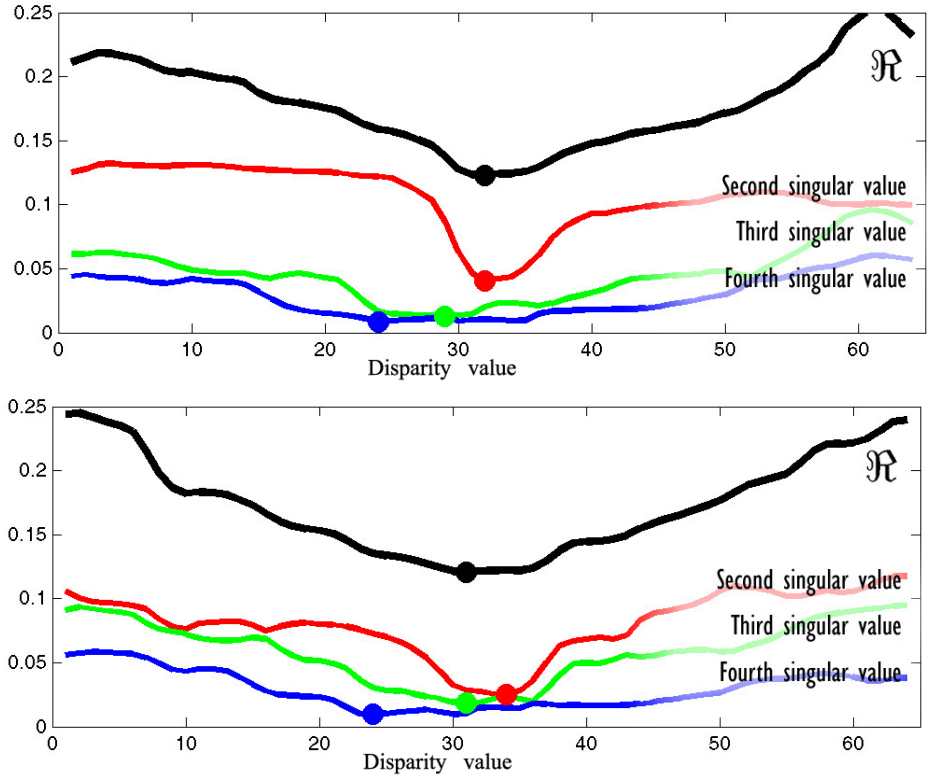


Figure 6.16: Normalized singular values for two particular scene points. The x-axis represents the disparity. Dots indicate the minimum on each curve. The moment has been scaled to fit on the same graph together with the singular values. Note that the moment is minimized together with a different singular value in each case.

is illuminated, disparity can be calculated unambiguously. The advantage of this approach as opposed to using a real laser range scanner is that the ground-truth data is automatically registered in the stereo cameras' coordinate frame.

Figure 6.17 shows a data set we captured. Note that we have not implemented sub-pixel disparity interpolation, so some of the groves on the surface are not visually noticeable in the depth map. Bad pixels around the silhouette (due to occlusions) are manually removed.

We generated depth results under varying patterns and compared the depth maps with the ground-truth data. If a pixel's disparity differs more than one pixel from the

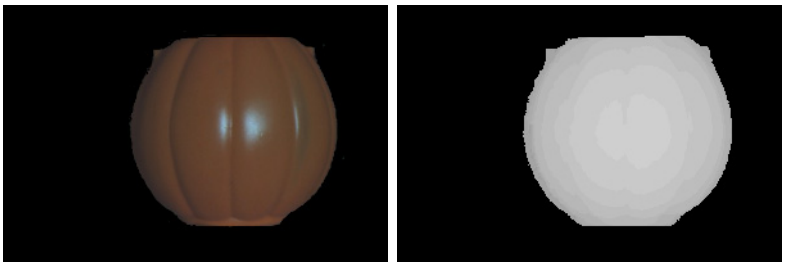


Figure 6.17: The ground truth dataset.Left is one color image and right is its corresponding depth map. Bad pixels due to occlusions are manually removed.

ground truth, we label it as a bad match. The error rates from two methods, one using the brightness constancy (BC) and the other using LTC, are summarized in Table 6.1 and Figure 6.18. In general, the error rates for both methods reduce as the number of lighting variations increases. This is not surprising because there are more data to work with. With just a few low-frequency lighting variations, the error rate from BC is very large and changes quite arbitrarily. Results from LTC are much better. On the other hand, with high frequency lighting, both BC and LTC can generate much more accurate results and the difference in error rates is much smaller. These rapid lighting variations in fact "mask" the surface reflectance properties. This is similar to the fact that regular structured light scanners using binary-coded patterns can get decent results from shiny objects. Nevertheless, LTC always outperforms BC in all testing cases.

Table 6.1: Error rate of depth maps computed with brightness constancy (BC) and light transport constancy (LTC). Different lighting patterns (as shown in Figure 6.4) are used for this evaluation.

# of lighting variations	Low-freq. (blob)						High-freq. (stripe)		
	2	3	4	8	16	32	8	16	32
BC Error (%)	54.5	65.0	61.1	40.7	44.1	21.3	3.44	3.95	3.25
LTC Error (%)	13.6	9.27	5.9	5.28	3.9	3.10	3.33	3.13	2.36

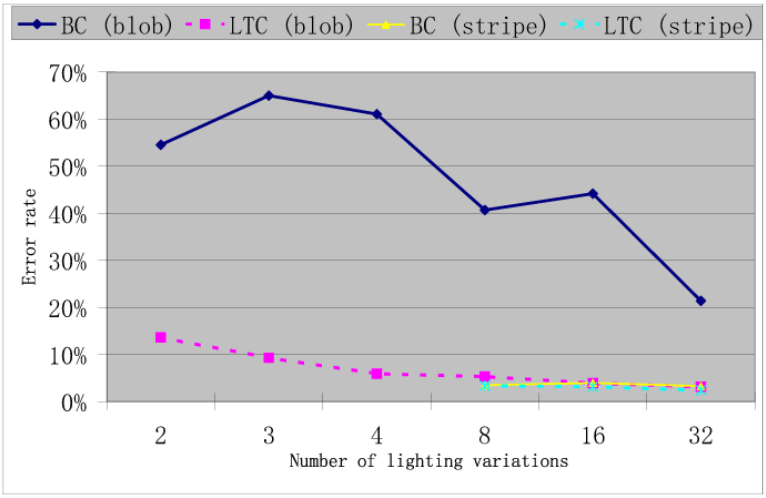


Figure 6.18: A plot of the error rates using data from Table 6.1.

6.3 Summary

Light transport constancy is a new invariant for multi-view stereo matching which allows the depth of surfaces with arbitrary BRDF to be computed. We introduce a rank constraint based on this invariant which allows stereo algorithms to combine observations of non-Lambertian surfaces from different viewpoints in a theoretically principled way.

Our rank constraint can be applied with as few as two cameras and two lighting configurations. In addition, unlike existing methods for non-Lambertian stereo, our method does not require that light sources be precisely calibrated nor does it require known calibration objects in the scene. The rank constraint implied by light transport constancy can easily be employed as a replacement to brightness constancy. Thus, whenever sufficient lighting variation is available, any existing stereo algorithms can be enhanced to allow matching of non-Lambertian surfaces. We have verified experimentally that stereo matching is possible using our rank constraint. In addition, we

show that it performs better than brightness constancy on a variety of scenes.

A few aspects of our work may limit the conditions under which light transport constancy can be used. The rank constraint requires multiple illumination conditions to be available. All previously existing methods for arbitrary BRDF stereo also require illumination variation [133, 140], and it is interesting to wonder if this is a fundamental requirement. In addition, we do not consider the issue of inter-reflection in our formulation. In scenes with strong inter-reflection (e.g., concave and shiny objects), some points may have a higher rank than the rest (consider inter-reflections as additional light sources). Experiments are needed to see if inter-reflection can be treated as a secondary effect or noise. Finally, the rank constraint is a multi-view constraint, and we do theoretically require more camera viewpoints than light source positions when the surface BRDF is truly arbitrary. However, the BRDF of most real surfaces is not arbitrary, and we have shown that BRDF complexity can be traded for lighting complexity. Thus an interesting avenue for future work would be to characterize the actual matrix rank, and thus the actual number of viewpoints required, for a wide class of naturally occurring scenes and lighting.

Chapter 7 Conclusions and Future Work

In this dissertation, we have focused on the classical stereo matching problem - estimating the scene depth information from a collection of calibrated images gathered from different viewpoints. We have presented novel dense stereo algorithms for high-quality depth estimation from images. In this chapter, we summarize our technical innovations and suggest areas for future work.

7.1 Innovations

This dissertation has introduced the following five innovations:

- **Two-Pass Approximation of the Bilateral Filtering Based Cost Aggregation.** We investigate the use of two separate 1D windows, one horizontal, and one vertical, to approximate the full bilateral filtering based cost aggregation approach originally described in [3]. Our approximation leads to low computational complexity and satisfactory cost-volume smoothing results. The two-pass approximation is also suitable for hardware acceleration. We propose a GPU implementation of this two-pass adaptive aggregation method and showed that the GPU version is orders of magnitude faster than its CPU counterpart.

- **Real-Time Stereo using Vertical Aggregation and Dynamic Programming.** For high-quality depth estimation in real-time, we propose to incorporate the two-pass cost aggregation scheme into a dynamic programming (DP)

stereo framework. We found that changing the window shapes from conventional squares to vertical rectangles allows overall smooth depth estimates, fine structures near depth discontinuities, and much less scanline inconsistency ("streaking") artifacts. A hybrid (GPU + CPU) implementation makes it one of the fastest stereo algorithms available.

- **GCPs-Based Regularization Prior for Global Stereo.** We propose a novel formulation for stereo reconstruction that makes use of constraints from sparse ground control points (GCPs). Prior constraints about the scene structure derived from the GCPs are incorporated into a global inference framework via an MRF formulation in a principled way. We demonstrated that using GCPs computed automatically from stable matching, our stereo model can improve the reconstruction accuracy without resorting to image segmentation, plane fitting, or additional sensors. Furthermore we showed that our stereo formulation is able to handle surfaces with different orders of smoothness, such as those with high-curvature details.

- **Fusion of Low Resolution LiDAR Data and High Resolution Imagery for 3D Reconstruction.** Based on our proposed stereo formulation in Chapter 5, we fuse low resolution LiDAR data acquired from a mobile range scanner with high resolution images captured by digital cameras for 3D reconstruction. In this scenario, we demonstrate that GCPs can be obtained from external sensors and our stereo model is able to improve the stereo matching quality by leveraging the constraints from sparse LiDAR measurements.

- **Light Transport Constancy for Stereo Correspondence Beyond Lambert.** We introduce a new matching invariant for stereo called *light transport constancy* (LTC) and use it to formulate a rank constraint for multi-view stereo. LTC does not require calibrated light sources or calibration objects in the scene and allows stereo matching to be performed precisely even when the scenes contain arbitrary surface BRDFs. Our new constraint can be used to provide BRDF invariance to any existing stereo method whenever appropriate lighting variation is available.

7.2 Future Work

At the end of each previous chapter (Chapters 4, 5,and 6), we discussed limitations of our proposed methods and suggested relative immediate issues for future work. In this section, we propose a few more ambitious research topics and share our impressions of future trends in stereo matching.

After roughly 40 years of research on stereo, many elements of stereo algorithms had, in many ways, matured. For instance, camera calibration, stereo geometry, and efficient methods for local correspondences search are well understood. Perhaps the most significant progress in the last decade has been the advance of global stereo methods that based on the MRF formulation [85]. In particular, the development of powerful optimization algorithms (e.g., graph cuts [165] and belief propagation [173]) and effective regularization priors (e.g., segment-based priors [122,174] and high-order smoothness priors [2, 88] has dramatically pushed the envelope of stereo research, giving substantially more accurate results than were previously possible. However,

nearly all top stereo algorithms were evaluated using the Middlebury benchmark data set [5, 175] which has been captured inside the laboratory with ideal lighting conditions, Lambertian materials and piecewise planer surfaces (Figure 2.4). On the other hand, depth estimation for outdoor environments is of greater relevance to applications but also more challenging. Typical difficulties that a stereo method needs to conquer in an outdoor scene (e.g. Figure 7.1) include large textureless regions (ground and sky), non-rigid scenes (pedestrians or vehicles), non-Lambertian reflectance (windows and metals), changing light conditions, and surface with high-curvature details, etc. These aspects form a particular challenge for outdoor stereo reconstructions. We believe that in the coming decade, the focus of stereo algorithms should turn to handling real-world images that has been acquired outside the laboratory without attempting to find simple or ideal cases. We also expect to see more complete benchmark data that contains realistic scenes, i.e., outdoor scenes for which active stereo is not applicable. The recent work by Strecha et al. [6]is a promising first step.

Another important research direction to explore is the potential of using machine learning and data driven approaches to help stereo reconstruction overcome its weakness. Recently, learning has been successfully applied to single image 3D reconstruction [31, 33]. Unlike stereo vision which reconstructs 3D via triangulation, depth estimates from monocular cues are entirely based on the evidence about the environment presented in a single image. A natural question to ask is whether one can combine the monocular cues with multi-view cues for improved depth estimation. We believe that for certain types of scene (e.g. urban environments), monocular cues and geometric-based stereo cues give largely orthogonal, and therefore complemen-

Figure 7.1: Image of a typical outdoor urban scene.

tary information about depth. Take the scene in Figure 7.1 for example, stereo should be able to predict correct disparities for building facades and thin structures which contain sufficient texture variations, however tends to fail for textureless regions such as the ground and sky. On the other hand, monocular cues which depend on the overall content of the image, are better at handling these homogeneously textured regions. Looking into the near future, investigating how monocular cues can be integrated with passive stereo to obtain better depth estimates than using stereo alone is, in our view, a very promising direction.

Appendix

We show in this appendix that the multiview rank constraint proposed in Chapter 6 is equivalent to the absolute difference of ratio images when only two viewpoints and two illumination conditions are present. That is, given image intensities from illumination conditions a and $k_1 a$ from camera A, and b and $k_2 b$ from camera B, we want to show that the observation matrix's second singular value has a minimum at the same disparity as $|k_1 - k_2|$.

The observation matrix can be written as: $\begin{bmatrix} a & b \\ k_1 a & k_2 b \end{bmatrix}$

Its second singular value s_2 can be calculated as:

$$s_2 = \frac{1}{2}\left[2(a^2 + k_1^2 + k_2^2 b^2 + b^2) - 2\sqrt{\begin{array}{c} a^4(k_1^2+1)^2 + b^4(k_2^2+1)^2 + \\ 2a^2b^2(k_1k_2+1)^2 - 2a^2b^2(k_1-k_2)^2 \end{array}}\right]^{\frac{1}{2}} \tag{A.1}$$

Let us define $d = k_1 - k_2$ such that it is positive, reversing the role of k_1 and k_2 if necessary. Note also that a, b, k_1, k_2 are all positive due to physical constraints.

It can be shown that $s_2 = 0$ if and only if $k_1 = k_2$, (given non-zero a and b). Similarly, it is obvious that $|k_1 - k_2| = 0$ only when $k_1 = k_2$. It remains to be shown that s_2 is related to d by a monotonic relationship, such that an increase in s_2 always implies an increase in d.

Now if we replace k_2 with $k_1 - d$ in equation (A.1) and take the derivative of s_2 with respect to d, we have:

$$\begin{aligned} \frac{\partial(s_2)}{\partial(d)} =& \frac{1}{4\sqrt{s_2}}[4(k_1+d)b^2 - \frac{1}{G}(4b^4((k_1+d)^2+1)(k_1+d) + \\ & 4a^2b^2(k_1(k_1+d)+1)k_1 - 4a^2b^2d], \end{aligned} \tag{A.2}$$

where

$$G = \sqrt{\begin{array}{l} a^4(k_1^2+1)^2 + b^4((k_1+d)^2+1)^2 + \\ 2a^2b^2(k_1(k_1+d)+1)^2 - 2a^2b^2d^2 \end{array}} \tag{A.3}$$

We need to show that this derivative is always positive, $\frac{\partial(s_2)}{\partial(d)} \geq 0$, which is the same as showing:

$$\begin{array}{rl} 4(k_1+d)b^2 > & \frac{1}{G}(4b^4((k_1+d)^2+1)(k_1+d)+ \\ & 4a^2b^2(k_1(k_1+d)+1)k_1 - 4a^2b^2d) \end{array} \tag{A.4}$$

Taking square of both sides and simplifying results in:

$$\begin{array}{l} 64a^2b^6d^2 + 128a^2b^6k_1^2d^2 + 64a^2b^6k_1^3d + 64a^2b^6k_1d^3 + \\ 64a^4b^4k_1^3d + 64a^2b^6k_1d + 64a^4b^4k_1d + 64a^4b^4k_1^2d^2 > 0 \end{array} \tag{A.5}$$

The above inequality holds true because all variables are positive, and thus the rank constraint is equivalent to using the absolute difference of the ratio images.

Bibliography

[1] D. Scharstein and R. Szeliski. www.middlebury.edu/stereo.

[2] B. M. Smith, L. Zhang, and H. Jin. Stereo matching with nonparametric priors in feature space. In *Proceedings of CVPR*, 2009.

[3] K.-J. Yoon and I.-S. Kweon. Adaptive support-weight approach for correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):650–656, 2006.

[4] A.F. Bobick and S.S. Intille. Large occlusion stereo. *International Journal of Computer Vision*, 33(3):181–200, 1999.

[5] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Proceedings of CVPR*, pages 195–202, 2003.

[6] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibrarion and multi-view stereo for high resolution imagery. In *Proceedings of CVPR*, 2008. http://cvlab.epfl.ch/ strecha/multiview/denseMVS.html.

[7] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

[8] M. Pollefeys, D. Nister, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton,

L. Wang, Q. Yang, H. Stewenius, R. Yang, G. Welch, and H. Towles. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78(2):143–167, 2008.

[9] C. Zhang, L. Wang, and R. Yang. Semantic Segmentation of Urban Scenes Using Dense Depth Maps. In *Proceedings of ECCV*, pages 708–721, 2010.

[10] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. Probabilistic fusion of stereo with color and contrast for bilayer segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28, 2006.

[11] Daniel Scharstein. *View synthesis using stereo vision*. Springer-Verlag Berlin, Heidelberg, ISBN 3-540-66159-X, 1999.

[12] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Proceedings of ICPR*, pages 15–18, 2006.

[13] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister. Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. In *Proceedings of CVPR*, pages 2347–2354, 2006.

[14] M. Gong, R. Yang, L. Wang, and M. Gong. A performance study on different cost aggregation approaches used in real-time stereo matching. *International Journal of Computer Vision*, 75(2):283–296, 2007.

[15] S. Forstmann, J. Ohya, Y. Kanou, A. Schmitt, and S. Thuering. Real-time stereo by using dynamic programming. In *Proc. of CVPR Workshop on Real-time 3D Sensors and Their Use*, 2004.

[16] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *Proceedings of CVPR*, 1998.

[17] H. Tao, H. S. Sawhney, and R. Kumar. A global matching framework for stereo computation. In *Proceedings of ICCV*, pages 532–539, 2001.

[18] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1):7–42, May 2002.

[19] D. Nister. *Automatic Dense Reconstruction from Uncalibrated Video Sequences*. PhD Thesis, Royal Institute of Technology KTH, Stockholm, Sweden, ISBN 91-7283-053-0, 2001.

[20] S. Zhu, L. Zhang, and B. M. Smith. Model evolution: An incremental approach to non-rigid structure from motion. In *Proceedings of CVPR*, 2010.

[21] J. Taylor, A. D. Jepson, and K. N. Kutulakos. Non-rigid structure from locally-rigid motion. In *Proceedings of CVPR*, 2010.

[22] J. Shi and C. Tomasi. Good features to track. In *the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 593–600, Washington, June 1994.

[23] M. Z. Brown, D. Burschka, and G. D. Hager. Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:2003, 2003.

[24] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of CVPR*, 2006.

[25] Y. Horry, K. Anjyo, and K. Arai. Using a spidery mesh interface to make animation from a single image. In *Proceedings of Siggraph*, page 225232, 1997.

[26] S. B. Kang. *ADepth Painting for Image-Based Rendering Applications*. Technical Report, Compaq Computer Corporation, Cambridge Research Lab, 1998.

[27] L. Zhang, G. Dugas-Phocion, J.-S. Samson, and S. M. Seitz. Single view modeling of free-form scenes. In *Proceedings of CVPR*, 2001.

[28] T. Lindeberg and J. Garding. Shape from texture from a multi-scale perspective. In *Proceedings of ICCV*, 1993.

[29] J. Malik and R. Rosenholtz. Computing local surface orientation and shape from texture for curved surfaces. *International Journal of Computer Vision*, 23(2):149–168, 1997.

[30] D. A. Forsyth. Shape from texture and integrability. In *Proceedings of ICCV*, 2001.

[31] A. Saxena, S. H. Chung, and A. Y. Ng. 3-D depth reconstruction from a single still image. *International Journal of Computer Vision*, 76(1), 2007.

[32] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *Proceedings of ICCV*, 2005.

[33] D. Hoiem, A. A. Efros, and M. Hebert. Recovering Surface Layout from an Image. *International Journal of Computer Vision*, 73(1):151–172, 2006.

[34] B.K.P. Horn. Obtaining shape from shading information. In *Proceedings of The Psychology of Computer Vision*, pages 115–155, 1975.

[35] B.K.P. Horn and M.J. Brooks. Shape and source from shading. In *International Joint Conference on Artificial Intelligence*, pages 932–936, 1985.

[36] B.K.P. Horn and M.J. Brooks. *Shape from Shading.* MIT Press, 1989.

[37] B.K.P. Horn. Understanding image intensities. *Artificial Intelligence*, 8(2):201–231, 1977.

[38] B.K.P. Horn and M.J. Brooks. The variational approach to shape from shading. *Computer Vision Graphics and Image Processing*, 33(2):174–208, 1986.

[39] Y.G. Leclerc and A.F. Bobick. The direct computation of height from shading. In *Proceedings of CVPR*, pages 552–558, 1991.

[40] Richard Szeliski. Fast shape from shading. *Computer Vision Graphics and Image Processing: Image Underst.*, 53(2):129–153, 1991.

[41] E. Prados and O. Faugeras. Shape from shading: a well-posed problem? In *Proceedings of CVPR*, 2005.

[42] E. North Coleman Jr. and R. Jain. Shape from shading for surfaces with texture and specularity. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 652–657, 1981.

[43] S. Bakshi and Y.H. Yang. Shape from shading for non-lambertian surfaces. In *Proceedings of ICIP*, pages 130–134, 1994.

[44] M. Asada, T. Nakamura, and Y. Shirai. Weak lambertian assumption for determining cylindrical shape and pose from shading and contour. In *Proceedings of CVPR*, pages 726–729, 1992.

[45] A. Ortiz and G. Oliver. Shape from shading for multiple albedo images. In *Proceedings of ICPR*, pages Vol I: 786–789, 2000.

[46] R. Zhang, P.S. Tsai, J.E. Cryer, and M. Shah. Shape from shading: A survey. *PAMI*, 21(8):690–706, 1999.

[47] R.J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980.

[48] E. North Coleman Jr. and R. Jain. Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry. *Computer Graphics Image Processing*, 18(4):309–328, 1982.

[49] H.D. Tagare and R.J.P. de Figueiredo. A theory of photometric stereo for a class of diffuse non-lambertian surfaces. *PAMI*, 13(2):133–152, 1991.

[50] A. Hertzmann and S.M. Seitz. Example-based photometric stereo: Shape reconstruction with general, varying brdfs. *PAMI*, 27(8):1254–1264, 2005.

[51] S. B. Kang, J. A. Webb, C. L. Zitnick, and T. Kanade. A multibaseline stereo system with active illumination and real-time image acquisition. In *Proceedings of ICCV*, 1995.

[52] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. *PAMI*, 27(2):296–302, 2005.

[53] L. Zhang, B. Curless, and S. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *Proceedings of CVPR*, pages 367–374, 2003.

[54] T. Zickler, P.N. Belhumeur, and D.J. Kriegman. Helmholtz stereopsis: Exploiting reciprocity for surface reconstruction. In *Proceedings of ECCV*, pages 869–884, 2002.

[55] T. Zickler, D.J. Kriegman J. Ho, J. Ponce, and P.N. Belhumeur. Binocular helmholtz stereopsis. In *Proceedings of ICCV*, pages 1411–1417, 2003.

[56] T. Darrell and K. Wohn. Pyramid based depth from focus. In *Proceedings of CVPR*, pages 504–509, 1988.

[57] E. Krotkov. Focusing. *International Journal of Computer Vision*, 1(3):223–237, 1987.

[58] A.P. Pentland. A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(4):523–531, 1987.

[59] M. Subbarao and G. Surya. Depth from defocus: A spatial domain approach. *International Journal of Computer Vision*, 13(3):271–294, 1994.

[60] Samuel W. Hasinoff and Kiriakos N. Kutulakos. Confocal stereo. In *Proceedings of ECCV*, pages 620–634, 2006.

[61] Li Zhang and Shree K. Nayar. Projection defocus analysis for scene capture and image display. *Proceedings of ACM Siggraph*, 25(3):907–915, 2006.

[62] B. Girod and S. Scherock. Depth from defocus of structured light. In *Proceedings of SPIE Conference on Optics, and Image Sensing for Machine Vision*, 1989.

[63] D. A. Forsyth and J. Ponce. *Computer Vision A Modern Approach.* Prentice Hall, Upper Saddle River, New Jersey, ISBN 0-13-085198-1, 2003.

[64] D. Scharstein. View Synthesis Using Stereo Vision. *Lecture Notes in Computer Science (LNCS)*, 1583, 1999.

[65] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[66] R. Sara. Finding the largest unambiguous component of stereo matching. In *Proceedings of ECCV*, pages 900–914, 2002.

[67] J. Cech and R. Sara. Efficient sampling of disparity space for fast and accurate matching. In *Intl. Workshop on Benchmarking Automated Calibration, Orientation,and Surface Reconstruction from Images*, 2007.

[68] D. Papadimitriou and T. Dennis. Epipolar line estimation and rectification for stereo image pairs. *IEEE Transactions on Image Processing*, 5(2):672–676, 1996.

[69] C. Loop and Z. Zhang. Computing Rectifying Homographies for Stereo Vision. In *Proceedings of CVPR*, pages 125–131, 1999.

[70] M. Pollefeys, R. Koch, and L. Van Gool. A Simple and Efficient Rectification Method for General Motion. In *Proceedings of ICCV*, pages 496–501, 1999.

[71] S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4):401–406, 1998.

[72] H. Hirschmuller, P. R. Innocent, and J. M. Garibaldi. Realtime correlation-based stereo vision with reduced border errors. *International Journal of Computer Vision*, 47:229–246, 2002.

[73] R. Zabih and J. Woodfill. Non-parametric Local Transforms for Computing Visual Correspondance. In *Proceedings of ECCV*, pages 151–158, 1994.

[74] H. Hirschmuller and D. Scharstein. Evaluation of cost functions for stereo matching. In *Proceedings of CVPR*, 2007.

[75] R. Yang and M. Pollefeys. Multi-resolution real-time stereo on commodity graphics hardware. In *Proceedings of CVPR*, pages 211–217, 2003.

[76] J. C. Kim, K. M. Lee, B. T. Choi, and S. U. Lee. A dense stereo matchings uing two-pass dynamic programming with generalized ground control points. In *Proceedings of CVPR*, 2005.

[77] M. Okutomi and T. Kanade. A Locally Adaptive Window for Signal Matching. *International Journal of Computer Vision*, 7(2):143–162, 1992.

[78] O.Veksler. Stereo Matching by Compact Windows via Minimum Ratio Cycle. In *Proceedings of ICCV*, pages 540–547, 2001.

[79] T. Kanade and M. Okutomi. A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–932, 1994.

[80] L. Wang, M. Liao, M. Gong, R. Yang, and D. Nister. High-quality real-time stereo using adaptive cost aggregation and dynamic programming. In *Proceedings of 3DPVT*, 2006.

[81] S. Mattoccia, S. Giardino, and A. Gambini. Accurate and Efficient Cost Aggregation Strategy for Stereo Correspondence Based on Approximated Joint Bilateral Filtering. In *Proceedings of ACCV*, 2009.

[82] F. Tombari, S. Mattoccia, L. D. Stefano, and E. Addimanda. Classification and Performance Evaluation of Different Aggregation Costs for Stereo Matching. In *Proceedings of CVPR*, 2008.

[83] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[84] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision*, 70(1):41–54, 2006.

[85] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. F. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):1068–1080, 2008.

[86] J. Sun, N.-N Zheng, and H.-Y. Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):787–800, July 2003.

[87] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, September 2004.

[88] O. J. Woodford, P. H. S. Torr, I. D. Reid, and A. W. Fitzgibbon. Global stereo reconstruction under second order smoothness priors. In *Proceedings of CVPR*, 2008.

[89] Y. Ohta and T. Kanade. Stereo by intra- and inter- scanline search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(2):139–154, 1985.

[90] I.J. Cox, S.L. Hingorani, S.B. Rao, and B.M. Maggs. A Maximum Likelihood Stereo Algorithm. *Computer Vision and Image Understanding (CVIU)*, 63(3):542–567, 1996.

[91] C. Sun. Fast stereo matching using rectangular subregioning and 3D maximum-surface techniques. *International Journal of Computer Vision*, 47(3):99–177, 2002.

[92] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister. Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):492–504, March 2009.

[93] G. Egnal and R. P. Wildes. Detecting binocular half-occlusions: empirical comparisons of five approaches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1127–1133, August 2002.

[94] L. Wang, H. Jin, R. Yang, and M. Gong. Stereoscopic inpainting: Joint color and depth completion from stereo images. In *Proceedings of CVPR*, 2008.

[95] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *Proceedings of CVPR*, 2007.

[96] Y. Nakamura, T. Matsuura, K. Satoh, and Y. Ohta. Occlusion detectable stereoocclusion patterns in camera matrix. In *Proceedings of CVPR*, 1996.

[97] S. Baker, R. Szeliski, and P. Anandan. A layered approach to stereo reconstruction. In *Proceedings of CVPR*, Santa Barbara, CA, June 1998.

[98] U. R. Dhond and J. K. Aggarwal. Structure from stereo-a review. *IEEE Trans. on Systems, Man and Cybernetics*, 19(6):1489–1510, 1989.

[99] S.T. Barnard and M.A. Fischler. Computational Stereo. *Computer Surveys*, 14(4):553–572, 1982.

[100] R. Yang and M. Pollefeys. Multi-Resolution Real-Time Stereo on Commodity Graphics Hardware. In *Proceedings of CVPR*, pages 211–218, 2003.

[101] R. Yang, M. Pollefeys, H. Yang, and G. Welch. A Unified Approach to Real-Time, Multi-Resolution, Multi-Baseline 2D View Synthesis and 3D Depth Estimation using Commodity Graphics Hardware. *International Journal of Image and Graphics (IJIG)*, 2003.

[102] R. Collins. A Space-Sweep Approach to True Multi-Image Matching. In *Proceedings of CVPR*, pages 358–363, June 1996.

[103] Y. Boykov, O. Veksler, and R. Zabih. A Variable Window Approach to Early Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12), December 1998.

[104] S.B. Kang, R. Szeliski, and J. Chai. Handling Occlusions in Dense Multi-view Stereo. In *Proceedings of CVPR*, 2001.

[105] O. Veksler. Fast Variable Window for Stereo Correspondence using Integral Images. In *Proceedings of CVPR*, 2003.

[106] R.K. Gupta and S.-Y. Cho. A Correlation-Based Approach for Real-time Stereo Matching. In *Proceedings of the 6th International Conference on Advances in Visual Computing*, 2010.

[107] R.K. Gupta and S.-Y. Cho. Real-time stereo matching using adaptive binary window. In *Proceedings of 3DPVT*, 2010.

[108] W. Yu, T. Chen, F. Franchetti, and J.C. Hoe. High performance stereo vision designed for massively data parallel platforms. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(11):1509–1519, 2009.

[109] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. Dodgson. Real-time Spatiotemporal Stereo Matching using the Dual-Cross-Bilateral Grid. In *Proceedings of ECCV*, 2010.

[110] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast Cost-Volume Filtering for Visual Correspondence and Beyond. In *Proceedings of CVPR*, 2010.

[111] K. He, J. Sun, and X. Tang. Guided Image Filtering. In *Proceedings of ECCV*, 2010.

[112] M. Gong and Y.-H. Yang. Near real-time reliable stereo matching using programmable graphics hardware. In *Proceedings of CVPR*, pages 924–931, 2005.

[113] J.C. Kim, K.M. Lee, B.T. Choi, and S.U. Lee. A dense stereo matching using two-pass dynamic programming with generalized ground control points. In *Proceedings of CVPR*, pages 1075–1082, 2005.

[114] O. Veksler. Stereo correspondence by dynamic programming on a tree. In *Proceedings of CVPR*, pages 384–390, 2005.

[115] C. Lei, J. Selzer, and Y. Yang. Region-Tree based Stereo using Dynamic Programming Optimization. In *Proceedings of CVPR*, 2006.

[116] Q. Yang, L. Wang, R. Yang, S. Wang, M. Liao, and D. Nister. Real-Time Global Stereo Matching using Hierarchical Belief Propagation. In *Proceedings of BMVC*, 2006.

[117] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proceedings of ICCV*, 2001.

[118] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *Proceedings of ECCV*, 2002.

[119] A. S. Ogale and Y. Aloimonos. Shape and the stereo correspondence problem. *International Journal of Computer Vision*, 65(3):147–162, 2005.

[120] G. Li and S. W. Zucker. Surface geometric constraints for stereo in belief propagation. In *Proceedings of CVPR*, 2006.

[121] S. Birchfield and C. Tomasi. Multiway cut for stereo and motion with slanted surfaces. In *Proceedings of ICCV*, pages 489–495, 1999.

[122] J. Sun, Y. Li, S. B. Kang, and H.-Y. Shum. Symmetric stereo matching for occlusion handling. In *Proceedings of CVPR*, pages 399–406, 2005.

[123] L. Xu and J. Jia. Stereo matching: An outlier confidence approach. In *Proceedings of ECCV*, 2008.

[124] D. Gallup, J-M. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *Proceedings of CVPR*, 2010.

[125] Z. Zhang and Y. Shan. A progressive scheme for stereo matching. In *Europ. Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, July 2000.

[126] M. Lhuillier and L. Quan. Match propagation for image-based modeling and rendering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1140 – 1146, 2002.

[127] Y. Wei and L. Quan. Region-based progressive stereo matching. In *Proceedings of CVPR*, pages 106–113, 2004.

[128] L. Wang, H. Jin, and R. Yang. Search space reduction for mrf stereo. In *Proceedings of ECCV*, pages 576–588, 2008.

[129] A. Blake. Specular Stereo. In *Proc. 9th Int. Joint Conf. Artificial Intell. (IJCAI)*, volume 2, pages 973–976, 1985.

[130] G. Brelstaff and A. Blake. Detecting Specular Reflections using Lambertian Constraints. In *Proceedings of ICCV*, pages 297–302, 1988.

[131] D. Bhat and S. Nayar. Stereo in the presence of specular reflection. In *Proceedings of ICCV*, pages 1086–1092, 1995.

[132] Y. Li, S. Lin, H. Lu, S.B. Kang, and H-Y Shum. Multibaseline Stereo in the Presence of Specular Reflections. In *International Conference on Pattern Recognition*, pages 573–576, 2002.

[133] Stephen Lin, Yuanzhen Li, Sing Bing Kang, Xin Tong, and Heung-Yeung Shum. Diffuse-Specular Separation and Depth Recovery from Image Sequences. In *Proceedings of ECCV*, pages 210–224, 2002.

[134] R. Yang, M. Pollefeys, and G. Welch. Dealing with Textureless Regions and Specular Highlights—A Progressive Space Carving Scheme Using a Novel Photo-consistency Measure. In *Proceedings of ICCV*, pages 576–584, 2003.

[135] H. Jin, S. Soatto, and A. Yezzi. Multi-view Stereo beyond Lambert. In *Proceedings of CVPR*, pages 171–178, June 2003.

[136] S. Magda, T. Zickler, D. Kriegman, and P. Belhumeur. Beyond Lambert: Reconstructing Surfaces with Arbitrary BRDFs. In *Proceedings of ICCV*, pages 297–302, 2001.

[137] T. Zickler, P. N. Belhumeur, and D. J. Kriegman. Helmholtz stereopsis: exploiting reciprocity for surface reconstruction. In *Proceedings of ECCV*, pages 869–884, 2002.

[138] T. Zickler, P. N. Belhumeur, and D. J. Kriegman. Toward a Stratification of Helmholtz Stereopsis. In *Proceedings of CVPR*, pages 548–554, 2003.

[139] T. Zickler, J. Ho, D. J. Kriegman, J. Ponce, and P. N. Belhumeur. Binocular Helmholtz Stereopsis. In *Proceedings of ICCV*, pages 1411–1417, 2003.

[140] Adrien Treuille, Aaron Hertzmann, and Steven M. Seitz. Example-Based Stereo with General BRDFs. In *Proceedings of ECCV*, pages 457–469, 2004.

[141] Aaron Hertzmann and Steven M. Seitz. Example-Based Photometric Stereo: Shape Reconstruction with General, Varying BRDFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1254–1264, 2005.

[142] J. A. Jalkio, R. C. Kim, and S.K. Case. Three dimensional inspection using multistripe structured light. *Optical Engineering*, 24(6):996–974, 1985.

[143] Li Zhang, Brian Curless, and Steven M. Seitz. Spacetime Stereo: Shape Recovery for Dynamic Scenes. In *Proceedings of CVPR*, pages 367–374, 2003.

[144] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):296–302, 2005.

[145] B. Carrihill and R. Hummel. Experiments with the intensity ratio depth sensor. *Computer Vision, Graphics, and Image Processing*, 32:337–358, 1985.

[146] T. Miyasaka and K. Araki. Development of real time 3-d measurement system using intensity ratio method. In *Proceedings of Machine Vision and Three Dimensional Imaging System for Inspection and Metrology II, Intelligent System and Advanced Manufacturing*, 2001.

[147] L. Wolff and E. Angelopoulou. Three-dimensional stereo by photometric ratios. *Journal of the Optical Society of America A*, 11(11):3069–3078, 1994.

[148] B. Curless and M. Levoy. Better optical triangulation through spacetime analysis. In *Proceedings of International Conference on Computer Vision*, pages 987–994, 1995.

[149] A. Ansar, A. Castano, and L. Matthies. Enhanced Real-time Stereo Using Bilateral Filtering. In *Proceedings of 3DPVT*, 2004.

[150] C. Tomasi and R. Manduchi. Bilateral Filtering for Gray and Color Images. In *Proceedings of ICCV*, 1998.

[151] J. Kopf, M.F. Cohen, D. Lischinski, and M. Uyttendaele. Joint bilateral upsampling. In *Proceedings of Siggraph*, 2007.

[152] S. Paris, P. Kornprobst, J. Tumblin, and F. Durand. Bilateral filtering: Theory and applications. *Foundations and Trends in Computer Graphics and Vision*, 4(1):1–73, 2009.

[153] T.Q. Pham and L.J. van Vliet. Separable bilateral filtering for fast video preprocessing. In *Proceedings of IEEE International Conference on Multimedia and Expo*, 2005.

[154] D. Geiger, B. Ladendorf, and A. Yuille. Occlusions and Binocular Stereo. *International Journal of Computer Vision*, 14(3):211–226, 1995.

[155] Microsoft. DirectX, 2003. http://www.microsoft.com/windows/directx.

[156] J. Salmen, M. Schlipsing, J. Edelbrunner, S. Hegemann, and S. Lueke. Real-time stereo vision: Making more out of dynamic programming. In *Proceedings of CAIP*, 2009.

[157] Y. Deng and X. Lin. A fast line segment based dense stereo algorithm using tree dynamic programming. In *Proceedings of ECCV*, 2006.

[158] Q. Yang, C. Engels, and A. Akbarzadeh. Near Real-Time Stereo for Weakly-Textured Scenes. In *Proceedings of BMVC*, 2008.

[159] K. Zhang, J. Lu, G. Lafruit, R. Lauwereins, and L. Van Gool. Real-time accurate stereo with bitwise fast voting on cuda. In *Proceedings of ICCVW*, 2009.

[160] R. T. Collins. A space-sweep approach to true multi-image matching. In *Proceedings of CVPR*, pages 358–365, 1996.

[161] M. Guttmann, L. Wolf, and D. Cohen-Or. Semi-automatic stereo extraction from video footage. In *Proceedings of ICCV*, pages 358–365, 2009.

[162] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. *Proceedings of SIGGRAPH*, pages 689–694, August 2004.

[163] T.A. Davis. http://www.cise.ufl.edu/research/sparse/umfpack/.

[164] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[165] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, November 2001.

[166] D. Nister. A Minimal Solution to the Generalised 3-Point Pose Problem. In *Proceedings of CVPR*, pages 560–567, 2004.

[167] T. Mitsunaga and S. K. Nayar. Radiometric Self Calibration. In *Proceedings of CVPR*, volume 1, pages 380–387, 1999.

[168] P. E. Debevec and J. Malik. Recovering High Dynamic Range Radiance Maps from Photographs. *Proceedings of ACM Siggraph*, pages 369–378, 1997.

[169] K. Fujii, M. D. Grossberg, and S. K. Nayar. A Projector-Camera System with Real-Time Photometric Adaptation for Dynamic Environments. In *Proceedings of CVPR*, pages 814–821, 2005.

[170] R. Ng, R. Ramamoorthi, and P. Hanrahan. All-frequency shadows using non-linear wavelet lighting approximation. *ACM Transactions on Graphics (SIGGRAPH)*, 22:376–381, 2003.

[171] S. Mann and R. Picard. Being Undigital with Digital Cameras:Extending Dynamic Range by Combining Differently Exposed Pictures. In *Proceedings of IST's 48th Annual Conference*, pages 442–448, 1995.

[172] Vladimir Kolmogorov. An implementation of graph-cuts stereo. http://www.cs.cornell.edu/People/vnk/software.html, 2003.

[173] J. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *Proceedings of IJCAI*, 2001.

[174] H. Tao and H. Sawhney. Global matching criterion and color segmentation based stereo. In *Proceedings of IEEE Workshop on Applications of Computer Vision (WACV)*, pages 246–253, 2000.

[175] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of CVPR*, pages 519–526, 2006.

VITA

NAME

Liang Wang

DATE OF BIRTH

August 8, 1981

PLACE OF BIRTH

Anshan, People's Republic of China

EDUCATION

- July 2004: B.S. in Computer Science, Beihang University, Beijing, China

SELECTED PUBLICATIONS

- L. Wang, M. Gong, C. Zhang, R. Yang, C. Zhang, and Y. H. Yang, Automatic real-time video matting using time-of-flight camera and multichannel poisson equations, *International Journal of Computer Vision*, 97(1): 104-121, 2012.

- L. Wang and R. Yang, Global stereo matching leveraged by sparse ground control points, In *Proc. of the IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, June 2011.

- L. Wang, C. Zhang, R. Yang, and C. Zhang, TofCut: Towards robust real-time foreground extraction using a time-of-flight camera, In *Proc. of the fifth International Symposium on 3D Data Processing, Visualization and Transmission*, May 2010.

- L. Wang, H. Jin, and R. Yang, Search space reduction for mrf stereo, In *Proc. of Europ. Conf. on Computer Vision*, October 2008.

- L. Wang, H. Jin, R. Yang, and M. Gong, Stereoscopic inpainting: Joint color and depth completion from stereo images, In *Proc. of the IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, June 2008.

- L. Wang, R. Yang, and J. E. James, BRDF invariant stereo using light transport constancy, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(9): 1616-1626, 2007.

- L. Wang, M. Liao, M. Gong, R. Yang, and D. Nister, High-quality real-time stereo using adaptive cost aggregation and dynamic programming, In *Proc. of the third International Symposium on 3D Data Processing, Visualization and Transmission*, June 2006.

- L. Wang, M. Gong, M. Gong, and R. Yang, How far can we go with local optimization in real-time stereo matching, In *Proc. of the third International Symposium on 3D Data Processing, Visualization and Transmission*, June 2006.