The Dissertation Committee for Ming Yao
certifies that this is the approved version of the following dissertation:

High-quality Dense Stereo Vision for Whole Body Imaging and Obesity Assessment

Committee:

Bugao Xu, Supervisor

Andrew Dunn

Christopher Jolly

Mia Markey

Pengyu Ren

High-quality Dense Stereo Vision for Whole Body Imaging and Obesity Assessment

by

Ming Yao, B.S.; M.S.; M.S.T.A.T.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2015

Acknowledgments

I wish to thank the multitudes of people who helped me through out my journey of graduate studies. I extend my sincere gratitude to my advisor, Dr. Bugao Xu, for his insightful guidance and constant encouragement. I genuinely appreciated him for persistently pushing me to achieve my potential, and I am deeply indebted to his trust and understanding that helped me reach my academic success.

I am thankful to my dissertation committee, including Andrew Dunn, Christopher Jolly, Mia Markey and Pengyu Ren, for their valuable comments and critiques. I also own sincere thanks to Dr. Andrew Dunn, the Graduate Advisor of the Department of Biomedical Engineering, for extending my employment as a Teaching Assistant, so that I could afford my tuition and make my ends meet.

I am very grateful to various individuals who provided us with resources and assistance to conduct our experiments. Particularly, Philip Stanforth gave us access to the DEXA scanner at the Texas Fitness Institute, and helped us with the scans. Dr. Gregory Reece allowed us to check out the GoScan 3D handheld scanner, with which we can perform addition measurements. Wenbin Ouyang recruited volunteers and helped to coordinate their visits. Special thanks also go to the anonymous volunteers for their contribution to this study.

I would like to thank my present and former colleagues for their intellectual and emotional support: Jingjing Sun, Wenbin Ouyang, Xiaowen Guo, Jerry Yan, Xun Yao, Qingguang Li, Yan Wan and Wurong Yu. I appreciate all the wonderful moments we spent together. They made these last several years truly memorable. My heartfelt thanks also go to Mrs. Jo Frederick, the elder who met me every week and shared with me her

life stories. From her, I gained strength and courage that would benefit me in a lifelong journey dealing with challenges and difficulties.

I would like to take this opportunity to mention my parents for bearing, raising, and loving me. I miss them in China, and I am in deep sorrow for not being able to go back to visit them in the past 6 years. I wish I could have spent more time with them. Last but not least, my deepest thanks go to my wife, Zuyun, for her unconditional love. I cannot thank her enough for her patience and support in my pursuit of the Ph.D. degree.

High-quality Dense Stereo Vision for Whole Body Imaging and Obesity Assessment

Publication No. _____

Ming Yao, Ph.D.
The University of Texas at Austin, 2015

Supervisor: Bugao Xu

The prevalence of obesity has necessitated developing safe and convenient tools for timely assessing and monitoring this condition for a broad range of population. Three-dimensional (3D) body imaging has become a new mean for obesity assessment. Moreover, it generates body shape information that is meaningful for fitness, ergonomics, and personalized clothing. In the previous work of our lab, we developed a prototype active stereo vision system that demonstrated a potential to fulfill this goal. But the prototype required four computer projectors to cast artificial textures on the body which facilitate the stereo-matching on texture-deficient surfaces (e.g., skin). This decreases the mobility of the system when used to collect a large population data. In addition, the resolution of the generated 3D images is limited by both cameras and projectors available during the project. The study reported in this dissertation highlights our continued effort in improving the capability of 3D body imaging through simplified hardware for *passive* stereo and advanced depth computation techniques.

The system utilizes high-resolution single-lens reflex (SLR) cameras, which became widely available lately, and is configured in a two-stance design to image the front and

back surfaces of a person. A total of eight cameras are used to form four pairs of stereo units. Each unit covers a quarter of the body surface. The stereo units are individually calibrated with a specific pattern to determine cameras' intrinsic and extrinsic parameters for stereo matching. The global orientation and position of each stereo unit within a common world coordinate system is calculated through a 3D registration step. The stereo calibration and 3D registration procedures do not need to be repeated for a deployed system if the cameras' relative positions have not changed. This property contributes to the portability of the system, and tremendously alleviates the maintenance task. The image acquisition time is around two seconds for a whole-body capture. The system works in an indoor environment with a moderate ambient light.

Advanced stereo computation algorithms are developed by taking advantage of high-resolution images and by tackling the ambiguity problem in stereo matching. A multi-scale, coarse-to-fine matching framework is proposed to match large-scale textures at a low resolution and refine the matched results over higher resolutions. This matching strategy reduces the complexity of the computation and avoids ambiguous matching at the native resolution. The pixel-to-pixel stereo matching algorithm follows a classic, four-step strategy which consists of matching cost computation, cost aggregation, disparity computation and disparity refinement.

The system performance has been evaluated on mannequins and human subjects in comparison with other measurement methods. It was found that the geometrical measurements from reconstructed 3D body models, including body circumferences and whole volume, are highly repeatable and consistent with manual and other instrumental measurements ($CV < 0.1\%$, $R^2 > 0.99$). The agreement of percent body fat (%BF) estimation on human subjects between stereo and dual-energy X-ray absorptiometry (DEXA) was found to be improved over the previous active stereo system, and the limits of agreement

with 95% confidence were reduced by half. Our achieved %BF estimation agreement is among the lowest ones of other comparative studies with commercialized air displacement plethysmography (ADP) and DEXA. In practice, %BF estimation through a two-component model is sensitive to body volume measurement, and the estimation of lung volume could be a source of variation. Protocols for this type of measurement should still be created with an awareness of this factor.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Obesity has been a growing health concern in the United States (U.S.), and many other countries. Obesity increases the likelihood of various diseases, particularly cardiovascular disease, type II diabetes, hypertension, osteoarthritis, and certain types of cancer [3, 4]. The World Health Organization (WHO) describes obesity as one of the most apparent, yet most neglected, public health problems that threaten to overwhelm both more and less developed countries [5]. The prevalence of obesity has made it necessary to develop a safe, reliable and convenient tool for efficiently assessing and monitoring this condition in the public health. WHO has accepted a Body Mass Index (BMI) as a quantitative scale to classify the severity of obesity.

BMI is calculated by dividing a person's weight in kilograms (kg) by the square of the persons height in meters (m). A person with BMI index higher than 25.0 kg/m^2 is considered as abnormal, while a BMI index greater than 30.0 kg/m^2 is considered as obese. Various techniques have been developed to assist BMI-based obesity assessment. For instance, densitometry methods including underwater weighting [6] and air displacement plethysmography [7] were accepted as standard methods for body density estimate, but their accuracy in estimating the body fat percentage was questioned because of its two-component model that only included fat and fat-free mass. In addition, BMI has become controversial in medical assessments because BMI was originally proposed as a simple mean of classifying sedentary individuals whose body compositions deviate from the av-

erage [8]. It fails to take into account age, body shape or body composition, all crucial factors in obesity designation and evaluation of associated health risks for the individual. In addition to the fact that BMI only correlates to the overall percent body fat, the distribution of fat is also an important factor in assessing health risk. It is believed that the accumulation of fat in abdominal section is associated with increased risk of cardiovascular disease and insulin resistance [9, 10]. Thus, with the same BMI, individuals with most of their weight above the waist line ("apple-shaped") have a higher risk of metabolic disorder than individuals with most of their weight below the waist line ("pear-shaped"). In this case, waist circumference gives a better prediction of the individual's health condition than BMI [11].

A whole body 3D imaging device is an ideal tool for obesity research. Because such a device captures the 3D profile of a person's exterior surface, so that computations can be used to calculate the volumes and the dimensions of various body parts. Such a device, commonly referred to as a body scanner, captures the surface profile through non-contact optical techniques. With these 3D surface data, a digital model representing the shape of the scanned body can be generated. Total and regional body volumes, as well as other measurements that are helpful in evaluating a person's fitness level, such as various circumferences, regional thicknesses and breadths can all be readily obtained from the 3D digital model.

1.2 Goals and Contributions

Popular technologies that are utilized in 3D imaging devices involve laserline triangulation, coded structured light, and stereo vision. A laserline scanner usually provides good resolution but it requires mechanical devices to "scan" the subject, thus the total capture time is limited by its scanning speed, and may require regular calibration due to the

movement of the laser projector. Coded structured light and stereo vision are both static technologies. The former uses active lighting to create multiple sets of light patterns for depth sensing, in which the light patterns may last for a few seconds. A stereo vision device captures 3D scene by taking stereo pictures, hence the 3D capture is the fastest among all three types of imaging techniques, although the depth computation in a stereo device is the most sophisticated among all.

With 3D body imaging techniques maturing, a complete system dedicated to 3D anthropometry for body composition assessment with convenience to use and good accuracy is still a challenge. The reason is multifold. First, a 3D imaging system that is accurate in measurement and robust in field use usually requires active lighting, such as laser or digital projector, for depth sensing. Their high price and bulkiness prevent them to be massively deployed, thus limited its accessibility to the general public. Second, most of the body imaging systems that are commercially available are limited to the use of clothing and animation industries [12, 13], the potential and value of this type of system have yet been widely recognized by body composition researchers and health care providers. Finally, software systems capable of body composition assessment are rarely available.

A previous project of developing a 3D anthropometry system based on stereo vision technology has been conducted [14]. It concludes that this technology is ready for practical use as a body measurement system dedicated to body composition assessment. In addition to volumetric measurements, other physical measurements and indirect measurement that are meaningful for body composition and health risk assessment, such as waist-hip ratio that is and indicator of central obesity, can all be obtained from the reconstructed 3D digital model more accurately and efficiently than from a tape measure. However, the practical application of this developed system is limited by the low resolution and specialized hardware that were designed for general computer vision tasks a decade ago.

In the previous system, a total of four pairs of monochronical video cameras are used and set apart at 12 feet away for whole body coverage. A digital projector is required in each stereo unit to add artificial texture on scanned surface to assist stereo matching. Having more hardware components increases the efforts in deploying such a system into the test field. Dedicated hardware component is also less flexible in replacement and upgrade.

The work reported in this dissertation is a continued effort in improving the capability of a stereo vision system by the utilizing high resolution consumer-grade cameras, and developing the state of the art stereo matching algorithms for 3D scene recovery. On the hardware side, active lighting devices have been eliminated in the new system, thanks to the higher resolution of stereo images in which the rich of skin textures, other than artificial pattern, provide adequate information for stereo matching. The new cameras have a larger viewing angle so that cameras can be placed closer to a subject, effectively reducing the space to deploy this system while maintaining the same coverage for imaging. On the software side, the developed stereo matching algorithm takes advantage of the highly detailed, chromatic stereo images, and incorporates sophisticated design concepts for robust stereo matching. As demonstrated by our system-wise evaluation, the proposed stereo matching algorithm together with dedicated surface reconstruction push the boundaries of stereo imaging to a new level.

1.3 Structure of the Dissertation

The remainder of the dissertation is divided into eight chapters. Chapter 2 provides background for this research. Current body composition techniques for body fat assessment are explored. Then the advantages and potential values of 3D body imaging system for body composition research are discussed.

Chapter 3 presents the design of a 3D body imaging system after briefly reviewing

current major techniques of 3D imaging. The principle of stereo vision is described, and related work in developing stereo vision based depth estimation algorithms are discussed.

Chapter 4 introduces the framework design of a stereo vision system being used for body imaging purpose. The hardware setup and system configuration are presented in this chapter. A revised camera calibration approach and 3D registration are also proposed. An accurate camera calibration is the foundation of stereo vision based depth estimation. The proposed camera calibration technique improves optical distortion correction introduced by camera lenses, and reduces the error in merging 3D data captured by multiple cameras that are positioned at different locations. A multi-scale stereo matching framework is also documented in this chapter. It provides an overlook of the stereo processing pipeline.

Chapter 5 and 6 documents the stereo matching algorithm which is the major challenge in developing such a system. Chapter 5 focuses on the computation of feature-matching costs between left and right images, and proposes a novel hybrid cost function combined with a sophisticated yet computationally efficient cost aggregation method to improve algorithm robustness in areas that have low texture and contrast. Chapter 6 deals with the optimization problem in assigning the correct depth value to each pixel of the stereo images based on the computed and aggregated matching costs from the previous steps.

Chapter 7 describes the processes that are essential for recovering a high quality 3D model from disparity data. These processes include sub-pixel refinement and surface reconstruction. Sub-pixel refinement produces smooth surface with noise suppression and geometrical detail enhancement. 3D surface reconstruction converts the dense 3D points into more manageable and efficient data representation for measurement and display.

In Chapter 8, the performance evaluation of the developed stereo matching algorithms is reported. System-wise performance is evaluated with volunteers by comparing

the measurements of various circumferences and volumes computed from 3D models to tape measurements, and data generated by other 3D scanners. Body fat percentage computed from our system is compared to dual-energy X-ray absorptiometry. A pilot study is carried out to test the accuracy and precision of the system.

Chapter 9 concludes the work and discusses the possible improvements for future study.

Chapter 2

Background

2.1 Introduction

Obesity is a medical disorder that is caused by excess body fat accumulated over time. Thus, it is commonly accepted that body fat assessment is the primary focus of body composition research. Body fat assessment plays an important role in weight management and health risk evaluation. In this chapter, we first give a brief overview on overweight and obesity and their associated threat to health complications. Then we review current methods and technologies of body fat assessment. Finally, we propose stereo vision as an potential alternative to accurate 3D body imaging for body fat assessment.

2.2 Overweight and Obesity

2.2.1 Health Risks from Rising Obesity

The major health risks that are associated with obesity are various chronic diseases including type II diabetes, hypertension, cardiovascular diseases, and certain types of cancer [3, 4]. For instance, every increment of BMI by 5 Kg/m^2 raises a man's risk of esophageal cancer by 52% and colon cancer by 24%, and in women, endometrial cancer by 59% and gall bladder cancer by 59% [15]. Evidence also indicates that excess body weight leads to non-fatal but disabling disorders such as osteoarthritis [16]. Excess body weight also contributes to many additional medical conditions, e.g., benign prostate hypertrophy [17], infertility [18], asthma [19, 20], and sleep apnoea [2].

Overweight individuals are at higher risk to to have hypertension and hyperlipi-

Figure 2.1: Obesity and its association to metabolic disorder and mortality. Reprinted from [1]

demia, which can lead to coronary artery disease and stroke. It has been estimated that more than 85% of hypertension cases arise in individuals with overweight or obesity. Obesity, especially central or visceral obesity, is strongly associated with increased insulin resistance and glucose intolerance. Visceral fat, which often wraps deep around the belly, plays a role in the metabolic syndromes that increase the risk of type II diabetes and cardiovascular disease. According to a recent finding [1], an estimated 21% of U.S. adults who have an "obese BMI" are metabolically unhealthy, while only 10% of U.S. obese adults are metabolically healthy. Figure 2.1 lists the metabolic disorders and mortality among individuals with normal and obese BMI.

Currently the prevalence of obesity in many population is greater at a much younger

age than in previous generations, this trend in obesity projects a growth in the proportion of the population suffering from chronic disabilities, and presents a potential threat to the increase in life expectancy that is achieved by medical and public health advances during the past century [21].

2.2.2 Threat to Population Health

The increased prevalence of overweight and obesity has become a worldwide health concern [22]. The obesity epidemic seemed to grow almost concurrently in most developed countries in the 1970s and 1980s [23]. Since then, other countries have joined the global trend in obesity prevalence in adults and children [24]. By 2008, an estimated 1.46 billion adults globally were overweight and 502 million adults were obese. Furthermore, an estimated 170 million children (age < 18) globally were classified as overweight or obese [25]. Despite signs of stabilization in some populations [17, 26], the negative effects of consistently high prevalence of obesity are extensive: societies are burdened by premature mortality, morbidity associated with many chronic disorders, and degrades of health-related quality of life.

Figure 2.2 shows the prevalence of overweight in adults and children in selected countries [2]. The U.S. and the U.K. have had the striking increases in the percentage of their populations with BMI in overweight and obese ranges. If such trend were to continue, it is estimated that about three out of four Americans and seven out of ten British people will be overweight or obese by 2020 [2].

2.2.3 Economic Impact

In addition to many chronic and acute health disorders incurred by excess body weight, a society is burdened by substantial cost in improving the health-related quality of life of its affected people, notably from increased health care costs and lost productivity.

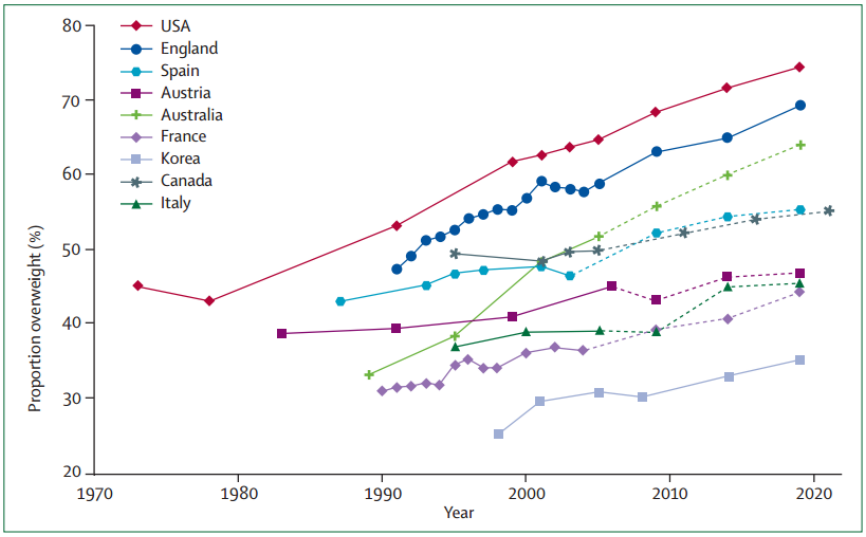


Figure 2.2: Past and projected prevalence of overweight (BMI $\geq$ 25 Kg/m^2). Reprinted from [2].

The medical costs of the care for obesity include various resources dedicated to managing obesity-related conditions, such as the costs incurred by excess use of ambulatory care, hospitalisation, drugs, radiological or laboratory tests, and long term care. In an review of the economic burden of obesity worldwide [18], it was found that obesity accounted for 0.7-2.8% of a country's total health care cost, and that obese individuals had medical costs 30% more than those with normal weight. The combination of developing obesity prevalence and the increased spending on obese people has been estimated to account for 27% of the growth in the U.S. health care spending between 1987 and 2001 [20]. This number is projected to double every decade to account for 16-18% of total health care spending by 2030 [19]. Another recent study [27] reported that, compared with normal weight individuals, obese patient incur 46% increased inpatient costs, 27% more physician visits and outpatient costs, and 80% increased spending on prescription drugs. The annual extra medical costs of obesity in the U.S. were estimated as $75 billion in 2003 [28] and accounted for 4-7% of total health care expenditure [29].

Besides the medical costs, society is also burdened by indirect costs from obesity as a direct result of decreased years of disability-free life, increased mortality before retirement, early retirement, disability pensions, and work absenteeism or reduced productivity. Several studies suggest that the monetary cost from lost productivity is several times higher than medical costs [30–32]. For U.S. employee, it was reported in [32] that annual missed workdays ranged from 0.5 more days for men who were overweight to 5.9 more days for men who were class III obese (BMI $\geq$ 40 Kg/m^2) than men of normal weight. Moreover, the estimated annual cost from presenteeism in men who were very obese was the equivalent of 1 month of lost productivity and cost employers $3800 per year.

2.3 Overview of Body Composition Assessment

Increased body fat is usually accompanied by increased total body mass, so BMI has been one of the most important indices to measure the relative weight of body mass, and has been commonly used to identify obesity. However, it was not originally invented as an index of obesity but is now widely employed as such in epidemiologic studies, because it can be easily measured. Even though, the accuracy of BMI as a body composition marker is controversial [33–35]. Some research data pointed out that BMI inadequately predicts percentage of body fat [33, 36], whereas others suggested that BMI may be useful to predict body fat indexed to height but not to predict percentage of body fat [37]. The inaccuracy of BMI serving as an obesity indicator lies in the inability of the BMI to distinguish body fat from muscle, bone and other non-fat body mass. In addition, the relationship between BMI and body fatness varies in gender, age and racial group. Furthermore, a consensus report by WHO warned researches that BMI must be interpreted carefully to avoid confusing muscularity with obesity [8]. Therefore, direct or indirect measurements of body fat could provide a significant improvement towards evaluation and diagnosis of obesity.

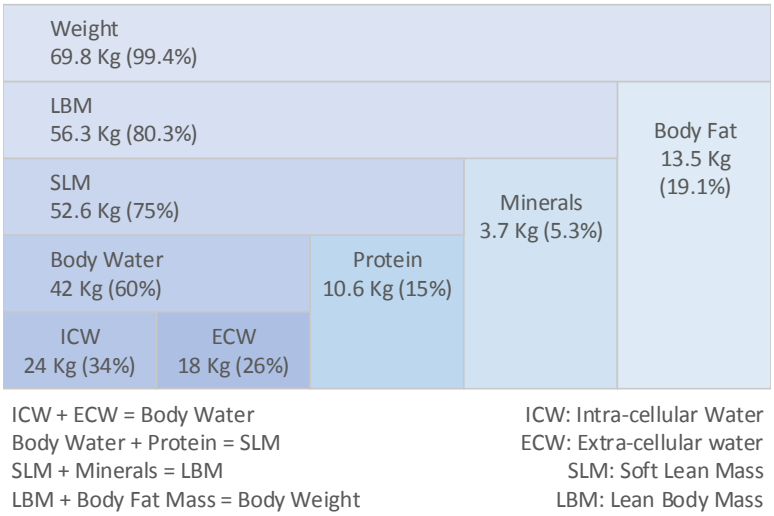


Figure 2.3: Illustration of body composition at molecular level.

2.3.1 Body Composition Models

In body composition research, a five-level model [38] was developed to provide a structural framework for studying human body composition. These five levels are atomic, molecular, cellular, tissue-organ and whole body. Among these five levels, the molecular level is most important because various methods for body composition assessment are derived on this level. Figure 2.3 illustrates the molecular components of a body at this level. The major components at this level include water, protein, mineral and fat. A simplified two-component model that partitions the body into fat mass (FM) and fat-free mass (FFM) is the most widely used approach to estimate body composition in adults. The lean body that contributes to the fat-free mass includes protein, mineral, and total body water (TBW). Protein is a main element of muscles and mineral is found mostly in bones. Body water consists of intra-cellular and extra-cellular water. Intra-cellular water (ICW) gives cell volume and extra-cellular water (ECW) is composed of blood, lymph, etc.

Within the two-component model, the proportion of FFM as water, protein and mineral is assumed to be constant. Then the percentage of body fat (%BF) can be calculated

by

$$\%BF = \left(\frac{C_{FM}}{D_b} - C_{FFM} \right) \times 100, \tag{2.1}$$

where C_{FM} and C_{FFM} are constants derived from fat mass density (D_{FM}) and fat-free mass density (D_{FFM}), D_b is the measured body density. The D_{FM} is relatively stable, because fat cells in humans are composed almost entirely of pure triglycerides with an average density of about 0.9 Kg/L. Most modern body composition laboratories today use the value of 1.1 Kg/L for the density of the FFM, with its theoretical composition of 72% water (density = 0.993 Kg/L), 21% protein (density = 1.34 Kg/L) and 7% mineral (density = 3.0 Kg/L) by weight. Different forms of (2.1) exist due to the slightly different composition for FFM used. Commonly used %BF calculation are Siri's formula [39]:

$$\%BF = \left(\frac{4.95}{D_b} - 4.50 \right) \times 100, \tag{2.2}$$

and Brozek's formula [40]:

$$\%BF = \left(\frac{4.57}{D_b} - 4.142 \right) \times 100. \tag{2.3}$$

Body composition estimates based on two-components model will be inaccurate when the assumptions that forms the basis for the model are not met. This may occur systematically with characteristics such as aging, pregnancy, weight reduction in obese people, athletic fitness, and in various disease states. This model was not suggested to be used with infants and young children as the proportions of FFM as water, protein, and mineral are constantly changing with growth. Some researchers considered four-component model a more accurate measure of body composition. The four-component model involves the measurement of body mass (or weight), total body volume, total body water, and bone mineral. However, specialized laboratory equipment is required to conduct the measurement, preventing its availability to many clinicians and researchers.

2.3.2 Underwater Weighting and Air Displacement Plethysmography

Both underwater weighting (UWW) and whole body air displacement plethysmography (ADP) are based on two-component body composition model. The goal of these methods is to estimate the total body volume in order to calculate the average body density. In the UWW method, a person is completely submerged in water and the volume of displaced water can be calculated by measuring weight difference before and after submerging in the water. Estimation of %BF from UWW has long been considered to be the best method available [39], especially in consideration of the cost and simplicity of the equipment.

An air displacement plethysmography device, commercially available as the BodPod (Life Measurement Instrument, Concord, CA), presents an alternative to UWW. ADP uses the same principles as the UWW, but introduces a densitometric method that is based on air displacement rather than on water immersion [7]. The measurement relies on Boyle's law which states that when temperature stays unchanged, air will increase its volume proportionally to decrease in pressure [41]. Reliability of APD method was found to be high for %BF and body density in adults [42, 43]. ADP offers several advantages over the UWW, including a fast, comfortable and safe measurement process, and is accommodating to various subject types, such as children, elderly and obese individuals.

2.3.3 Bioelectrical Impedance Analysis

Bioelectrical impedance analysis (BIA) measures the impedance or resistance to a small electrical current as it travels through the body's water with dissolved electrolytes. It assumes that 73% of the body's FFM is water, thus an estimate of total body FFM can be acquired from TBW. Single-frequency BIA is the most common use for assessing TBW and FFM, but its ability in distinguishing the distribution of ICW and ECW is limited.

The advantages of BIA include its portability and ease of use, relative low cost and safety, which make it attractive for large-scale studies. The accuracy of BIA is also affected by gender, age, health condition, race or ethnicity [44], and level of fitness, in which TBW and relative ECW are greater in obese individuals [45].

2.3.4 Dual-energy X-ray Absorptiometry

Dual-energy X-ray absorptiometry (DEXA) utilizes a three-component model of bone, lean soft tissue, and fat to estimate body composition [46, 47]. It measures X-ray photon energy attenuation through different types of body components. The radiation exposure from a whole body DEXA scan ranges from 0.04 to 0.86 mrem [48, 49], which is equivalent to between 1 and 10% of a chest radiograph. Thus DEXA technique is accepted as a noninvasive measurement method that can be applied in humans of all ages. The advantages of DEXA include good accuracy and reproducibility, and it provides regional assessment of body composition and nutritional status in disease states. Estimation of body fat by DEXA was found to be strongly related to estimation via a four-component model through criterion method in 78 subjects [50]. No significant difference was found between these methods. The correlation between DEXA and UWW was strong for both man and women following water loss and gain [51]. However, in a study of 110 men and 225 women, DEXA was shown to overestimate body fat in men and underestimate in women [52], as compared to UWW.

Assumptions associated with DEXA in %BF estimation include: the assumed constant attenuation of fat and bone, the uniform attenuation model across regional thickness (e.g., chest, leg and arm) on soft-tissue estimates, and the uniform fat content in that fat in analyzed area (nonbone-containing area) is comparable with the fat in unanalyzed area (bone-containing area) [53]. The limitation associated with these assumptions, when not met, includes errors in the estimation of fat mass, lean and bone in both regional and whole

body values. Estimate of fat mass may also be influenced by a person's trunk thickness in that error increases as the individual's trunk thickness increases. In longitudinal studies of persons who undergo significant changes in body composition, DEXA measures can be biased [54].

2.3.5 Computed Tomography and Magnetic Resonance Imaging

X-ray computed tomography (CT) and magnetic resonance imaging (MRI) allow the estimation of adipose tissue, skeletal muscle, and other internal tissues and organs. Their primary application has been in quantifying the distribution of adipose tissue into visceral, subcutaneous, and more recently intermuscular depots [55]. The application of these depots may help health care provider evaluate cardiovascular disease risk [56]. A further application of MRI has been in dissecting the FFM compartment for the quantification of specific high metabolic rate in organs *in vivo* (e.g., liver, kidneys, heart, spleen, pancreas, and brain) to improve our understanding of resting energy expenditure [57].

The limitations of CT and MRI include high costs owing to equipment and large data processing requirements, and individuals with large body size cannot fit within field-of-view. Neither CT nor MRI is capable of accommodating persons with BMI $> 40\ \mathrm{Kg/m^2}$. The field-of-view for most MRI scanners is limited to 48×48 cm. This becomes a significant limitation when there is a need to image persons before treatments such as bariatric surgery, which typically involves persons with BMI greater than $40\ \mathrm{Kg/m^2}$.

2.3.6 3D Photonic Scanner

The need for accurate measurement of body shape and body dimensions in a cost-effective manner has resulted in the development and application of a range of digitized optical methods to capture three-dimensional photonic images of an individual. The estimation of %BF through 3D photonic scanner (3DPS) is based on two-component body

composition model, with goals to generate values for total and regional body volumes and dimensions. Other measurements, such as height, various circumferences, segment lengths, and surface areas can all be calculated from the 3D model by using dedicated algorithms. This technique provides a more efficient, more objective and more comprehensive way for body dimension measurement than conventional tape anthropometry. Further, the 3D surface acquisition is non-contact and non-invasive, the 3D model is reusable so new measurements can be extracted whenever needed. The accuracy of a laser based 3DPS for the measurement of body volume, circumference, lengths and %BF compared with UWW and tape measures was reported in [58]. The 3DPS systems offer a novel approach for epidemiologic research into associations between body shape and health risks and outcome.

2.4 3D Body Imaging for Body Composition Estimation

Table 2.1 summarizes the advantages and disadvantages of various methods for body composition estimation. An ideal solution to acquire %BF is a system that is accurate and reliable in measurement, cost-effective in operation, and can be easily deployed into test field. Our proposed technique falls into the category of 3D photonic scanner that captures the 3D body surface and computes %BF through a two-component body model.

The capturing of highly detailed 3D surface of human body is of interest in multiple disciplines, including artistic 3D animation, customized fashion design, clinical use, obesity research, or for fitness purpose. The demand of capturing a high-quality 3D surface has intrigued extensive research in 3D acquisition technology, optical device design, and rendering techniques. However, capturing an accurate 3D body with minimum error and low hardware cost is still a challenge for computer vision and graphics researchers. This is because the human body if rich in surface geometrical features that requires sophisticated capturing and modeling techniques to reconstruct the realistic representation.

Table 2.1: The primary measurements, advantages and disadvantages of body composition estimation methods in humans.

Method	Direct measurements	Advantages	Disadvantages
UWW	Weight difference before and after submerging	Inexpensive and accurate	Uncomfortable, unaccommodating
ADP	Total body volume	Relatively high accuracy, fast	Reduced accuracy for individuals in disease states, expensive equipment
BIA	Total body water	Inexpensive, portable, simple, safe, quick	Population specific, poor accuracy in individuals and groups
DEXA	Total and regional body fat, lean mass and bone mineral content	Accurate, especially for limbs	Bias in body size and fitness, expensive equipment
CT	Specific regional bone density	High accuracy and reproducibility	High-radiation exposure, expensive equipment
MRI	Total and regional adipose tissue , skeletal muscle, organs, lipid content in liver and muscle	High accuracy and reproducibility for whole body and regional adipose tissue and skeletal muscle	Expensive
3DPS	Total and regional body volume	Can accommodate extremely obese persons, easy to use, suitable for both research and clinical applications	Technology is maturing but few scanners are available

UWW, underwater weighting; ADP, air displacement plythsmography; BIA, bioelectrical impedance analysis; DEXA, dual-energy X-ray absorptimetry; CT, computed tomography; MRI, magnetic resonance imaging; 3DPS, 3Dphotonic scanning.

Current 3D body imaging technologies can be broadly categorized into two classes: those with active lighting and those are passive in capturing. Popular 3D imaging techniques based on active lighting are laser, structured light, and gradient-based illumination. These technologies are usually robust because the depth information is computed from the augmented light pattern that is purposefully projected onto the imaged surface, thus they are insensitive to the native color and texture properties of the surface. However, they require purpose-built illumination devices and often utilize time-multiplexing. Both laser scanner and structured light imaging devices are based on profile measurements sampled across the imaged surfaced. A series of images are captured when the light pattern shifts overtime. Gradient illumination also requires multiple captures when the lights are projected from several directions. On the other hand, stereo vision based passive imaging does not require artificial lighting, therefore is more flexible in configuration and requires

less effort in deployment. Stereo imaging is done through one single-shot, thus is more convenient for the imaged individual to remain steady during the capture. Technical details of each 3D acquisition method as well as their applications are briefly reviewed in the following subsections.

2.4.1 3D Capturing Techniques

2.4.1.1 Laser Scanning

There are two types of technologies available for a laser scanner to detect depth, time-of-flight (ToF) and triangulation. The principle behind a ToF camera is the laser range sensor that resolves distance based on the recorded time of the round-trip of a pulse of light, with the known speed of light. A laser is usually used to emit a pulse of light and the amount of time before the reflected light is seen by a detector is measured. The accuracy of a ToF 3D device depends on the precision of the measurement of round-trip time. As light has a speed of approximately 3×10^8 meters per second, it takes 3.3 picoseconds to travel 1 millimeter. Since ToF is point based measurement, a scanner has to scan the field of view one point at a time by changing the laser direction to scan different points. The change of view direction is usually done by a rotating mirror because it can be operated very fast and with great accuracy. A typical ToF laser scanner can perform distance measurement at 10,000–100,000 points per second.

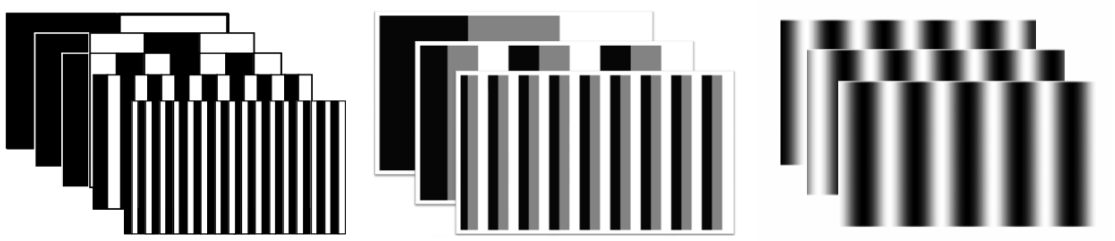
With respect to ToF 3D scanner, the triangulation laser projects a laser spot on the scanned surface and observes the spot through a camera. Depending on how far away the laser reaches to a surface, the laser spot appears at different places in the camera's image plane. This technique is called triangulation because the laser spot on the scanned surface, the camera and the laser projector form a triangle. The length of one side of the triangle, e.g., the distance between the camera and the laser projector is known. The angle of the laser projector corner is also known. The angle of the camera corner can be determined

by looking at the location of the laser spot's image on camera's photonic sensor. These three pieces of information fully determine the shape and size of the triangle and gives the location of the laser spot corner of the triangle. In order to scan a 2D surface, a laser beam projector instead of laser point projector is used to sweep across the imaged surface.

ToF and triangulation scanners each have strengths and weaknesses that make them suitable for different applications. ToF devices are capable of measuring very long distances, typically on the order of kilometers. Triangulation devices usually have a limited range of operation which is at a few meters, but their accuracy is relatively high and can achieve to a resolution on the order of tens of micrometers. In most cases, a low resolution laser scan can finish within less than a second. But high resolution scans, which may require millions of samples, can take several seconds. This leads to distortion from motion. Since each point is sampled at a different time, any motion in the subject, or the scanner, will distort the collected data. Recently, there has been research on compensating for distortion from small amounts of vibration [59] and distortions due to motion or rotation [60]. However, motion correction for body scanning still remain unsolved, due to the difficulty in estimating the body movement during a full body scan.

2.4.1.2 Structured Light

Structured light 3D scanner projects a light pattern on the scanned surface and observes the deformation of the pattern from a camera that slightly offsets from the pattern projector [61]. The principle of depth estimation in structured light is similar to the laser system and is also based on triangulation. Numerous techniques for surface imaging by structured light are currently available. In a more general sense, all techniques can be classified into two categories, sequential (multiple-shot) or single-shot. When the imaged 3D target is static and the image acquisition does not impose stringent constraint on the capturing time, multiple-shot techniques can be used and may often result in more reliable



(a) (b) (c)

Figure 2.4: Patterns used in structured light 3D imaging. (a) Sequential binary-coded patterns; (b) Gray-level coding; (c) Sinusoidal fringe pattern for phase shift 3D imaging.

and accurate results. On the other hand, if an imaging task requires the capture of an target that is in motion, single-shot techniques have to be used to acquire a snapshot of the image at a particular time instance.

Popular patterns used in multiple-shot capture include binary patterns, gray level patterns, and phase shift. The binary pattern (or, binary coding) [62, 63] uses black and white stripes to form a sequence of projections, such that each point on the imaged surface possesses a unique binary code that differs from any other codes of different points. In general, N pattern frames can code 2^N stripes. In other words, the horizontal resolution of the capture is determined by the finest stripes in the pattern series. Figure 2.4a shows a simplified 5-frame projection pattern.

To effectively reduce the number of patterns that are needed to obtain a high-resolution 3D image, gray-level patterns are developed [64]. For example, one can use M distinct levels of intensity, instead of only two in the binary code, to produce unique coding of the projection patterns. Figure 2.4b shows an example of 3-frame pattern with three levels of gray scale.

Binary coded pattern and gray-level coded pattern both use discrete step color patterns. On the other hand, phase shift is a type of fringe projection method which uses *continuously* colored patterns for imaging [65, 66]. A set of sinusoidal patterns is projected

onto the object surface (Figure 2.4c). The intensity of each pixel (x, y) in the images of three projected sinusoidal patterns are described as

$$\begin{aligned} I_1(x,y) &= I_0(x,y) + A\cos[\phi(x,y) - \theta], \\ I_2(x,y) &= I_0(x,y) + A\cos[\phi(x,y)], \\ I_3(x,y) &= I_0(x,y) + A\cos[\phi(x,y) + \theta], \end{aligned} \tag{2.4}$$

where $I_1(x,y)$, $I_2(x,y)$, and $I_3(x,y)$ are the intensities of three fringe patterns, $I_0(x,y)$ is the DC component (background), A is the modulation signal amplitude, $\phi(x,y)$ is the *unwrapped*, i.e., continuous and monotonically increasing, phase that we are looking for, and θ is a constant phase-shift between the consecutive pattern frames. Once the phase shifted images are captured, a process called phase unwrapping is used to convert the relative, wrapped phase $\phi'(x,y)$, $\phi' \in [0, 2\pi)$, to the absolute, unwrapped phase $\phi(x,y)$.

The wrapped phase information $\phi'(x,y)$ can be retrieved from the intensities in the three fringe pattern images[1]:

$$\phi'(x,y) = \arctan\left[\sqrt{3} \times \frac{I_1(x,y) - I_3(x,y)}{2 \times I_2(x,y) - I_1(x,y) - I_3(x,y)}\right]. \tag{2.5}$$

The discontinuity of the arc tangent function at 2π can be removed by adding or subtracting multiples of 2π on the ϕ' value, this unwraps the relative phase and generates the absolute phase value at pixel (x,y). The 3D coordinates can be calculated based on the difference of the absolute phase value between measured phase and the phase from a reference plane [62].

2.4.1.3 Stereo Vision

Stereo vision works similarly in concept to human binocular vision. In a traditional passive stereo setup, two cameras placed horizontally apart from one another are

[1]The standard `arctan()` function produces results in the range $(-\pi/2, \pi/2)$. In practice, an alternative function, the `atan2()` should be used instead to produce results in the range $(-\pi, \pi]$, which can be mapped to $[0, 2\pi)$ by adding 2π to negative results.

used to obtain two different views of the same scene. By comparing these two images, the relative depth information can be obtained in the form of disparities, which are inversely proportional to the distances from the camera to the imaged objects. The primary computation involved in a stereo vision system is a process called stereo matching, which finds the pixel-wise correspondences between left and right images. Various algorithms have been developed to improve the accuracy and robustness of stereo matching [67, 68], because pixel correspondences would be weak in texture-less regions, in images where great amount of noise presents, and between matched feature pixels whose colors are not consistent due to different gains and biases used in image sensors. Multiple variant of global optimization were proposed to reduce pixel-wise matching error, and to advance the state of the art of stereo matching [67, 69].

Compared to laser scanner and structured light, stereo vision system captures a 3D scene in one shot, which is as quickly as taking a picture on the camera. The capability of the fast scene capture makes stereo vision technology a great solution for body imaging, because it is difficult to have the scanned subject to remain static and to avoid any involuntary body movement for a period of time. Stereo vision system is passive in nature, it does not require any artificial lighting. Furthermore, since high resolution cameras are becoming more affordable, the total hardware cost in building a stereo vision system is getting lower. On the other hand, the primary disadvantage associated with stereo vision is the complexity of the algorithm it uses to recover the 3D scene. The quality of reconstructed 3D scene largely depends on the richness of textures on the imaged surface. The computational complexity of the 3D depth calculation in stereo vision is directly proportional to the size of stereo pictures and the depth of the scene. As a result, stereo vision faces great challenge in real-time application, and sometime requires parallel processing to fully utilize the computing power of modern massively paralleled computation infrastructure.

2.4.2 3D Acquisition Systems

3D body scanners are transforming the ability to accurately measure a person's body size, shape, and skin surface area. Originally developed primarily for the clothing and movie industries, 3D scanner's noninvasive nature and ease of use make them appealing for broad clinical applications and large scale epidemiological surveys [70]. Research on building accurate and reliable methods to capture human bodies began in the middle and late 1980s.

Today, numerous body scanners have emerged on the market, the majority of which are based on laser scanning and structured light technologies, primarily because of their robustness benefited from active lighting in depth sensing. According to a review [71] of body scanners conducted in 2007, there were over 50 companies around the world that are developing and producing systems for 3D measurement of human body at the time of writing. Systems and products for body scanning were developed and produced in three regions: North America, Europe and Asia. The majority of structured light systems were developed in Europe, mainly in Germany and UK. Whereas laser scanning systems were developed and produced in North America and Asia.

Cyberware developed the earliest body scanners for face scanning [72]. The scanning system is composed of a laser line projector and a camera. It rotates 360 degrees around the subject's head to capture a 3D image. The system was used for visual effects in the movie *The Abyss*, produced by 20th Century Fox in 1989, to digitize the face of two actors. Depending on the body parts to be measured, the type of movement and the number of laser-camera units varies. Later on, Cyberware extended the capability of their system to perform whole body scan by utilizing four vertically-moving scanning units. The whole body scanner of Vitus^{smart} (Vitronic, Germany) consists of three scanner units that also moves vertically along three pillars. A foot scanner of Yeti (Vorum Research Corp.,

Canada) is composed of three units, which moves horizontally, two laterally and one from the bottom.

In fact, the first 3D whole body scanning system, named the Loughborough Anthropometric Shadow Scanner (LASS) [73], was developed by the University of Loughborough (UK) in 1989. It uses white light projection instead of laser projection. Four vertical lines are projected onto the scanned subject simultaneously and the images are captured by multiple cameras. The system is rotated horizontally to cover the whole body. Structured light based body scanner can be built into a static setup. However, the field of measurement of such scanning devices is limited, thus multiple sensor units are needed to provide whole body coverage. The NX-16 body scanner available from $[TC]^2$ (US) consists of 16 sensors and every four of them are stacked at each corner of the cubic scanning booth to cover partial of the body. NX-16 uses sinusoidal strip pattern for 3D surface capture and the depth estimation is based on the phase shift computed at each pixel. Similar principles have been applied in other systems, such as the body scanner Capturor (InSpeck, Canada), which can measure surfaces with maximal size of half part of the human body, e.g., upper torso. Customers can customize their body scanner with Capturor in terms of the number of sensors to be used. With high resolution digital camera being available nowadays, the resolution of a structured light based 3D scanner has been greatly improved. The Mephisto EX (4DDynamics, Belgium) utilizes an HDTV machine vision camera with a resolution of 1920×1080 pixels at 8 bits color depth as the main geometry camera. An optional Canon DSLR camera can be used along with the geometry camera to capture texture maps. A total of four scanner units are deployed at four corners for whole body coverage. The Mephisto EX body scanner reaches to a point accuracy of 0.15 mm (average). A major disadvantage associated with structured light scanner is that multiple units cannot be used simultaneously, since the light pattern from one sensor unit interferes with each other's.

Figure 2.5: A structured light based body scanner from 4DDynamics with four Mephisto EX scanner units. Each units consists of an HDTV machine vision camera as the main geometry camera, a digital projector, and a Canon DSLR texture camera.

Practically, this means that multiple units have to be used serially. This implies an extension of the acquisition time. Each sensor units in Mephisto EX scanner takes about one second to capture a surface. A total of four seconds is required for a whole body scan. In order to prevent measurement error cause by subject's movement during a scanner, a software based motion compensation is included in its 3D model construction.

Since Microsoft® released the Kinect gaming device for its XBox console in 2010, its potential being a low-cost depth sensor has been extensively explored in the field of body scanning. Kinect uses non-visible infrared light pattern and achieves depth estimation through light coding. The details of the light coding technique has not been disclosed from this developer PrimeSense (Israel), but researchers and developers speculate that depth is calculated by triangulation against a known pattern from its infrared projector [74]. The pattern may be unique to each of the individual Kinect, and is acquired at a known depth during the manufacturing process. $[TC]^2$ released a Kinect based body scanner KX-16 as the successor to its NX-16 scanner, and for the first time announced a whole body scanner under $10,000 price point across the industry. KX-16 uses 16 Kinect sensors and applies the

same configuration as the NX-16. Similar system is available from Size Stream (US), which uses 18 Kinect sensors. Size Stream body scanner configures all its sensors into two 3×3 matrices, one placed in the front of the subject, and one in the back. Other Kinect based body scanners include Styku (US) and Bodymetrics (UK), both of which utilize less sensor unites and focus on the apparel industry. Kinect devices are calibrated during manufacturing with a proprietary algorithm. The calibrated parameters are stored in the devices's internal memory and are used by the official drivers to perform the reconstruction. Although adequate for casual use such as during games, the manufacturer's calibration does not correct the depth distortion. Thus depth camera calibration [75,76] is usually required for a system to be used for measurement purpose.

Apart from the active lighting 3D scanning technologies, passive stereo vision 3D imaging technique is maturing over the past few years. The robustness and accuracy of the stereo vision system greatly benefits from the availability of high resolution digital cameras. Canfield Scientific, Inc (US) developed a family of VECTRA® 3D imaging systems for face and partial body surface acquisition. The VECTRA H1 uses a camera with a split-optical path stereo lens for facial imaging, while the VECTRA XT uses three pairs of stereo sets for frontal upper torso imaging. A stereo vision system requires a well-illuminated environment for the best quality of stereo images. The VECTRA systems are built with light panels so the reflection of light on skin surface is minimum. Since a stereo unit works best for the frontal-parallel surface and human body is full of curves, a whole body stereo vision system requires multiple stereo unit configured around the body in order to get the complete body surfaces. The Infinite-Realities (UK), a 3D scanning and character creation studio, reported the deployment of a single shot whole body scanning systems with 115 Canon DSLR cameras. All cameras are hardware synchronized and arranged around the scanned subject (Figure2.6).

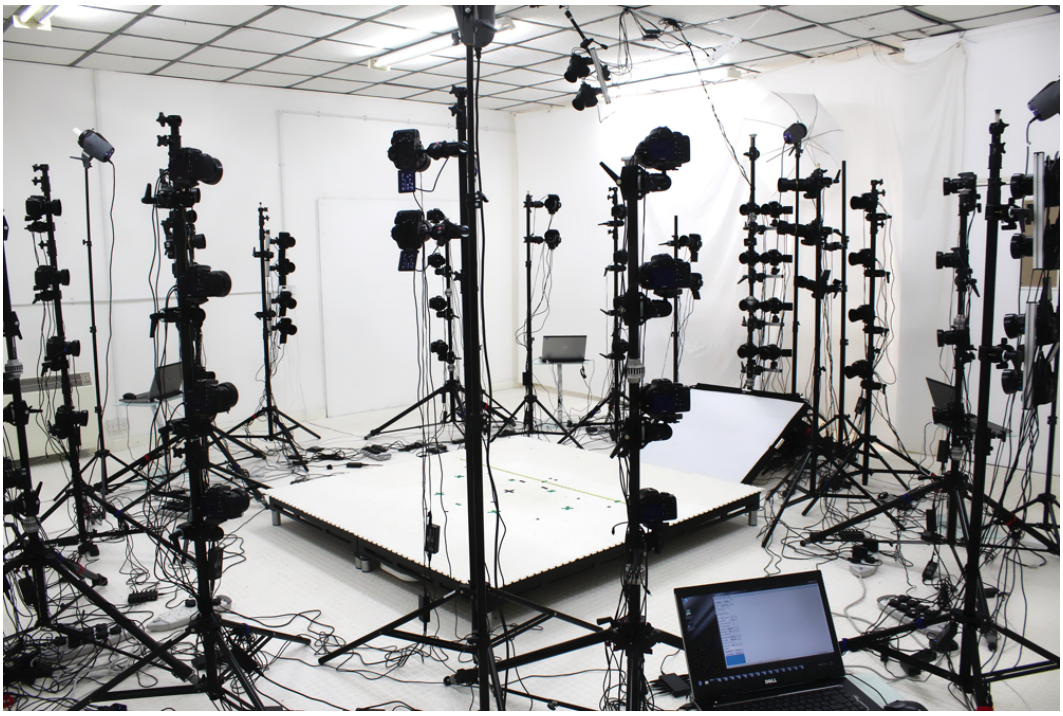


Figure 2.6: The stereo vision 3D imaging system deployed by Infinite-Realities (UK). System consists of 115 Canon DSLR cameras and studio lighting equipment.

2.5 Summary

The health threat and economical burden caused by the prevalence of obesity has triggered the need of a robust and piratical solution for obesity assessment and monitoring. 3D body imaging provides a convenient, noninvasive and radiation-free alternative for body dimension measurement. 3D body imaging techniques have been maturing over the years. We reviewed popular technologies and pointed out the advantages and disadvantages of each. Stereo vision has been one of the most active research topics in the computer vision community, and is becoming the technology of choice for depth sensing in a wide range of applications because of its fast image capturing, compact size and low cost. In the following chapter, we will focus on the principles of stereo vision and highlight the challenges in the application in body imaging.

Chapter 3

Stereo Vision Principles

This chapter provides a background knowledge about the depth estimation from stereo vision. The methods related to the problem of depth estimation from stereo images are discussed first, then a brief review of recent advances of stereo is given. Most of stereo algorithms follows a four-step framework. The key building blocks of this framework is discussed. This chapter ends with a description of the quality metrics we use in evaluating the performance of our developed stereo algorithm. System-wise evaluation is out of the scope of this topic and will be covered in later chapter. Much of the content in this chapter is at fundamental level and may be safely skipped for readers who have working experiences with stereo.

3.1 Depth Estimation from Images

Stereo vision recovers 3D shape from images taken under controlled lighting conditions. The depth estimation is based on the principle of multi-view triangulation, which is inspired from human binocular vision. By following this principle, a point's 3D position can be reconstructed by intersecting the lines of sight of the corresponding pixels in multiple images. Stereo vision assume the camera parameters are known and seeks to compute pixel correspondence for dense 3D reconstruction. Reviews of recent advances in this field can be found in [67, 77, 78]. The limitation of stereo vision comes from the difficulty in finding the correct pixel-to-pixel correspondence. Correspondence based stereo methods perform well when the imaged surface is Lambertian and contains rich texture. But

they may fail for surfaces that are non-Lambertain or Lambertian with little texture. Non-Lambertain surfaces are reflective and result in inconsistent color when they are viewed from different angles. Lambertain surfaces with little texture introduce ambiguities for the correspondence because pixels within a neighborhood are similar in color and is difficult to assign a one-to-one correspondence without regional information, such as the size and shape of the color block. Modern stereo methods resolve matching ambiguities by assuming the imaged surface is smooth or by applying planar prior model [79] for the imaged 3D shape. Nevertheless, obtaining accurate and robust depth estimation from stereo vision remains a very active and challenging field of research with the computer vision community. Since stereo matching is the main focus of this dissertation, Section 3.2 and 3.3 will provide more detailed background information.

3.2 Preliminaries

3.2.1 Image Formation

Without the loss of generosity, we use perspective projection to describe the image formation, through which a 3D scene is projected onto a 2D image plane and objects in the distance appear smaller than objects close by. This projection model can be presented by a pinhole camera (Figure 3.1a). Light from a feature point in the scene passes through a pinhole and forms an inverted image of the scene on the image plane. The pinhole camera model describes the mathematical relationship between the coordinates of a 3D point and its projection on the 2D image plane. The pinhole camera is widely adopted in the field of computer vision because it resembles closely the image formation process of a real camera. However, the image of a 3D scene is inverted in the pinhole camera model. To further simplify the image formation process and to prevent the image being inverted, the pinhole camera model can be redefined by placing the image plane in front of the focal point. An image is formed when light rays from a feature point in the 3D scene pass through

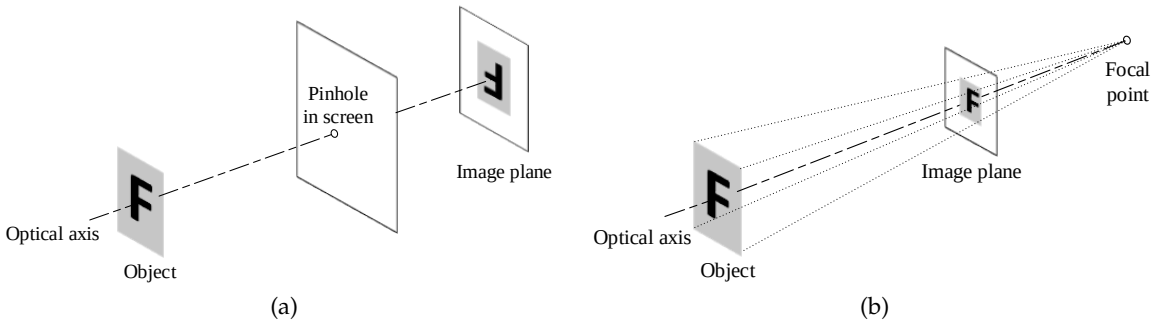


Figure 3.1: Perspective projection through (a) pinhole camera geometry: each ray of light passes through a common center of projection and intersects the image plane; (b) simplified camera model: each ray of light passes through the image plane and converges at the focal point.

the ”imaginary” image plane and converges in the focal point. Figure 3.1b illustrated the simplified camera model.

The primary difference between these projection models and real cameras is that real cameras have a lens instead of a point. Geometrical distortions introduced by the lens are not accounted for by the simple pinhole model. Fortunately, lens distortion can be corrected by a non-linear image transformation with camera parameters computed from camera calibration [80]. The pinhole camera also does not take into account the blurring of unfocused objects caused by lenses and finite sized apertures. This generally requires the 3D scene to be well focused for computer vision application, such as stereo matching.

3.2.2 Binocular Stereo Geometry

So far we have discussed how an image is formed through perspective projection. We now turn to the binocular stereo cameras and introduce important parameters for stereo correspondences and depth estimation. Stereo vision works similar in concept to human binocular vision, as shown in Figure 3.2. Since the two cameras observe the object from two different views, the captured left and right images are not the same due to perspective projection. The relative displacement of the same feature point in the two

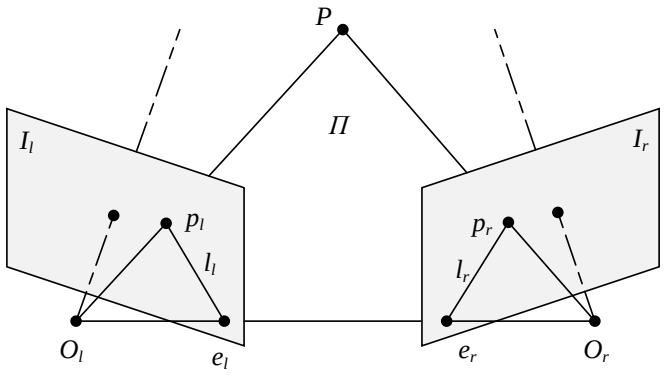


Figure 3.2: Epipolar geometry of binocular stereo vision. The 3D feature point P, the optical canters O_l and O_r, and the two image points p_l and p_r all lie in the same plane Π.

images is called the disparity, which is used to calculate the depth of the feature point with respect to the camera. As a convention adopted in this dissertation, we use subscripts "l" and "r" to denote the properties and measurements that are related to the left and right camera, respectively.

Assume P is an arbitrary feature point in a 3D scene, p_l and p_r are two images of P observed by two cameras with optical centers O_l and O_r, respectively. The feature point P and two optical centers define the *epipolar plane* Π. As is illustrated in Figure 3.2, the point p_r lies on the line l_r where Π and the right image plane intersect. The line l_r is the *epiploar line* associated with the point p_l, and it passes through the point e_r. Likewise, the point p_l lies on the epipolar line l_l associated with the point p_r, and the line l_l passes through the intersection e_l.

The points e_l and e_r are called the *epipoles* of the two cameras. The epipole e_r is the projection of the optical center O_l in the right image observed by the right camera and vice versa. Thus, if p_l and p_r are images of the same point, then p_r must lie on the epipolar line associated with p_l. This epipolar constraint plays a fundamental role in stereo matching because the search of correspondences can be restricted to one-dimensional instead of the whole image space, greatly reduces search range. The epipolar geometry can be easily

computed from camera parameters which can be acquired through camera calibration [80, 81].

Figure 3.3 shows a simple epiplar geometry that results from two cameras with identical focal length and coplanar image planes. In this scenario, corresponding epipolar lines are parallel to the horizontal axis of image planes and matched pixels p_l and p_r have the same y-coordinates. This special configuration greatly simplifies the correspondence problem since the explicit search of epipolar lines is no longer required. In addition, for area-based stereo matching approaches, two rectangular regions surrounding matched feature pixels can be evaluated directly without the need of image warping or interpolation. Most of stereo systems adopt this configuration to take advantage the simplified epipolar geometry. However, in practice it is technically difficult to install two identical cameras so that they sit at the same horizontal level and their image planes are coplanar. To achieve a geometrical equivalent to the simplified epipolar geometry with equal focal lengths, we can rectify the left and right images by re-projecting them to a specific coplanar plane which is equidistance to the baseline $\overline{O_l O_r}$. Rectification of stereo images can be achieved by applying image warping using two 3×3 homographies computed from the camera parameters [82–84].

Given two rectified images and the known simplified epipolar geometry, the position of a 3D feature point P can be determined by intersecting two rays $\overline{O_l p_l}$ and $\overline{O_r p_r}$. The correspondence between p_l and p_r is related by a disparity value d. The disparity is defined as the horizontal difference of two matched pixel coordinates as $d = x_r - x_l$. Note that $y_r \equiv y_l$ since corresponding pixels must be on the same vertical position for rectified images. Figure 3.3 illustrates how the depth of an arbitrary 3D point is computed from disparity under the simplified epipolar geometry. Denote the 3D point $P(X, Y, Z)$ and its

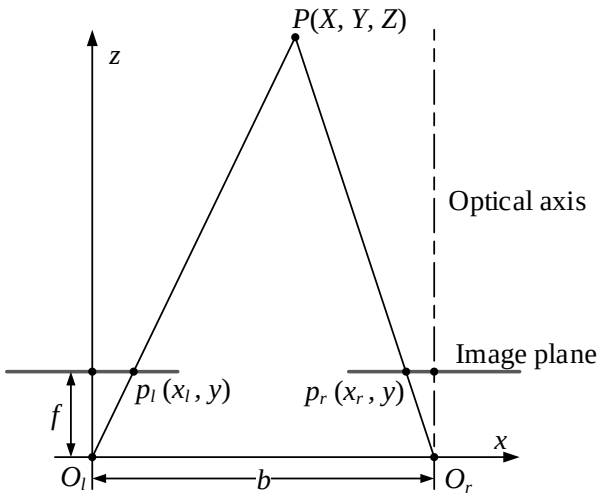


Figure 3.3: Stereo geometry in parallel-axis stereo vision. The disparity of a scene point P and its depth Z is related by $Z = -fb/d$.

2D images $p_l(x_l, y)$ and $p_r(x_r, y)$, we have

$$\frac{x_l}{f} = \frac{X}{Z} \text{ and } \frac{x_r}{f} = \frac{X+b}{Z} \tag{3.1}$$

from the relationship of similar triangles. The constants f and b denote the camera focal length and baseline, respectively. The disparity $d = x_r - x_l$ is proportional to focal length and baseline, and inversely proportional to the depth Z.

3.2.3 Stereo Correspondence

Stereo correspondence refers to the one-to-one match of the images of the same feature point between left and right views. Solving the stereo correspondence, or in other words, for each pixel in the reference image finding its corresponding matching points in the other image, is the primary task of binocular stereo vision. It is commonly adopted by most researchers that the image scene is composed of object with Lambertian surfaces and brightness consistency is assumed in order to establish the matching criteria for correspondences. The intensity consistency describes that corresponding points on a Lambertian

surface have the same intensity if they are viewed from different viewpoints. In practice, the Lambertian or intensity consistency assumption does not always hold for real-world scenes. Specularities, reflections, and transparency typically introduces problems to stereo matching algorithms. Even when the Lambertian assumption is true, stereo correspondence still remains a challenging task for the following reasons:

- **Sensor gain and bias.** The imaging sensors used to capture images from different viewpoints may have different gain and bias in their photonic response. This introduces color difference between the correspondence.
- **Repetitive patterns and textureless regions.** The intensity-consistency constraint is no longer valid for scenes that contain repetitive patterns or textureless retions.
- **Occlusions.** Occluded pixels, i.e., points visible from only one camera, does not have a match in the other view, thus should not be matched. Correctly identifying and handling occluded pixels is important for dense stereo vision.
- **Non-frontal-parallel surfaces.** Surfaces that are not parallel to camera's image plane may result in reduced resolution and blurring. The area of a non-frontal-parallel surface visible to left and right viewpoints are also different.
- **Depth discontinuities.** Preserving sharp depth discontinuities along object boundaries is especially important for some applications such as 3D reconstruction.
- **Noise.** Noise is unavoidable. There are always uncertain intensity values due to light variations, out-of-focus, and sensor noise introduced by the image formation process.

Traditionally, stereo matching algorithms are classified into two categories, feature-based and area-based. Feature-based approaches only establishes correspondences for distinct feature pixels that can be robustly distinguished and unambiguously matched

[85–87]. Other features, such as Scale-Invariant Feature Transform (SIFT) and Features from Accelerated Segment (FAST) corners can also be used for sparse stereo matching [88]. These features can typically be detected and matched at high speed, making this type of correspondence a viable solution for real-time robotic application. While feature points can be matched with high confidence, these methods are limited to spare or semi-dense depth estimation. Area-based approaches consider larger image regions that contain richer information than individual pixels to generate more stable matches. The matching function used in the area-based method typically computes the dissimilarity between support regions in stereo images. A major problem associated with area-based approaches is that they assume pixels within the support region have the same disparity. This is not valid for pixels near depth discontinuity or non-frontal-parallel surfaces. Therefore, in order to get accurate depth estimation the size and shape of matching windows should be carefully determined.

3.3 A Framework for Stereo Matching Algorithms

Following the taxonomy and evaluation of dense stereo matching algorithms reviewed by Scharstein and Szeliski [67], stereo matching algorithms generally follow four steps:

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

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

3.3.1 Matching Cost Computation

All stereo algorithms relies on a cost criteria which measure the similarity between pixels to establish pixelwise correspondences. A matching cost is a metric indicating how likely two pixels correspond to the same scene point. A low matching cost indicate a high confidence in the pixel-to-pixel correspondence. Matching cost computation is very often based on the absolute differences (AD), squared differences (SD), or Birchfield and Tomasi's (BT) sampling insensitive difference [89] of intensities and colors. Since these costs are sensitive to radiometric differences, costs based on image gradients are also used [90].

Besides the above methods, there are filter based cost function that are designed to tolerant global intensity variations caused by gain and exposure difference, image noise, different camera settings, etc. Images are preprocessed with certain types filters and then the filtered images are matched using common cost criteria, such as AD and SD. Popular filters include Laplacian of Gaussian (LoG) [91], rank filter [92], and mean filter. Normalized Cross Correlation (NCC) is another method for measuring matching cost. The normalization within a correlation support area effectively compensate variations in gain and bias. The main limitation 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 the work of Hirschmuller and Scharstein [93]

3.3.2 Cost Aggregation

Pixelwise cost calculation is generally ambiguous and wrong matches can easily have a lower cost than correct ones, due to noise, imaging sensor gain and bias, and so forth. Therefore, an additional constraint is usually applied to support smoothness by penalizing or rejecting changes of neighboring disparities. Local area-based methods ag-

gregate the matching cost by summing over a support region. A support region is typically a rectangular window centered at the current pixel. Conventional 2D aggretation methods smooth the cost volume by computing the weighted average of matching cost using box or Gaussian filters [94]. The advantage of using linear filters, such as box filter, for cost aggregation is that the 2D convolution process is separable and very fast implementation can be achieved. However, these methods tend to blur object boundaries with the fixed size of the support window. To avoid the blurring artifacts near depth discontinuities, shiftable windows [95, 96], windows with adaptive sizes [97, 98] or adaptive weights [99, 100] have been developed.

3.3.3 Disparity Computation and Optimization

Disparity computation and optimization refers to the methods of assigning a correct disparity value to a pixel. In general, these methods can be categorized into two major classes: local method and global method. In the local method, the disparity value at a pixel location is simply selected by a local Winner-Take-All strategy, that is, the disparity associated with the minimum aggregated cost at each pixel is chosen. In this scenario, the accuracy of selecting the correct disparity value largely depends on the quality and effectiveness of the cost computation and cost aggregation stage.

In contrast, global method make explicit assumptions about the scene that the imaged surfaces are piecewise smooth (except for object boundaries) and neighboring pixels should have very similar disparities. This assumption is generally true and the constraint used to enforce piecewise smooth is referred to as the *smoothness constraint* in the stereo vision literature. Global methods are usually formulated in an energy-minimization framework. Global methods are less sensitive to noise and textureless regions and are in general more robust than local methods since prior constraints provide regularization for regions difficult to match. However, global methods are usually more computationally intensive

than local methods.

3.3.4 Disparity Refinement

Disparity refinement is usually the last stage of a stereo matching algorithm, and it is done as a post-processing for checking the consistency, removing peaks and isolated values, interpolating gaps, or increasing the accuracy by subpixel interpolation.

Chapter 4

Framework Design of a Stereo Vision System

In this chapter, we describe the framework of the stereo vision system we have developed for 3D body imaging. As has been concluded from the previous chapter, stereo matching remains as a great challenge in the field of research. The quality of stereo images is crucial for a successful depth estimation. Stereo images must be taken in a controlled lighting condition that the scene is well illuminated but free from specular reflection, objects in the scene must be rich in surface texture, and the image should be corrected from lens distortion and properly rectified to enforce epipolar geometry. This chapter presents the setup of our stereo vision system and the calibration technique we apply in order to capture high quality pictures for dense stereo matching. An overview of our developed stereo algorithm will also be presented.

4.1 System Setup

The primary task of this research is to develop a robust solution for 3D body imaging, so that such a technology can benefit obesity study in terms of body shape monitoring and %BF estimation for a broad range of population. With this in mind, the engineering focuses in developing this type of system are cost, portability, and accuracy. To reduce the cost and shorten the duration of development, we have used off-the-shelf components. The basic unit of the system is a stereo unit that consists of a pair of digital Single-Lens Reflective (DSLR) cameras. Multiple stereo units are needed for whole body imaging. Our previous work on a rotary laser scanner [101] indicates that full body reconstruction can be

made from two scanning units that are placed in front and back of a subject, respectively. Later on, our developed active stereo vision system with artificial pattern projection [102] adopted this configuration but used two stereo units on one side to cover upper and lower body. A total of four stereo units were needed to provide whole body coverage. The similar construction of our active stereo system has been used in this study. However, this developed stereo system utilizes the natural skin texture that are readily available from a scanned subject as the stereo matching primitives, rather than the artificial projected random pattern used in our previous active stereo system.

Our hardware setup is illustrated in Figure 4.1. Two DSLR cameras are fixed on an aluminum plate through their tripod mounting holes to form a stereo pair. The optical axes of these two cameras are in parallel, this will reduce the amount of image distortion during rectification process. The baseline of the stereo unit is set to be around 150 mm. A large baseline can increase the disparities and eventually improve the depth resolution. However, a large baseline will cause a reduced common field-of-view, adversely reduce the coverage to the 3D scene. Two stereo units are mounted on an stainless steel pole vertically to provide coverage of one side of body. The stainless steel pole is attached to a metal base and is placed about 1.1 m away from the scanned subject in order to image a subject not taller than 1.9 m. The distance from the stereo units to the subject is largely determined by the fan angle of camera lens and the expected subject's height, thus it could be adjusted accordingly. A space of about 2.4×1.5 m will provide sufficient room for such a system to work.

The cameras we used in our systems are Canon EOS Rebel T3i (Canon Inc., Japan) DSLR cameras with 18-megapixel imaging sensor. The camera comes with a 18-55 mm lens. In order to cover a large field-of-view with limited distance, 18 mm focal length is applied on all cameras. To capture well focused image, each camera will take a few shots

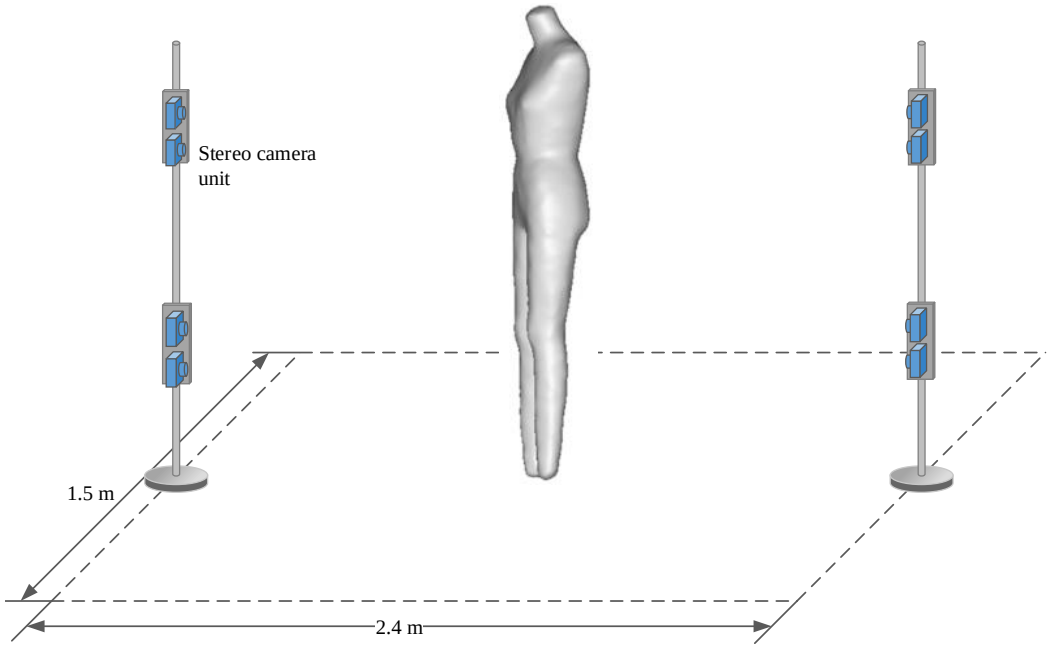


Figure 4.1: Schematic illustration of the system setup. The stereo vision system consists of four stereo units, and has eight cameras in total.

with ”Auto focusing” turned on to capture a random target placed at the imaging site where subject will stand. Once a well-focused image is acquired, we switch the focusing to ”Manual” mode and leave the focusing ring fixed at the best-focus position. This ensures the internal camera parameters stay unchanged during camera calibration and stereo picture taking. The advantages of DSLR camreras over less-expensive point-and-shoot cameras are larger imaging sensor and static lens constructions. A larger imaging sensor results in higher optical resolution, thus a higher signal-to-noise ratio (SNR) can be achieved. In addition, once good focus is achieved, the DSLR camera lens can be swithed to manual focusing and will stay in static, while a point-and-shoot camera’s lens will retract everytime it is switched off and deploy when it is switched on. The motion of the lens causes slight changes in internal camera parameters, thus may requires frequent calibration.

Our system is set up in a room with top ceiling lights and outdoor ambient lights through windows. Uneven illumination may occur in this casual setting, for example a subject's shoulder may appear brighter than his lower leg because the shoulder received more light from the ceiling lights. To reduced this uneven illumination, on-board camera flash is used so that the contribution from in-door light sources are reduced. Other researchers reported image capturing techniques by setting up cameras in a dark room and uses very long exposure time with a single flash for face imaging [103]. This allows the capture of highly synchronized images to minimize involuntary body motion, because camera shutters are all released and they are all waiting for the flash to fire. Another benefit of single flash image capturing is the ease to detect the region-of-interest (the scanned subject) based on the intensity, because foreground target is usually close to the flash and receives more photonic energy than background. A major disadvantage of this exposure strategy is the failure to meet the Lambertian surface constraint in certain areas, especially on face which is oily for some subjects. Pixels in the reflective area are usually white-out and causes mismatch in the disparity map. An effective solution to reduce reflection is to eliminate point-based light sources and to increase the ambient light or use surface-based light sources, such as light umbrellas or diffusers. However, this requires profession studio equipments and raises the effort in setting up such a system. The strategy we developed for picture taking takes the advantage of ambient lighting and balances the overall exposure with camera on-board flash. This combined method effectively reduces uneven illumination and significantly minimizes skin reflection.

The cameras are connected to a computer via USB cables, through which communication and data transfer are handled. Image capturing are trigged from our developed camera control software. All cameras are set to "Av" mode, in which the brightness of captured pictures is determined by the duration of exposure. To avoid the interference of

multiple flash firing during the same time, a short delay of 300 ms is applied between each capture made from cameras attached to the same pole. The total time needed to finish a whole body imaging is expected to be within two seconds.

4.2 System Calibration

System calibration involves two stages: camera calibration and 3D registration. The camera calibration calculates the intrinsic and extrinsic parameters of the cameras and determines the relative position and orientation between two cameras in a stereo setup. The 3D registration finds out the poses of each stereo units in a user defined world coordinate system, so that 3D surfaces captured from each individual stereo units can be merged into a common coordinate system.

4.2.1 Stereo Calibration

The camera calibration is a procedure of calculating the intrinsic and extrinsic camera parameters through feature point correspondences via nonlinear transformation from a user defined 3D pattern coordinate space to the 2D image coordinate spaces. The intrinsic parameters include the effective horizontal and vertical focal lengths f_x, f_y of the lens, the principle point (u_0, v_0) which describes the decentering of the lens, the radial lens distortion coefficients k_1, k_2, k_3 and the tangential lens distortion coefficients τ_1, τ_2. The distortion coefficients along with the principle point are useful in correcting geometrical distortion introduced by imperfect lens. The focal length is essential in estimating depth from disparity. The extrinsic parameters can be described by a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$, which define the camera pose with respect to the calibration pattern.

When working with perspective projection in computer vision or computer graphics, it is customary and convenient to use homogeneous coordinates. Mathematically, each

point in homogeneous coordinates is extended by an extra coordinate $s \neq 0$ that maps the point to a line through the origin in a space whose dimension is one unit higher than that of the original space. For example, a 2D image point (u, v) and a 3D scene point (X, Y, Z) can be represented by vectors $\begin{bmatrix} su & sv & s \end{bmatrix}^T$ and $\begin{bmatrix} sX & sY & sZ & s \end{bmatrix}^T$, respectively. Homogeneous coordinates allow us to express perspective projection of a 3D scene point onto a 2D image plane using the following linear equation:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{M} \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{4.1}$$

where $\begin{bmatrix} X & Y & Z \end{bmatrix}^T$ are the coordinates of a 3D point in the pattern coordinate space, $[u, v]^T$ are the coordinates of the projection on the image in pixels, and s is an arbitrary scale factor. $\begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}$ is a 4×3 matrix of extrinsic parameters, in which $\mathbf{R}$ is the rotation matrix defined on Euler angles (α, β, γ) and constructed with Rodrigues's rotation formula, $\mathbf{t}$ is the translation vector between the pattern coordinate system and the image coordinate system. $\mathbf{M}$ is called the camera intrinsic matrix, and is defined by

$$\mathbf{M} = \begin{bmatrix} f_x & k & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{4.2}$$

in which k is the skewness of the axes in the image plane and is usually 1 for most of cameras.

The image projection model defined by (4.1) is convenient because it is based on a linear transform and its parameters can be estimated as a closed-form solution. However, the non-linear optical distortion is not included in (4.1), thus (4.1) alone is insufficient for a complete camera calibration. In reality, a feature point detected on an image is geometrically distorted. Creating a classic model that includes radial distortion involves four steps:

1. Let $\begin{bmatrix} X & Y & Z \end{bmatrix}^T$ be the coordinates of a feature point that is defined on the calibration pattern, and $\begin{bmatrix} x & y & z \end{bmatrix}^T$ be the same point transformed into camera's coordinate space with rotation **R** and translation **t**,

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{4.3}$$

2. The perspective projection of this point in 2D undistorted, normalized image coordinates $\begin{bmatrix} x' & y' \end{bmatrix}^T$ is

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \frac{1}{z} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{4.4}$$

3. The transformation to link the undistorted, or corrected, coordinates $\begin{bmatrix} x' & y' \end{bmatrix}^T$ to he distorted coordinates $\begin{bmatrix} x'' & y'' \end{bmatrix}^T$ is a non-linear distortion function of parameters $\delta = \begin{bmatrix} k_1 & k_2 & k_3 & \tau_1 & \tau_2 \end{bmatrix}$.

$$\begin{bmatrix} x'' \\ y'' \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix} + \begin{bmatrix} \mathbf{D}_x^{(r)} + \mathbf{D}_x^{(t)} \\ \mathbf{D}_y^{(r)} + \mathbf{D}_y^{(t)} \end{bmatrix}, \tag{4.5}$$

where $\mathbf{D}_x^{(r)}$ and $\mathbf{D}_y^{(r)}$ describes the radial lens distortion,

$$\begin{bmatrix} \mathbf{D}_x^{(r)} \\ \mathbf{D}_y^{(r)} \end{bmatrix} = (k_1 r^2 + k_2 r^4 + k_3 r^6) \begin{bmatrix} x' \\ y' \end{bmatrix}, \tag{4.6}$$

$\mathbf{D}_x^{(t)}$ and $\mathbf{D}_y^{(t)}$ describes the tangential lens distortion,

$$\begin{bmatrix} \mathbf{D}_x^{(t)} \\ \mathbf{D}_y^{(t)} \end{bmatrix} = \begin{bmatrix} 2\tau_1 x'y' + \tau_2(r^2 + 2x'^2) \\ 2\tau_2 x'y' + \tau_1(r^2 + 2y'^2) \end{bmatrix}, \tag{4.7}$$

with $r^2 = x'^2 + y'^2$.

4. The 2D digital image coordinates (u, v) with lens distortion can then be calculated as

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{M} \begin{bmatrix} x'' \\ y'' \\ 1 \end{bmatrix}. \tag{4.8}$$

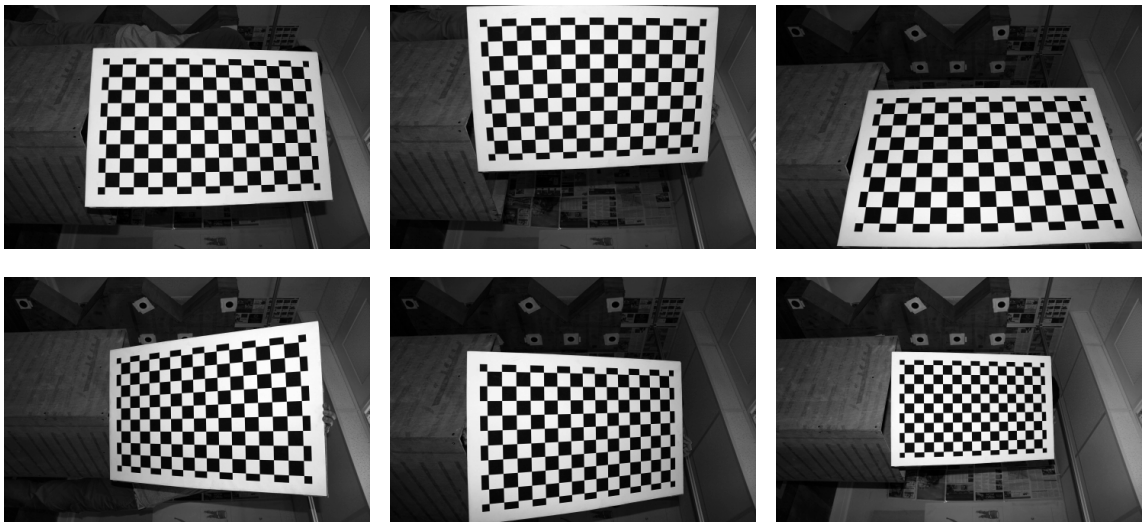


Figure 4.2: A set of images for camera calibration. Images are shown for individual camera calibration. Stereo calibration requires a set of image pairs for both left and right cameras.

Our camera calibration follows the technique originally proposed by Zhang [104], and uses a planar chessboard pattern to establish feature point correspondences. The pattern was printed by a high-quality poster printer and was attached to a rigid planar board. In includes 17×11 blocks, and the size of each block is 40×40 mm. The fabrication error was controlled under 0.2 mm. The 160 internal corners are used as feature points, and their locations on captured images are detected up to subpixel accuracy. The target should be placed at different positions and orientation. A set of images captured by one of the cameras are shown in Figure 4.2. The typical calibration errors are between 0.4–0.8 pixels. Once each camera is calibrated individually, a stereo calibration is applied to compute a rectification matrix for each stereo pair. The rectification matrix is used to reproject the distortion-corrected images onto the coplanar imaging planes to achieve simplified eqipolar geometry as discussed in Section 3.2.2. All the camera calibration procedures were implemented with OpenCV [105].

4.2.2 Global Registration

The 3D surfaced reconstructed from a stereo pair is defined in the reference camera's 3D coordinate system. Each stereo unit has its own camera coordinate system. The goal of the 3D registration is to compute the transformation between the reference camera's 3D coordinate system to a common world coordinate system, so that 3D surface data from each stereo unit can be merged. Since this transformation does not change the Euclidean distance between any points, it follows the rigid body model and involves rotation and translation only (no scaling). To determine a rigid body transformation, theoretically three non-collinear points are sufficient. Let $\{\mathbf{p}_i = \begin{bmatrix} x_i & y_i & z_i \end{bmatrix}^T | i = 1,2,3\}$ and $\{\mathbf{P}_i = \begin{bmatrix} X_i & Y_i & Z_i \end{bmatrix}^T | i = 1,2,3\}$ be the coordinates of three non-collinear points in the camera and world coordinate system, respectively. The registration task is to find a transformation that maps a point from the camera coordinate space to the world coordinate space

$$\mathbf{P}_i = \mathbf{R}^* \mathbf{p}_i + \mathbf{t}^*, \tag{4.9}$$

where $\mathbf{R}^*$ is a rotation matrix and $\mathbf{t}^*$ is a translation vector. Because there are measurement error in determining the point coordinates, the transformation can only be solved in a least-square fashion by minimizing the following error

$$\sum_{i=1}^{3} \|\mathbf{P}_i - (\mathbf{R}^* \mathbf{p}_i + \mathbf{t}^*)\|^2 \tag{4.10}$$

According to Haralick and Shapiro [106], the problem of finding the rotation and translation transformations by which one or more camera coordinate space can be made to correspond to a world coordinate space is defined as the *absolute orientation* problem. Horn [107] proposed a closed-form solution to this problem. The method has been so successful that there has been limited improvement in this area [108]. The steps to solve the absolute orientation problem is outlined as follows:

1. Calculate the point coordinates with respect to their centroids,

$$\mathbf{p}_i' = \mathbf{p}_i - \overline{\mathbf{p}}, \tag{4.11}$$

$$\mathbf{P}_i' = \mathbf{P}_i - \overline{\mathbf{P}}, \tag{4.12}$$

where $\overline{p}$ and $\overline{P}$ are the centroids of the points in the camera and world coordinate spaces, respectively. Now the new centroids of the points are $\mathbf{0}$ in both coordinate spaces.

2. The plane containing the points in the camera coordinate system is rotated to coincide with the place containing the points in the world coordinate space, so that

$$\mathbf{p}_i'' = \mathbf{R}_1 \mathbf{p}_i', \tag{4.13}$$

in which $\mathbf{R}_1$ is the rotation matrix that can be determined from the normals of the two plane.

3. An in-plane rotation $\mathbf{R}_2$ is sought that minimizes

$$\sum_{i=1}^{3} \left\| \mathbf{P}_i' - \mathbf{R}_2 \mathbf{p}_i'' \right\|^2. \tag{4.14}$$

4. The rotation and translation that relate the camera coordinate system to the world coordinate system is

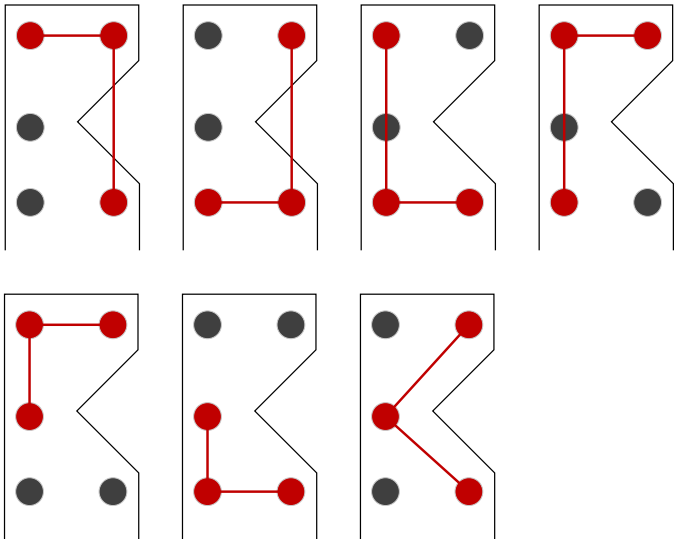
$$\mathbf{R}^* = \mathbf{R}_2 \mathbf{R}_1, \tag{4.15}$$

and

$$\mathbf{t}^* = \overline{\mathbf{P}} - \mathbf{R}^* \overline{\mathbf{p}}. \tag{4.16}$$

Horn's method for absolute orientation problem works with 3 feature points. An optimal solution can be achieved by running the Horn's methods for multiple times over different feature point combinations. We have designed a registration target with five

(a) The target for 3D registration.



(b) The combinations of three feature points selected out of five for each absolute orientation computation.

Figure 4.3: The 3D registration target and the feature points attached on the surface of the target.

feature points available to each of the stereo units, so that registration only requires one shot of the target. As is shown in Figure 4.3a, the top five circles are visible to the stereo unit that is to cover the upper body, and the bottom five circles are visible to the lower stereo unit. The two circles lie in the middle are shared between both stereo units. The center of each circle provides a feature point for the absolute orientation computation. We pick three circles out of five to run the Horn's method, and the orientations and translations computed from each point sets are averaged to generate an optimized global solution. The combinations of selecting three feature points out of five from the registration target are illustrated in Figure 4.3b;

Figure 4.4 shows the results of the 3D registration for one of our stereo unit. The stereo image pair captured by this unit is first rectified, then circle centers are detected from the images. Once the transformation between the reference camera's coordinate system

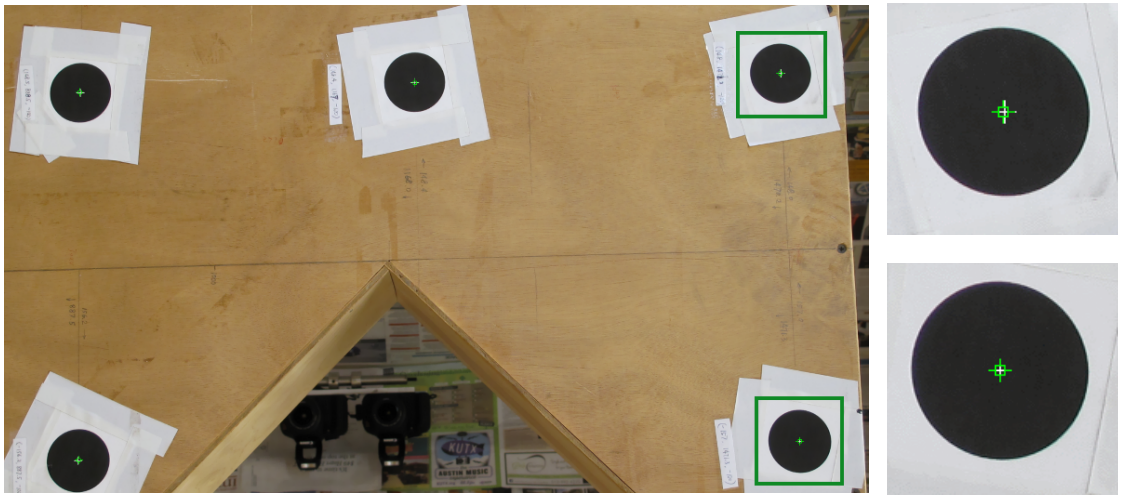


Figure 4.4: The results of 3D registration. The white crosses are the centers of circles detected from the image. The green crosses represent the back-projection of the circle centers (defined in the world coordinate system) transformed with the computed global rotation and translation. The agreement between white and green crosses indicates the accuracy of the global transformation. Right column: zoom-ins of the crosses highlighted on the picture.

and the world coordinate system is computed, the feature points (circle centers) measured on the registration target are back-projected onto the images. Ideally, the projected points should coincide with the circle centers that are detected from the image. The agreement between the detected points and the back-projected points indicates the accuracy of the absolute orientation.

4.3 Stereo Matching Algorithm Overview

4.3.1 Technical Challenges

The research proposed in this dissertation aims to make depth estimation from stereo images more accurate and robust for demanding applications that requires precise, reliable, and dense depth estimates. Towards this end, we address two key challenges for reconstructing dense scene structure using stereo matching and contribute several novel algorithms that are motivated by the specific requirements and limitations imposed by the

application of body imaging.

Our developed 3D body imaging system is a passive stereo vision system, which does not involve any artificial lighting for depth estimation purpose. The natural skin texture provides the matching primitives. Even with high resolution cameras, the quality of skin texture is not on par with artificial pattern found in an active system. Thus, the primary challenge this research addresses is to resolve the stereo matching ambiguities to achieve high-accuracy in depth estimation. A practical stereo algorithm has to deal with matching ambiguity results from inconsistent lighting during image capturing, sensor noise in image formation, homogeneous or repeated texture, and unmatched pixels due to occlusion. A robust stereo matching strategy must be able to accommodate all these characteristics in captured images.

The benefit of better texture in higher resolution stereo images comes with the increased computational cost, because more pixels are to be processed. With a total of four pairs of stereo units being used in our system and each camera captures pictures at 18-megapixel resolution, there are roughly 1.5×10^8 pixels to be processed to generate the 3D surfaces for a scanned subject. Stereo matching on high-resolution images is challenging because an algorithm may suffer from both long processing time and heavy memory consumption. A matching algorithm needs to search every possible disparity step for every pixel to determine the best match. The time cost for the algorithm is $\mathrm{O}(W \times H \times D)$, with W and H being the width and the height of the image, D being the range of disparity. The time cost increases by the power of 3 as the size of the image increases. For example, the stereo matching algorithm proposed by Mei, *et al.* [109], which is ranked number two on Middlebury website [110] in term of matching accuracy, requires 15 seconds to process both of the "Teddy" and the "Cones" images (450×375 pixels) in a non-parallel implementation. If the same algorithm is applied on an 18-megapixel image of size 5184×3456,

the projected processing time would be 4.5 hours. The scale of increase also applies to the memory consumption. In order to design a stereo matching algorithm that can handle high-resolution images, novel strategies are needed to improve both the time and memory efficiencies.

4.3.2 Multi-scale Matching

To develop a whole body stereo vision system with comparable capability to other 3D vision applications with reduced computational complexity, our proposed stereo matching algorithm takes the advantage of the multi-scale, coarse-to-fine strategy to address the stated challenges. A significant benefit of applying multi-scale matching is to utilize the matching result from a lower resolution scale as an initial guess for the subsequent scale. This prevents unnecessary search along the whole disparity space for a possible match, greatly reducing the time complexity. To reduce matching ambiguity, we design the algorithm such that both localized texture details and the texture gradient at the neighborhood of a matching feature point will be taken into account in the computation of the matching cost, enforcing a non-local optimization during matching.

In our multi-scale stereo matching framework, an image pyramid is first constructed by successive Gaussian filtering and down-sampling by a factor of two from original images. A total of four scales was applied, and the image resolution at the top of the pyramid is $\frac{1}{8}$ of the original size. The number of layers of the pyramid is flexible and can be configured as a parameter of the stereo matching. The criteria is that the major body surface features are still visible at the lowest resolution images.

Given a pyramid of stereo images, stereo matching starts from the top level of the pyramid. This is referred to as *coarse match*, since it matches large scale features and generates a low resolution disparity map at that scale. Our coarse match performs a full

disparity range search for every pixel in the image. This allows the algorithm to discover 3D surfaces at any depth within the predetermined depth-of-interest.

The disparity map computed from a lower resolution level provides input to the next higher resolution level, where it is used to constrain the disparity search range for match, and so on for the highest resolution level for the pyramid. A disparity map computed at a previous scale only contains values in integer format. To produce an estimate of the map at a higher resolution level, the map is first up-sampled by a factor of two with nearest-neighbor interpolation. This results in a new disparity map that matches the size of images at the new scale. Next, the value of each element in the map is scaled by a factor of two, so that the disparity value at the new pixel location is scaled properly.

In order to constrain the search range for the new match, we took a strategy that differentiate pixels that matched with high confidence, and pixels that were originally mismatched and were interpolated in the previous scale. Figure 4.5 illustrates the concept of generating the disparity search range from a disparity map of the previous resolution level. For pixels that passed left-right check in the previous scale, we are certain that these pixels were matched with high confidence, thus their new disparity value in the current scale should be close to their estimates with the only error being the error from the nearest-neighbor interpolation introduced in the up-sampling step. Thus a ± 1 relaxation is applied to these pixels. For those pixels that did not pass left-right check in the previous scale, their disparity values were interpolated from their neighbors whose texture information is similar. Even though constraints were applied in the interpolation as is described in Section 6.3, it is still possible that interpolated disparity deviates from the true value. This often occurs on curved or slanted surfaces with low texture, where disparity changes over the whole surface. For these pixels, the full disparity range that corresponds to the scene is assigned to give these pixel the opportunity to find the true disparity value with surface

texture at a higher resolution.

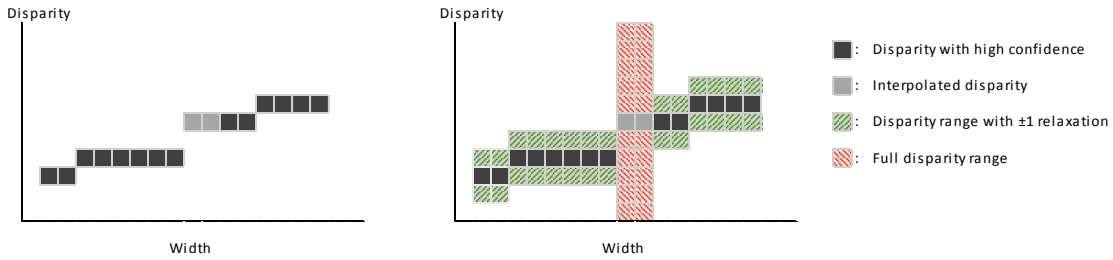


Figure 4.5: Generating disparity search ranges from the disparity estimates computed from a previous resolution scale. *Left*: the disparities of one row of elements within a disparity map. *Right*: the search ranges at each element location based on the confidence from a previous match.

The coarse match computes a 3D matching cost volume with the base of the volume matches to the size of image and the height corresponds to disparity ranges at this resolution level. Thus the complexity of the coarse match in big O notation is $\mathrm{O}(W_j H_j D_j)$, in which j denotes the resolution scale, W_j and H_j are the width and height of stereo images at scale j, D_j is the disparity range at scale j. Subsequent stereo matches on higher resolution scales only perform on fixed disparity range, i.e., [estimate - 1, estimate + 1] for pixels matched with high confidence. Since these pixels cover the most of the 3D surfaces, the complexity of the subsequent matches becomes $\mathrm{O}(W_j H_j)$. This effectively reduces the computation by an order of magnitude.

The cost volume in the subsequent matching becomes irregular in shape rather than a rectangular prism in the coarse match stage. This is the result of variable disparity range at each pixel location. This feature requires a flexible data structure to represent the cost volume, and the modification of our cost aggregation procedure to handle the discontinuity of disparity values between neighboring pixels. In additional to reduced computation, the benefit of these added algorithm complexity is the reduced memory footprint, which also achieves an order of magnitude of saving. As the stereo matching reaches down to the bottom of the pyramid, the memory consumption may becomes a major constraint

(tens of gigabytes of usage) if the cost volume is defined as fixed height. The work flow of our multi-scale stereo matching framework is presented in Figure 4.6.

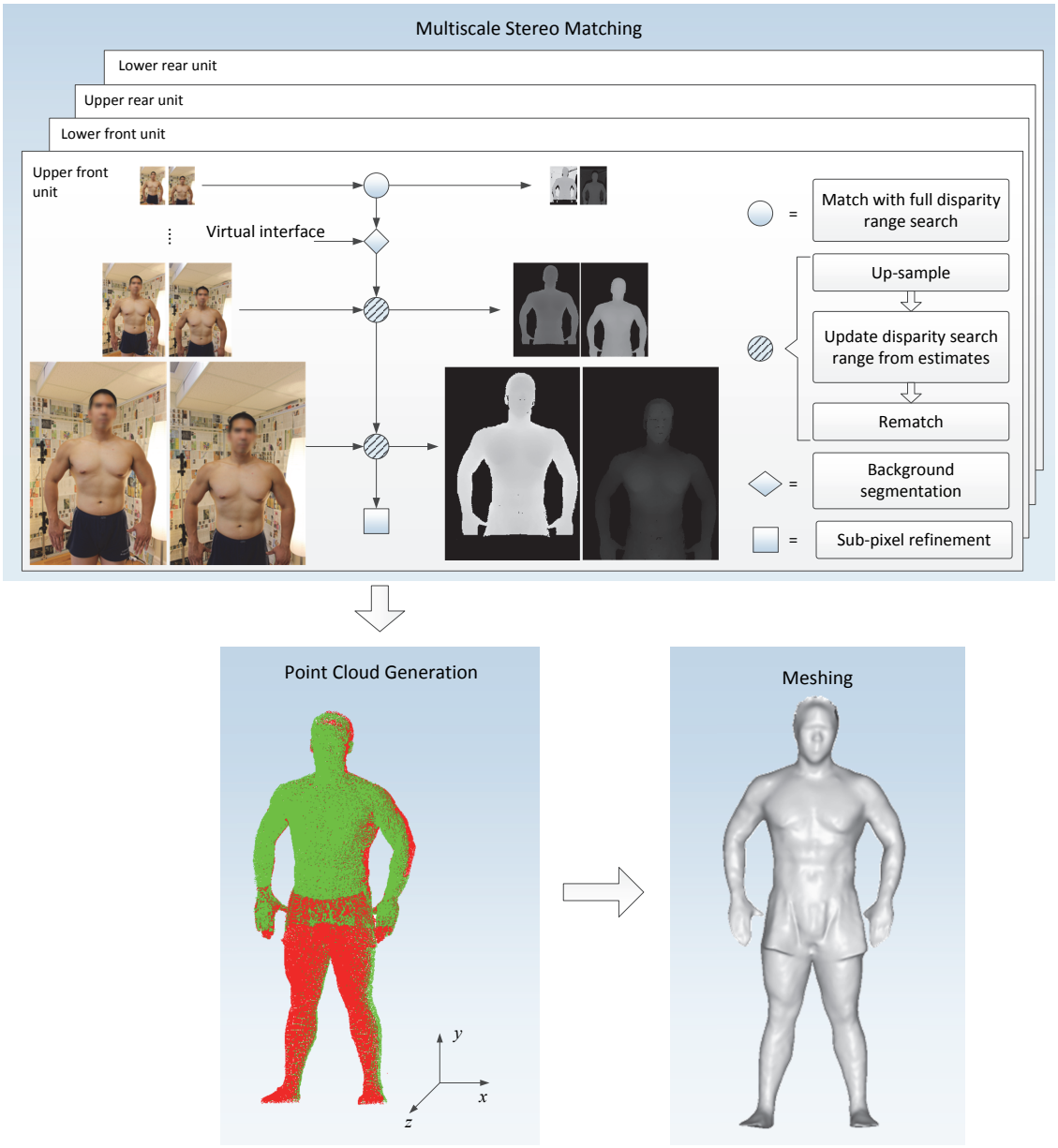


Figure 4.6: The work flow of our multi-scale stereo matching framework.

Multi-scale stereo matching is a crucial strategy in handling large images for dense disparity estimation. It reduces the amount of computation, thus saves total processing

time as well as memory consumption, making the matching problem solvable on a desktop computer in a reasonable time (a few minutes). Once the matching on the highest resolution scale is done, a sub-pixel enhancement process based on quadratic polynomial interpolation is performed to reduce the errors caused by discrete disparity steps. The final disparity results are obtained by enhancing surface fine geometric details through bilateral filtering.

4.3.3 Virtual Interface and 3D Background Segmentation

The virtual interface is the combination of surfaces in disparity space that correspond to surfaces in the 3D space which segment an imaged person from the rest of the space. The purpose of introducing a virtual interface is to provide a mechanism to automatically specify the disparity search range which is short enough to avoid unnecessary computation, but must be guaranteed to cover the depth of the ROI. To simply the computation of the virtual interface, we define four planes that are placed in the front, rear, top and bottom of the space where an imaged person will be standing in. The two side planes are not required because objects beyond them are invisible to the cameras. Figure 4.7 shows the arrangement of the virtual planes.

The origin of the world coordinate system, O_W, is at the center of the floor plane, and the positive Z_W-axis points to the frontal stereo units. To divide the 3D space into the foreground and background, three of the four planes are applied. For example, the bottom, top and rear planes are used for the frontal stereo units, and the bottom, top and frontal planes are used for rear stereo units. To convert the virtual planes in the 3D space to the virtual interface in the disparity space, the essential task is to compute the disparity map of the 3D planes. Detailed instructions were provided in [14] in which a background disparity map was computed with left camera being the reference of a stereo pair. In this proposed stereo matching framework, in order to compute a disparity map with the right

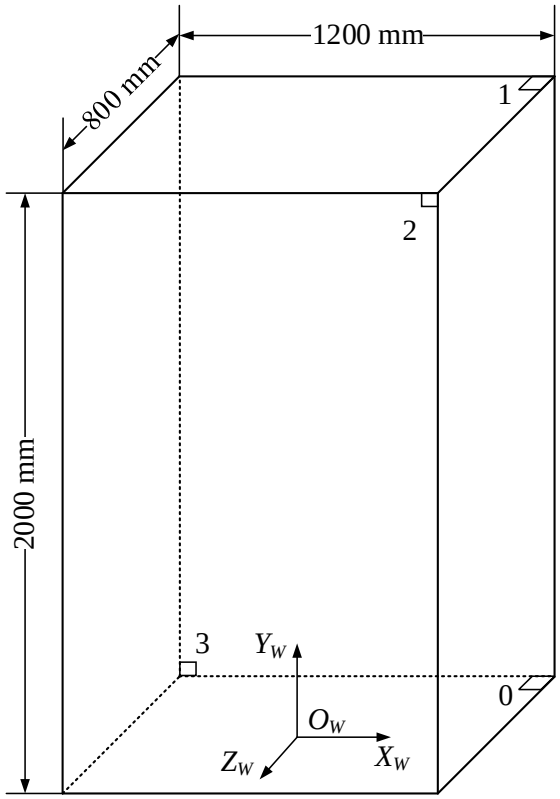


Figure 4.7: The virtual interface that defines the 3D region of interest. Four virtual planes are utilized: bottom, top, front and rear.

camera as the reference for left-right check purpose, the background disparity for the right camera should also be computed. The steps to generate the background disparity for the right camera is presented below.

Figure 4.8 shows a 3D plane, Π, being viewed by stereo cameras configured in parallel-axis setup. O_l and O_r are focal points of the left camera and right camera, with baseline distance of b. The normal of the plane Π is $\mathbf{n} = \begin{bmatrix} n_x & n_y & n_z \end{bmatrix}^T$. Without loss of generality, the plane is defined in the left camera's coordinate system with the normal $\mathbf{n}$ and the perpendicular distance from the origin s. Let $\mathbf{X}_l$ and $\mathbf{X}_r$ be the left and right camera

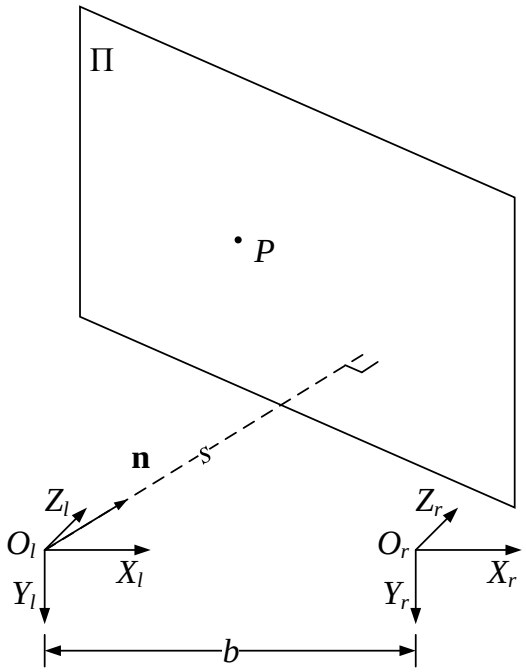


Figure 4.8: The homography that is induced by a 3D plane observed by a pair of stereo cameras.

coordinates of an arbitrary point P on Π. Thus, $\mathbf{X}_l$ and $\mathbf{X}_r$ satisfies

$$\mathbf{X}_r = \mathbf{H}\mathbf{X}_l, \tag{4.17}$$

with

$$\mathbf{H} = \mathbf{R} + \frac{1}{s}\mathbf{t}\mathbf{n}^T. \tag{4.18}$$

$\mathbf{R}$ and $\mathbf{t}$ are the relative rotation and translation of the right camera with respect to the left camera. $\mathbf{H}$ is the homography related with Π. Specifically, for the parallel-axis stereo geometry, $\mathbf{R} = \mathbf{I}$, $\mathbf{t} = [-b \quad 0 \quad 0]^T$, and thus we have

$$\mathbf{H} = \begin{bmatrix} 1 - \frac{b}{s}n_x & -\frac{b}{s}n_y & -\frac{b}{s}n_z \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{4.19}$$

and

$$\mathbf{H}^{-1} = \begin{bmatrix} \frac{1}{1-\frac{b}{s}n_x} & \frac{\frac{b}{s}n_y}{1-\frac{b}{s}n_x} & \frac{\frac{b}{s}n_z}{1-\frac{b}{s}n_x} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{4.20}$$

Denote $\tilde{\mathbf{x}}_l = [x_l \quad y_l \quad f]^T$ and $\tilde{\mathbf{x}}_r = [x_r \quad y_r \quad f]^T$, which are the homogeneous coordinates of the images of point P in the left and right image planes, respectively. Then according to the perspective projection, we have $\lambda_l \tilde{\mathbf{x}}_l = \mathbf{X}_l$ and $\lambda_r \tilde{\mathbf{x}}_r = \mathbf{X}_r$, where λ_l and λ_r are scalar values. In addition, $\lambda_l = \lambda_r$ stands for the parallel-axis stereo geometry. Then by replacing $\mathbf{X}_l$ and $\mathbf{X}_r$ in (4.17), we obtain

$$\tilde{\mathbf{x}}_l = \mathbf{H}^{-1}\tilde{\mathbf{x}}_r. \tag{4.21}$$

By combining (4.20) and (4.21) and rearrange, we can compute the disparity by

$$d = x_l - x_r = \frac{1}{1-\frac{b}{s}n_x}\begin{bmatrix} 1 & \frac{b}{s}n_y & \frac{b}{s}n_z \end{bmatrix}\begin{bmatrix} x_r \\ y_r \\ f \end{bmatrix} - x_r. \tag{4.22}$$

In practice, it is easier to define the plane Π in the global world coordinate system, so it is necessary to transform it into camera's coordinate system for background segmentation. We assume the plane equation in the world coordinate system is

$$\hat{\mathbf{n}}^T \mathbf{X}_W = \hat{s}, \tag{4.23}$$

in which $\hat{\mathbf{n}}$ is the plane normal defined in the world coordinate system, and $\hat{s}$ is its distance to the world coordinate system origin. The transformation between the camera and world coordinate system is

$$\mathbf{X}_W = \mathbf{R}^* \mathbf{X}_C + \mathbf{t}^*, \tag{4.24}$$

in which we assume the camera coordinate system is defined on the left camera, i.e., $\mathbf{X}_C = \mathbf{X}_l$, $\mathbf{R}^*$ and $\mathbf{t}^*$ are camera coordinate systems's rotation and translation with respect to the

Table 4.1: Planes of virtual interface. Plane parameters are defined in the world coordinate system.

	$\hat{\mathbf{n}}$	$\hat{s}$ (mm)
Plane 0 (floor)	$[0 \quad 1 \quad 0]^T$	2
Plane 1 (roof)	$[0 \quad 1 \quad 0]^T$	2000
Plane 2 (front)	$[0 \quad 0 \quad 1]^T$	400
Plane 3 (rear)	$[0 \quad 0 \quad -1]^T$	400

world coordinate system and are obtained through 3D registration. Then by inserting (4.24) to (4.23), we obtain

$$\left(\hat{\mathbf{n}}^T\mathbf{R}^*\right)\mathbf{X}_C = \hat{s} - \hat{\mathbf{n}}^T\mathbf{t}^*. \tag{4.25}$$

Comparing to $\mathbf{n}^T\mathbf{X}_C = s$, we obtain the plane parameters in the camera coordinate system,

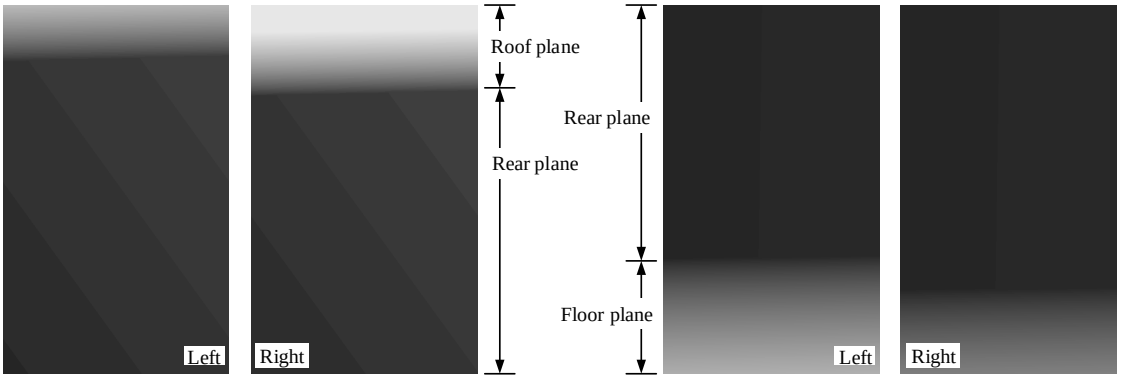
$$\mathbf{n} = \hat{\mathbf{n}}^T\mathbf{R}^*, \tag{4.26}$$

and

$$s = \hat{s} - \hat{\mathbf{n}}^T\mathbf{t}^*. \tag{4.27}$$

Table 4.1 defines the four planes that serve as virtual interface for foreground and background segmentation. The floor plane has been slightly lifted off the ground by 2 mm to separate the body form the ground. Examples of the computed background disparity maps are shown in Figure 4.9, in which Figure 4.9(a) are the maps from the upper stereo unit and Figure 4.9(b) are the maps from the lower stereo unit. The grayscale values of these maps have been scaled to highlight the variations within each map. Pixels of light color indicate they are close to the stereo unit, while pixels of dark color indicate they are far away. It can be observed from these disparity maps that the rear plane of our virtual interface is visible to both the upper and the lower stereo unit, while the roof plane

is only visible to the upper unit and the floor plane is only visible to the lower unit. The right-camera in the upper unit sees more roof plane than the left-camera, because the right-camera is mounted higher in elevation. The same applies to the lower unit that the left-camera in the lower unit sees more floor plane. An interesting feature that is revealed by these pairs of background disparity maps is that the variation of surface depth of the rear plan shows a diagonal pattern in the upper unit, while the pattern in the lower unit is horizontal. This is caused by the fact that our upper stereo unit has slight rotation angles around both the y- and z-axis with respect to the world coordinate system, but the lower unit has near zero rotation around the z-axis.



(a) Upper frontal stereo unit.

(b) Lower frontal stereo unit.

Figure 4.9: The background disparity maps computed for two frontal stereo units. Light pixel value indicates near range, and dark pixel value indicates far range. The roof plane and rear plane are visible to the upper unit, while the floor plane and rear plane are visible to the lower unit.

4.4 System-wise Innovation

Considering the prevalence of obesity, a convenient, reliable, safe and relatively inexpensive device is necessary for timely assessment and monitoring fitness in public health. The stereo vision system for obesity assessment proposed here represents four advances in the field of 3D body imaging:

1. Stationary setup that does not require any moving objects, and is capable of fast image capturing. The imaged subject is only required to remain stationary for one second;
2. Simple and low cost hardware that is easy to be reconfigured and deployed;
3. Improved calibration method specifically designed to be used with high-resolution cameras, and is able to be conducted in the test field;
4. Robust and efficient depth estimation algorithms with innovative functions for %BF estimation.

Compared to other popular 3D imaging solutions, stereo vision is the most flexible solution that requires less hardware but is capable of capturing high-resolution 3D images. It is static and the depth sensing is non-active thus no artificial lighting device is required. Our proposed stereo vision system is built upon consumer-grade, inexpensive cameras only. The overall cost in hardware is among the lowest in all types of 3D imaging devices.

4.5 Summary

The framework of the proposed body imaging system has been described in this chapter. We have set up a prototype with consumer-grade components. The construction of our system is quite simple, since it only involves cameras and its associated mounting accessories. The system can be easily disassembled, transported and reassembled. A two-stage system calibration method has been described. The system does not need to be calibrated frequently, as long as the camera parameters and camera positions remain unchanged. This property improves the portability of the system and reduces the cost of maintenance. The whole body image capturing only take about one second, greatly reduces the effect of motion. The 3D reconstruction is based on natural skin texture. The stereo matching is performed in a multi-scale framework. A full-range search of the op-

timized disparity values is only conducted at the highest scale with smallest image size. The matching results from a previous scale provides a good guess of the disparity for the next scale. This effectively reduces the amount of computation and saves processing time. It also improves the matching accuracy because large scale features, which usually cause less ambiguity, are matched first to generate a coarse map. The coarse map is then refined at a finer scale with surface textures are higher resolution are available. The details of our stereo matching algorithms are presented in the following chapters.

Chapter 5

Matching Cost Computation and Aggregation

Matching cost computation and aggregation are the first two steps in a four-step stereo matching framework. For a dense stereo matching, a matching cost is computed at each pixel for all disparities within the search range. It evaluates the similarity between the pixel-to-pixel correspondence. The cost aggregation connects the matching costs with a certain neighborhood to reduce mismatches by supporting smoothness. This chapter presents the cost computation and aggregation methods we developed for our 3D imaging system. We starts with a brief review of related work, and then describe our method that meets the requirements of our specific application.

5.1 Related Work

Stereo correspondence methods rely on matching costs for computing the disparities between matched pixels from left and right images. The simplest matching costs assume constant intensities or colors at matched pixel locations, but robust cost metric should compensate for certain radiometric differences and noise. Radiometric differences can be caused by different gain and bias settings between imaging sensors. This type of differences can be compensated by radiometric calibration. However, radiometric calibration requires special equipment and may not be possible in all situations. Further differences may be caused by non-Lambertian surfaces, for which the amount of reflected light depends on the viewing angle. While such differences can be reduced by making the stereo baseline smaller, this is limited by the physical dimensions of cameras. Small baseline also

reduces the geometric accuracy of the reconstruction. Thus, a practical stereo matching algorithm requires radiometric robustness.

Pixel-based matching costs include absolute differences (AD), squared differences (SD), and sampling-insensitive absolute difference [111]. Window-based matching costs include sum of absolute or squared differences (SAD/SSD) and normalized cross-correlation (NCC). NCC is generally more robust than SAD and SSD, because NCC accounts for gain differences in the matching windows due to normalization. Zero-mean versions of window-based costs, such as ZSAD, ZSSD and ZNCC, are developed to compensate for the bias in pixel intensities. Alternatively, bias can also be reduced by filtering the images before matching using a mean filter, computing a gradient magnitude image [112], or Laplacian of Gaussian (LoG) [91,113], which generates a smoothed second order derivative magnitude map. Unfortunately, all of these filters result in a blurred disparity image.

The weakness of a window-based methods is the inability to differentiate outliers that occur near object boundaries. For example, a window-based method will take into account background pixels when computing the matching cost for edge pixels of a foreground object. Nonparametric matching costs [92, 114, 115] were introduced for being robust against outliers near object boundaries. However, since nonparametric costs rely only on the relative ordering or pixel values, they are also invariant under all radiometric changes that preserve this order. In order words, a matched relative ordering of pixel values from two texture patches may be not necessarily sufficient to justify these two patches are from the same part of an object. But when combined with radiometric difference based measurement, nonparametric matching costs are more robust than when they are used alone. The Rank and Census methods [92] can be implemented as a filter followed by a comparison using the absolute difference or Hamming distance. Ordinal measurements [114] compute the distance of rank permutations of corresponding windows.

Another category of methods tries to explicitly model the complex radiometric relationships between images. Mutual Information (MI) has been introduced in computer vision by Viola and Wells [116]. Later work on MI in window-based stereo methods [117–119] demonstrated its power to model complex radiometric relationship. Others used approximations of MI [120] for a segment-wise stereo matching. It has been found [118, 119] that large windows are needed for collecting enough data to accurately estimate the joint probability distribution, but large windows lead to blurring at object boundaries. Towards this end, a hierarchical method [118] was proposed for estimating probability priors over the whole image at a lower resolution. These priors are fused with pixel values collected from smaller matching windows, which result in a reliable probability distribution. A pixel-based MI (without matching windows) in a global graph cuts stereo method has also been reported [121]. The probability distribution is iteratively calculated over the whole image using a prior disparity, which is random at the beginning. It has been shown [122] that a hierarchical calculation of pixelwise MI is as accurate as iterative calculation, but performance-wise MI is more computationally intensive.

According to the taxonomy [67], stereo matching algorithms are generally classified into two categories: local and global algorithms. In a local algorithm, the disparity computation at a given pixel location depends only on the intensities or colors with a local regions. All local algorithms require cost aggregation and usually make implicit smoothness assumption by aggregating supports. Global algorithms, on the other hand, make explicit smoothness assumptions and compute the best disparities by solving an optimization problem. Such algorithms typically skip the cost aggregation, but rather seeks a disparity solution that minimizes a global cost function. Popular global methods include Dynamic Programming (DP) [96, 123, 124], Belief Propagation (BP) [125, 126] and Graph Cut (GC) [79, 127]. Unlike local algorithms, global algorithms estimate the disparity at one

pixel using the disparity estimates at all other pixels.

Cost aggregation methods are traditionally performed locally by averaging matching costs over a support region. The fastest local cost aggregation method is unnormalized box filtering which runs in linear time (with respect to the number of image pixels) using integral image [128]. The major drawback is that it blurs across depth edges. Yoon and Kweon [99] demonstrated that edge-aware filters like bilateral filter [129] are very effective for preserving depth edges and Yang *et al.* [130] used bilateral filter for depth super-resolution. However, a full-kernel implementation of the bilateral filter is computationally expensive.

A number of approximation methods have been developed to accelerate bilateral filtering, including Paris and Durand's fast bilateral filter [131], Porikli's $O(1)$ bilateral filter [132] and Yang's real-time bilateral filters [133, 134]. These methods rely on quantization, and will degrade the performance as demonstrated in [135]. Paris and Durand's method was implemented on graphics processing unit (GPU) and was evaluated in stereo matching. However, the depth map accuracy is much lower than the full-kernel implementation [99]. Recently, He *et al.* [136] proposed a new edge-aware filter called guided image filter. Unlike bilateral filter, its runtime is linear with respect to the number of image pixels, and was demonstrated [137] to outperform all the other local methods on Middlebury benchmark [110] both in speed and accuracy.

The stereo images captured by our developed body imaging system may contain homogeneous texture regions and may show inconsistent lighting conditions due to the casual illumination setting. The matching cost computation and aggregation methods that are used in our study are designed to be robust to tolerant these image characteristics. This chapter provides the detailed description of our cost computation and aggregation methods.

5.2 Matching Cost Computation

5.2.1 Overview

The matching cost is calculated for a base (left) image pixel $\mathbf{p}$ from its potential correspondence pixel $\mathbf{q} = e_{bm}(\mathbf{p}, d)$ of the match (right) image. The function $e_{bm}(\mathbf{p}, d)$ represents the epipolar line in the right image for the left image pixel $\mathbf{p}$ with the line parameter d. For rectified images, we have $e_{bm}(\mathbf{p}, d) = \begin{bmatrix} p_x + d & p_y \end{bmatrix}^T$ with d as disparity.

An important consideration in selecting a cost function is the size and shape of the area that is considered for matching. The robustness of matching is increased with large area. However, the implicit assumption of constant disparity inside the area is violated at discontinuities, which leads to blurring object borders and fine structures. Although certain shapes and techniques can be used to reduce blurring, it cannot be avoided. Therefore, the assumption of constant disparities in the vicinity of $\mathbf{p}$ is not always reliable. To balance the performance of matching accuracy and the robustness in dealing with matching ambiguity, we propose a hybrid cost function that consists of there terms: cost of normalized cross-correlation $C_{\text{NCC}}(\mathbf{p}, d)$, cost of background suppressed color absolute difference $C_{\text{AD}}(\mathbf{p}, d)$, and cost of census $C_{\text{C}}(\mathbf{p}, d)$. The combined cost function is in the form of

$$C(\mathbf{p}, d) = \rho(C_{\text{NCC}}, \lambda_{\text{NCC}}) + \rho(C_{\text{AD}}, \lambda_{\text{AD}}) + \rho(C_{\text{C}}, \lambda_{\text{C}}), \tag{5.1}$$

where $\rho(C, \lambda)$ is a robust function on variable C:

$$\rho(C_{[\cdot]}, \lambda_{[\cdot]}) = 1 - \exp\left[-\frac{C_{[\cdot]}(\mathbf{p}, d)}{\lambda_{[\cdot]}}\right]. \tag{5.2}$$

The purpose of this function is twofold: first, it maps different cost measures to the range $[0, 1]$, such that (5.1) won't severely biased by one of the measures; second, it allows customizable control on the impact of the outliers with the parameter λ. This computation is done for every pixel at every possible disparity. $C(\mathbf{p}, d)$ is usually called the matching cost volume.

5.2.2 NCC with Adaptive Support

Traditional window-based NCC tends to blur the depth discontinuities because of outliers within the fixed window. To improve depth discontinuity in disparity maps, adaptive window sizes and shapes should be used. To reduce computational complexity for the NCC, traditional acceleration methods use two-dimensional integral image technique. However, this technique is inapplicable to NCC computation over non-rectangular support regions. NCC is computationally intensive without the effective acceleration. So here we present a fast NCC computation over shape- and size-adaptive support regions. First, pixelwise shape-adaptive support regions are constructed using a cross-based approach. Then, the NCC computation is transformed and effectively accelerated using an orthogonal integral image technique.

5.2.2.1 Cross-based Adaptive Support Region

To decide the pixelwise support regions $U(\mathbf{p})$ for pixel $\mathbf{p}$ in the left image and $U(\mathbf{q})$ for pixel $\mathbf{q}$ in the right image, we adopt an approach [138] that is built on upright crosses. As shown in Figure 5.1, a cross for a kernel pixel $\mathbf{p}$ composes of four arms with lengths of $\{h_{\mathbf{p}}^{-}, h_{\mathbf{p}}^{+}, v_{\mathbf{p}}^{-}, v_{\mathbf{p}}^{+}\}$. The support region at pixel $\mathbf{p}$ is constructed in two steps. The pixel at the end of left arm, $\mathbf{p}_l$, is determined by two following rules:

1. $D_r(\mathbf{p}_l, \mathbf{p}) < \tau$, where $D_r(\mathbf{p}_l, \mathbf{p})$ is the radiometric difference between $\mathbf{p}_l$ and $\mathbf{p}$, and τ is a pre-set threshold. The radiometric difference is defined as

$$D_r(\mathbf{p}_l, \mathbf{p}) = \max_{i \in \{R,G,B\}} |I_i(\mathbf{p}_l) - I_i(\mathbf{p})|. \tag{5.3}$$

2. $D_s(\mathbf{p}_l, \mathbf{p}) < L$, where $D_s(\mathbf{p}_l, \mathbf{p})$ is the spatial difference (or, distance) between $\mathbf{p}_l$ and $\mathbf{p}$, and L is a preset maximum length measured in pixels. The spatial distance is defined as $D_s(\mathbf{p}_l, \mathbf{p}) = |\mathbf{p}_l - \mathbf{p}|$.

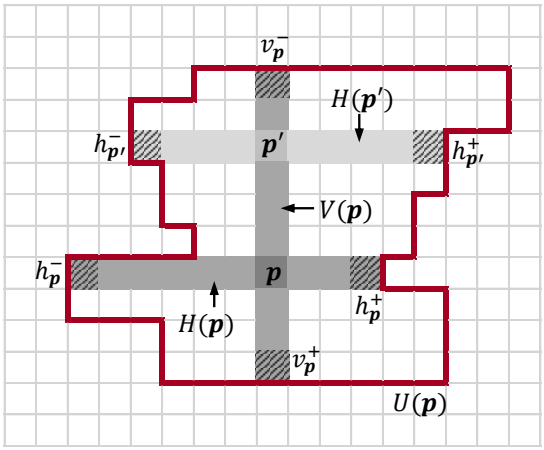


Figure 5.1: The adaptive support region $U(\mathbf{p})$ at pixel $\mathbf{p}$ is constructed by merging multiple horizontal segments $H(\mathbf{p}')$ along the vertical segment $V(\mathbf{p})$.

The two rules pose constraints on radiometric similarity and arm length with parameters τ and L. The right, up and bottom arms of $\mathbf{p}$ are built in the same way. The support region $U(\mathbf{p})$ is constructed by merging multiple horizontal segments $H(\mathbf{p}')$ along the vertical segment $V(\mathbf{p})$, where $\mathbf{p}'$ is a support pixel from $V(\mathbf{p})$. Due to the orthogonal construction of the cross, the complete map of support regions for each pixel in the image can be computed conveniently.

The accuracy of cross-based matching cost algorithm is closely related to the parameters τ and L, since they control the shape of the support regions. Large textureless regions may require large τ and L values to include enough color variation, but simply increasing these parameters for all the pixels would introduce more errors at depth discontinuities. We therefore enhance the cross construction with a dual-threshold scheme:

1. $D_r(\mathbf{p}_l, \mathbf{p}) < \tau_1$ and $D_r(\mathbf{p}_l^+, \mathbf{p}_l) < \tau_1$;
2. $D_s(\mathbf{p}_l, \mathbf{p}) < L_1$;
3. $D_r(\mathbf{p}_l, \mathbf{p}) < \tau_2$, if $L_2 < D_s(\mathbf{p}_l, \mathbf{p}) < L_1$.

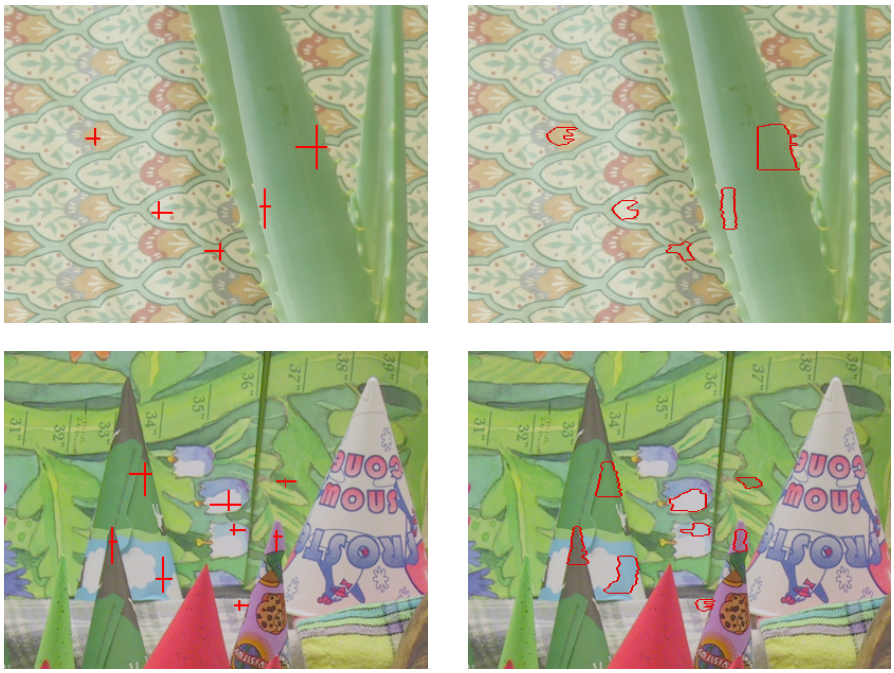


Figure 5.2: Construction of cross-based local support regions on the *Aloe* and *Cones* images. Left column: pixelwise adaptive crosses are constructed from local support skeletons for each kernel pixel. Right column: the shape-adaptive local support regions, which approximate local texture structures, are dynamically generated by integrating multiple horizontal arms of neighboring crosses.

Rule 1 restricts not only the radiometric difference between $\mathbf{p}_l$ and $\mathbf{p}$, but also the radiometric difference between $\mathbf{p}_l$ and its predecessor $\mathbf{p}_l^+$ on the same arm, such that the arm won't run across an edge in the image. Rule 2 and 3 allow more flexible control on the arm length. We use a large L_1 to include enough pixels for textureless regions. But when the arm length exceed a preset value L_2 ($L_2 < L_1$), a much stricter threshold value τ_2 ($\tau_2 < \tau_1$) is used for $D_r(\mathbf{p}_l, \mathbf{p})$ to make sure that the arm only extends in regions with very similar color. Examples of the adaptive support regions from cross bounds are shown in Figure 5.2. Parameters used to compute cross arms are $L_1 = 30$, $L_2 = 20$, $\tau_1 = 15$, and $\tau_2 = 8$. The local support regions approximate local texture structures with great consistency.

5.2.2.2 NCC Computation Acceleration

To measure the correlation between two signals $P = \{p_i | i = 1, \ldots, N\}$ and $Q = \{q_i | i = 1, \ldots, N\}$, NCC computes the following correlation coefficient,

$$C_{P,Q} = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}} \tag{5.4}$$

where $\bar{p}$ and $\bar{q}$ are the mean values of the elements from P and Q. The numerator represents the cross-correlation term and the denominator normalizes the coefficient to unit length.

Direct computation of NCC is computationally intensive and the time cost is proportional to the support region size. Assume the average size of the support region is S, the computational complexity to match two images with image size of M and disparity range D is $\mathrm{O}(M \times S \times D)$. Since M, S and D are usually large, the complexity is often prohibitive for fast stereo matching. Accelerating NCC over shape-adaptive matching region can be achieved by orthogonal integral image technique with a computational complexity of $\mathrm{O}(M \times D)$, greatly accelerating the computing process.

First, the numerator and the denominator of (5.4) can be transformed as follows,

$$\sum_i (p_i - \bar{p})(q_i - \bar{q}) = \sum_i p_i q_i - \frac{\sum_i p_i \sum_i q_i}{N} \tag{5.5}$$

$$\sum_i (p_i - \bar{p})^2 \sum_i (p_i - \bar{p})^2 = \left[\sum_i p_i^2 - \frac{\left(\sum_i p_i\right)^2}{N} \right] \times \left[\sum_i q_i^2 - \frac{\left(\sum_i q_i\right)^2}{N} \right] \tag{5.6}$$

Equation (5.5) and (5.6) suggest that the essential computational component is to sum the first and second order variables. In the case of stereo matching based on tow-dimensional signals, the computational component can be generally represented as

$$G_f(\mathbf{p}) = \sum_{(x,y) \in U(\mathbf{p})} f(x, y), \tag{5.7}$$

with $f(x, y) = I_l(x, y)$, $I_r(x, y)$, $I_l^2(x, y)$, $I_r^2(x, y)$, $I_l(x, y) I_r(x, y)$ for G_{I_l}, G_{I_r}, $G_{I_l^2}$, $G_{I_r^2}$, and $G_{I_l} G_{I_r}$, respectively.

An orthogonal integral image technique can be used to accelerate the computational component in the general form above. The accumulation over a two-dimensional shape-adaptive region is first decomposed into two consecutive orthogonal one-dimensional summing. Then each summing is further accelerated with integral image technique. The complete computation flow can be summarized in four steps. For the simplicity presentation, we use $\mathbf{p}$ and (x, y) interchangeably to denote a pixel ($\mathbf{p}$) at location (x, y).

Step 1 A horizontal integral image $F_f^H(x, y)$ is built on the image of $f(x, y)$, accumulating values at each row as

$$F_f^H(x, y) = \sum_{0 \le m \le x} f(m, y) = F_f^H(x-1, y) + f(x, y). \tag{5.8}$$

$F_f^H(x, y)$ can be iteratively computed with only one addition. When $x = 0$, $F_f^H(-1, y) = 0$.

Step 2 Based on $F_f^H(x, y)$, we can compute the horizontal integral $G_f^H(\mathbf{p})$ at a pixel location $\mathbf{p}$ as follows,

$$G_f^H(\mathbf{p}) = F_f^H(x_\mathbf{p} + h_\mathbf{p}^+, y_\mathbf{p}) - F_f^H(x_\mathbf{p} - h_\mathbf{p}^- - 1, y_\mathbf{p}). \tag{5.9}$$

Step 3 Taking the $G_f^H(x, y) = G_f^H(\mathbf{p})$ as the input, a vertical integral image F_f^V is built to store the cumulative column sum as

$$F_f^V(x, y) = \sum_{0 \le n \le y} G_f^H(x, n) = F_f^V(x, y-1) + G_f^H(x, y). \tag{5.10}$$

Step 4 The final result $G_f(\mathbf{p})$ for the pixel $\mathbf{p} = (x_\mathbf{p}, y_\mathbf{p})$ is computed with one final subtraction

$$G_f(\mathbf{p}) = F_f^V(x_\mathbf{p}, y_\mathbf{p} + v_\mathbf{p}^+) - F_f^V(x_\mathbf{p}, y_\mathbf{p} - v_\mathbf{p}^- - 1). \tag{5.11}$$

By taking I_l, I_r, I_l^2, I_r^2, $I_l I_r$ as the input image function of f, we get G_{I_l}, G_{I_r}, $G_{I_l^2}$, $G_{I_r^2}$, and $G_{I_l I_r}$, respectively. $C_{\text{NCC}}(\mathbf{p}, d)$ can be computed as

$$C_{\text{NCC}}(\mathbf{p}, d) = \frac{G_{I_l I_r}(\mathbf{p}) - \frac{G_{I_l}(\mathbf{p}) G_{I_r}(\mathbf{q})}{N}}{\sqrt{\left[G_{I_l^2}(\mathbf{p}) - \frac{(G_{I_l}(\mathbf{p}))^2}{N}\right] \times \left[G_{I_r^2}(\mathbf{q}) - \frac{(G_{I_r}(\mathbf{q}))^2}{N}\right]}} \tag{5.12}$$

where the pixel $\mathbf{q}$ in the right image is related to the pixel $\mathbf{p}$ in the left image with disparity d, with $\begin{bmatrix} x_\mathbf{p} + d & y_\mathbf{p} \end{bmatrix}^T = \begin{bmatrix} x_\mathbf{q} & y_\mathbf{q} \end{bmatrix}^T$. $N = \|U(\mathbf{p})\|$ is the size of the support region at pixel $\mathbf{p}$. Note that $I_l^2(x,y)$ and $I_r^2(x,y)$ can be pre-computed independent of the disparity value d, while $I_l(x,y)I_r(x,y)$ is computed at each iteration.

5.2.3 Cost of Census

Our second cost term is the census transform [92]. Census encodes local image structures with relative orderings of the pixel intensities other than the intensity values themselves, and therefore tolerates outliers due to radiometric changes and image noise. Given a pixel $\mathbf{p}$ in the image and a disparity value d, we use a 9×7 window centered at $\mathbf{p}$ to encode each pixel's local structure in a 64-bit string. If a neighbor pixel's intensity is higher than the kernel pixel $\mathbf{p}$, the corresponding bit in the 64-bit string is set to 1, and 0 otherwise. The cost of the match with census transform $C_C(\mathbf{p}, d)$ is defined as the Hamming distance of the two bit strings that stand for pixel $\mathbf{p}$ and its correspondence $\mathbf{q}$ that is related by disparity d. The Hamming distance counts the number of bits that differ in the two bit strings.

The census transform rely solely on the comparison between a neighbor pixel and the kernel, and is therefore invariant under changes in gain or bias. If a small count of pixels in a local neighborhood have a very different intensity distribution than the rest majority of pixels, only comparisons involving a small member of pixels are affected. Such pixels do not make a contributions proportional to their intensity, but proportional to their number.

In a recent review by Hirschmüller and Scharstein [93], census shows the best overall results in local and global stereo matching methods. However, the census transform could also introduce matching ambiguities in image regions with repetitive or similar lo-

cal structures. To handle this problem, more detailed information should be incorporated in the measure. For image regions with similar local structure, the color information might help alleviate the matching ambiguities. While for regions with similar color distributions, the census transform over a window is more robust than pixel-based measures, such as absolute difference. This observation suggests the incorporation of pixel based measure for overall robustness.

5.2.4 Background Suppressed Color AD

Our third cost term is the background suppressed absolute color difference. AD is a point based matching cost, thus it preserves depth discontinuity. However, it assume brightness constancy for corresponding pixels, which may not always be satisfied. To improve the robustness of AD, we consider background subtraction by *bilateral filtering* (*BilSub*). The bilateral filter sums neighboring values weighted according to proximity and color similarity.

5.2.4.1 Bilateral Filter

The bilateral filter is a filtering technique to smooth an image while preserving edges [129]. Its basic idea is very similar to Gaussian convolution: value of 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 blurring kernel. Given an image I and a kernel pixel $\mathbf{p} \in I$, the support weight $w(\mathbf{p}, \mathbf{k})$ of $\mathbf{p}$'s neighbor $\mathbf{k}$ is written as:

$$w(\mathbf{p}, \mathbf{k}) = \exp\left(-\frac{\|I(\mathbf{p}) - I(\mathbf{k})\|}{\sigma_r} - \frac{\|\mathbf{p} - \mathbf{k}\|}{\sigma_s}\right), \tag{5.13}$$

where $\|I(\mathbf{p}) - I(\mathbf{k})\|$ and $\|\mathbf{p} - \mathbf{k}\|$ represent the radiometric dissimilarity and the spatial distance between $\mathbf{p}$ and $\mathbf{k}$, respectively. The bilateral filter is controlled by two parameters

σ_r and σ_s. These two values control the influence from radiometric similarity and spatial proximity. An image filtered by a bilateral filter $BF(\cdot)$ is defined by

$$BF[I(\mathbf{p})] = \frac{\sum_{\mathbf{k}\in\Omega_\mathbf{p}} [w(\mathbf{p},\mathbf{k})I(\mathbf{p})]}{\sum_{\mathbf{k}\in\Omega_\mathbf{p}} w(\mathbf{p},\mathbf{k})}, \tag{5.14}$$

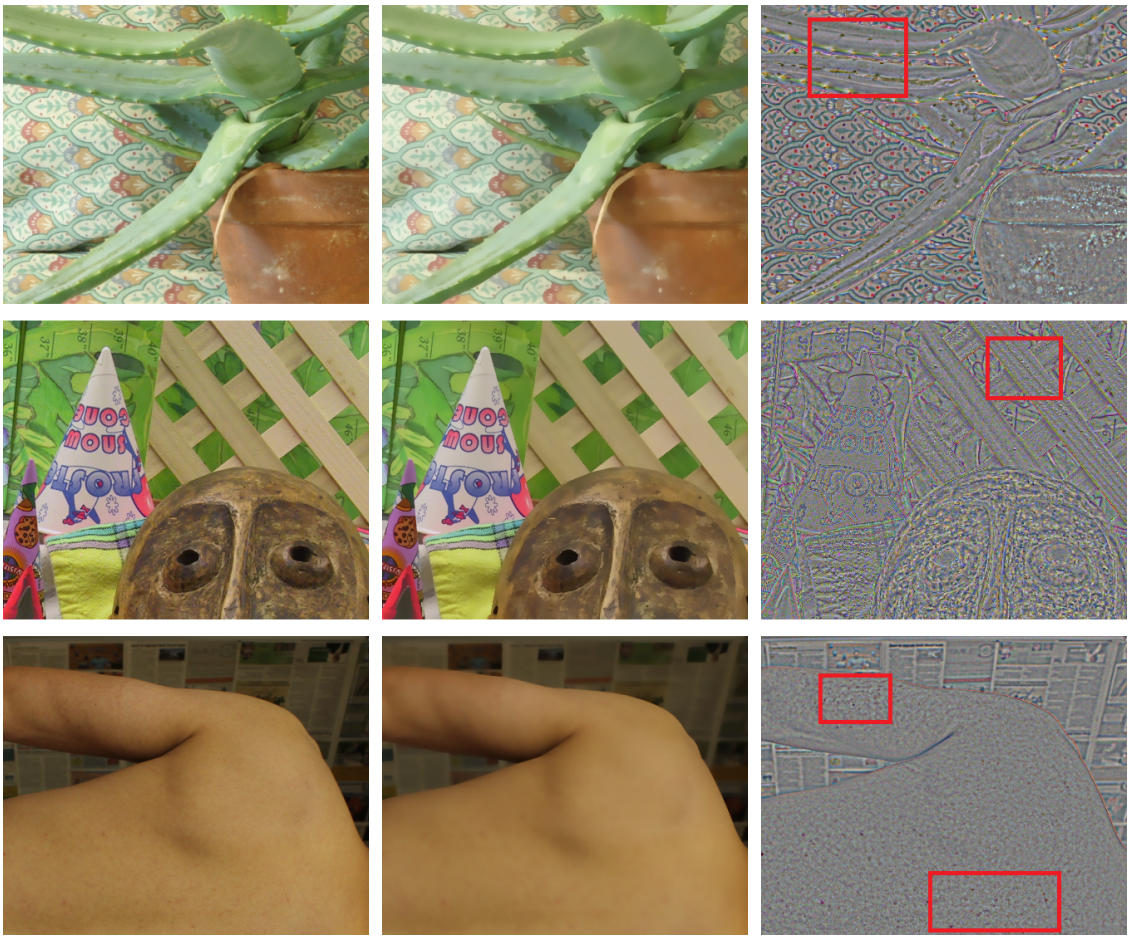
where $\Omega_\mathbf{p}$ denotes the set of all pixels in the support region and the normalization factor $\sum_{\mathbf{k}\in\Omega_\mathbf{p}} w(\mathbf{p},\mathbf{k})$ ensures support weights sum to one. More interesting properties, implementation details, and applications of bilateral filtering can be found in [139].

5.2.4.2 Background Subtracted AD

Background subtraction is implemented by subtracting from each value the corresponding value of the bilateral filtered image:

$$I_{\text{BilSub}}(\mathbf{p}) = I(\mathbf{p}) - BF[I(\mathbf{p})]. \tag{5.15}$$

This effectively removes a local offset without blurring high-contrast texture differences that may correspond to depth discontinuities. We use a kernel of 15×15 pixels in our bilateral filtering. The standard deviation of spatial distance is set to $\sigma_s = 3$. It defines the amount of smoothing. The standard deviation of radiometric distance is set to $\sigma_r = 20$. It prevents smoothing over high-contrast texture differences. On intensity images, the radiometric distance is computed as the absolute difference of intensities as defined in [140]. On color images, distance in CIELab space was originally suggested in [129]. Our approach however measure the chromatic difference in the RGB color space for simplicity and efficiency. Examples of background subtraction by bilateral filtering is shown in Figure 5.3. The edge preserving blurring effect can be observed in Figure 5.3b that only neighbor pixels whose color are similar to the kernel contribute to the filtering. In the background subtracted images of Figure 5.3c, the local bias and gain in each individual image are suppressed, and texture details on object surfaces are enhanced.



(a) Originals. (b) Bilateral filtered. (c) Background subtracted with bilateral filtering.

Figure 5.3: Examples of background subtraction with edge preserving bilateral filtering. Local bias and gain in each individual image are suppressed, and texture details on object surfaces are enhanced (highlighted in red) in the background subtracted images.

5.3 Cost Aggregation

Pixelwise cost calculation is generally ambiguous and wrong matches can easily have a lower cost than correct ones, due to noise, and so forth. Therefore, additional constraints should be added that supports smoothness by penalizing changes of neighboring disparities. Our cost aggregation strategy adopts the method originally proposed by Hirschmüller [122] that utilize multiple paths around a pixel for aggregation. We show

that by refining the parameters along the aggregation path, this method can produce aggregated results comparable to the adaptive weight method with much less computation time.

5.3.1 Definition of Pixelwise Energy for Aggregation

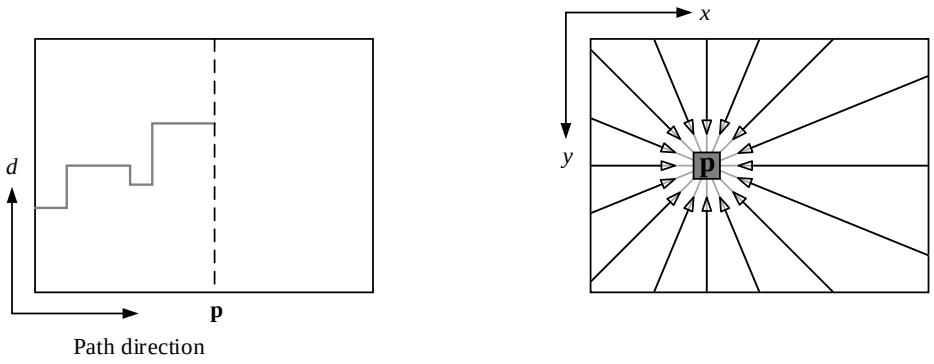
Within the multi-path cost aggregation framework, the pixelwise cost and the smoothness constraints are expressed by defining the energy $E(D)$ that depends on the disparity image D:

$$E(D) = \sum_{\mathbf{p}} \left[C(\mathbf{p}, d) + \sum_{\mathbf{k} \in \Omega_{\mathbf{p}}} P_1 T\left(|D(\mathbf{p}) - D(\mathbf{k})| = 1\right) + \sum_{\mathbf{k} \in \Omega_{\mathbf{p}}} P_2 T\left(|D(\mathbf{p}) - D(\mathbf{k})| > 1\right) \right] \tag{5.16}$$

The first term of (5.16) is the sum of all pixel matching costs for the disparities of D. The second term adds a constant penalty P_1 for all pixel $\mathbf{k}$ in the neighborhood $\Omega_{\mathbf{p}}$ of $\mathbf{p}$, for which the disparity changes only by one step. The third term adds a larger constant penalty P_2, for all larger disparity changes. The function $T(\cdot)$ takes in a boolean expression and return 1 if its value is True, and 0 otherwise. Using a lower penalty for small changes permits an adaptation of slanted or curved surfaces in the 3D scene. The constant penalty for all large changes, which are independent of their sizes, preserves discontinuities, since discontinuities are often visible as intensity changes.

5.3.2 Multipath Aggregation

The searching for a disparity image D that minimizes an energy function $E(D)$ is a 2D global minimization problem, and is NP-hard for many discontinuity preserving energies [141]. In contrast, the minimization along individual image rows in 1D can be performed efficiently in polynomial time using DP [89,142]. However, DP solutions generally



(a) Minimum cost path $L_{\mathbf{g}}(\mathbf{p}, d)$.

(b) 16 paths from all directions $\mathbf{g}$ for a pixel at $\mathbf{p}$.

Figure 5.4: Aggregation of costs in disparity space.

suffer from streaking effects, due to difficulty in relating the 1D optimization of individual image rows to their neighbor rows in a 2D image. The problem is that very strong constraints in one direction along image rows are combined with none or much weaker constraints in the other direction, that is, along image columns.

This leads to the idea of aggregating matching costs in 1D from *all* directions equally. The aggregated, or smoothed, cost $S(\mathbf{p}, d)$ for a pixel $\mathbf{p}$ and disparity d is calculated by summing the costs of all 1D minimum cost paths that end in pixel $\mathbf{p}$ at disparity d, as shown in Figure 5.4. These paths through matching cost volume are projected as straight lines onto the left image but as non-straight lines onto the corresponding right image, according to disparity changes along the paths. It is noteworthy that only the cost along the path is of interest, but not the path itself.

The cost $L'_{\mathbf{g}}(\mathbf{p}, d)$ along a path traversed in the direction $\mathbf{g}$ of the pixel $\mathbf{p}$ at disparity

d is defined recursively as

$$L'_{\mathbf{g}}(\mathbf{p},d) = C(\mathbf{p},d) + \min[L'_{\mathbf{g}}(\mathbf{p}-\mathbf{g},d), \\ L'_{\mathbf{g}}(\mathbf{p}-\mathbf{g},d-1)+P_1, \\ L'_{\mathbf{g}}(\mathbf{p}-\mathbf{g},d+1)+P_1, \\ \min_i L'_{\mathbf{g}}(\mathbf{p}-\mathbf{g},i)+P_2] \tag{5.17}$$

The pixelwise matching cost $C(\mathbf{p},d)$ is computed from our three-component cost computation as is presented in Section 5.2. The rest of the terms add the lowest cost of the previous pixel $\mathbf{p}-\mathbf{g}$ of the path, adjusted with appropriate penalty if depth discontinuity occur. This aggregation implements the behavior of (5.16) along a 1D path. Adding costs along an arbitrary path would not allow us to enforce *ordering* and *visibility* constraint, because they cannot be applied for the paths that are identical to epipolar lines. We will leave these constraint to subsequent processes. The values of $L'_{\mathbf{g}}$ increase constantly along the path, which may lead to very large values. However, (5.17) can be modified by subtracting the minimum path cost of the previous pixel from the whole term

$$L_{\mathbf{g}}(\mathbf{p},d) = C(\mathbf{p},d) + \min[L_{\mathbf{g}}(\mathbf{p}-\mathbf{g},d), \\ L_{\mathbf{g}}(\mathbf{p}-\mathbf{g},d-1)+P_1, \\ L_{\mathbf{g}}(\mathbf{p}-\mathbf{g},d+1)+P_1, \\ \min_i L_{\mathbf{g}}(\mathbf{p}-\mathbf{g},i)+P_2] - \min_j L_{\mathbf{g}}(\mathbf{p}-\mathbf{g},j). \tag{5.18}$$

This adjusted cost aggregation does not change the actual path through disparity space, since the subtracted value is constant for all disparities at a given pixel location at $\mathbf{p}$. Thus the disparity step that has the lowest cost at pixel $\mathbf{p}$ does not change. The costs $L_{\mathbf{g}}$ are summed over paths in all directions $\mathbf{g}$:

$$S(\mathbf{p},d) = \sum_g L_{\mathbf{g}}(\mathbf{p},d). \tag{5.19}$$

We selected a total of 16 paths covering 360° of a pixel for a good coverage of the 2D image. Paths that are not horizontal, vertical, or diagonal are implemented by going one step horizontal or vertical followed by one step diagonally. This will not generate paths that are evenly distributed around 360°, but it avoids the interpolation of costs between adjacent pixels.

5.3.3 Aggregation with Adaptive Penalties

During the aggregation along a specific path, P_1 and P_2 are two parameters for penalizing the disparity changes between neighboring pixels. While P_1 penalize small disparity change ($|\Delta d| = 1$), P_2 penalize large disparity change ($|\Delta d| > 1$). As suggested in [122], instead of using a constant value, P_2 can be made adaptive to the intensity gradient, that is,

$$P_2 = \frac{P_2^*}{|I(\mathbf{p}) - I(\mathbf{k})|}, \tag{5.20}$$

for neighboring pixels $\mathbf{p}$ and $\mathbf{k}$ in the reference (left) image, where P_2^* is a chosen constant. equation (5.20) is an inverse function of absolute radiometric difference between to neighboring pixels. It generates a large penalty value when the absolute difference is small due to its non-linearity, and it is a continuous function with respect to the absolute difference. However, this function only depends on the radiometric differences in the reference image, ignoring the differences in the match image. It may reject disparity changes from an incorrect value at the previous pixel to the correct one at the current pixel, when the pixel color barely changes in the reference image. This behavior can be corrected by checking the radiometric differences in both the reference image and the match image. Instead of taking the inverse of difference, we apply a step function based on radiometric differences $D_1 = D_r(\mathbf{p}, \mathbf{p} - \mathbf{g})$ in the reference image and $D_2 = D_r(\mathbf{p} + d, \mathbf{p} - \mathbf{g} + d)$. $D_r(\cdot, \cdot)$ is the same function as is defined in (5.3), then P_1 and P_2 can be adaptively adjusted:

Table 5.1: Parameters setting for our cost computation and aggregation methods.

Parameters	Values	Descriptions
λ_{NCC}	1.0	The control parameters for robust cost function $\rho(C_{[\cdot]}, \lambda_{[\cdot]})$
λ_{AD}	30	
λ_{Census}	1.0	
L_1	22	Arm lengths for calculating adaptive support regions
L_2	10	
τ_1	20	Thresholds of color difference for adaptive support
τ_2	6	
σ_r	20	Variance for radiometric difference in bilateral filtering
σ_s	3	Variance for spatial distance in bilateral filtering
P_1^*	1.0	Penalties to the costs at disparity discontinuity in cost aggregation
P_2^*	3.0	
τ_{Agg}	15	Threshold of radiometric difference to determine disparity discontinuity

1. $P_1 = P_1^*$, $P_2 = P_2^*$, if $D_1 < \tau_{\text{Agg}}$, $D_2 < \tau_{\text{Agg}}$;
2. $P_1 = P_1^*/4$, $P_2 = P_2^*/4$, if $D_1 < \tau_{\text{Agg}}$, $D_2 \geq \tau_{\text{Agg}}$;
3. $P_1 = P_1^*/4$, $P_2 = P_2^*/4$, if $D_1 \geq \tau_{\text{Agg}}$, $D_2 < \tau_{\text{Agg}}$;
4. $P_1 = P_1^*/10$, $P_2 = P_2^*/10$, if $D_1 \geq \tau_{\text{Agg}}$, $D_2 \geq \tau_{\text{Agg}}$.

In the above rules, P_1^*, P_1^* are constants, and τ_{Agg} is a threshold value for radiometric difference. This ensures that a fairly large penalty will be applied to disparity change when radiometric differences between two neighboring pixels in both reference image and match image are small, while a relatively small penalty will be applied when radiometric differences between neighboring pixels are large. For any cases in between these two conditions, a median penalty will be applied. But still, it has always to be ensured that $P_2^* \geq P_1^*$.

The results of applying adaptive penalties at depth discontinuities for cost aggregation are shown in Figure 5.5. Parameters for these methods are given in Table 5.1, which are kept constant for all test image pairs. Incorrectly matched pixels have been removed from all disparity maps, therefore the black areas in Figure 5.5 represent either occluded

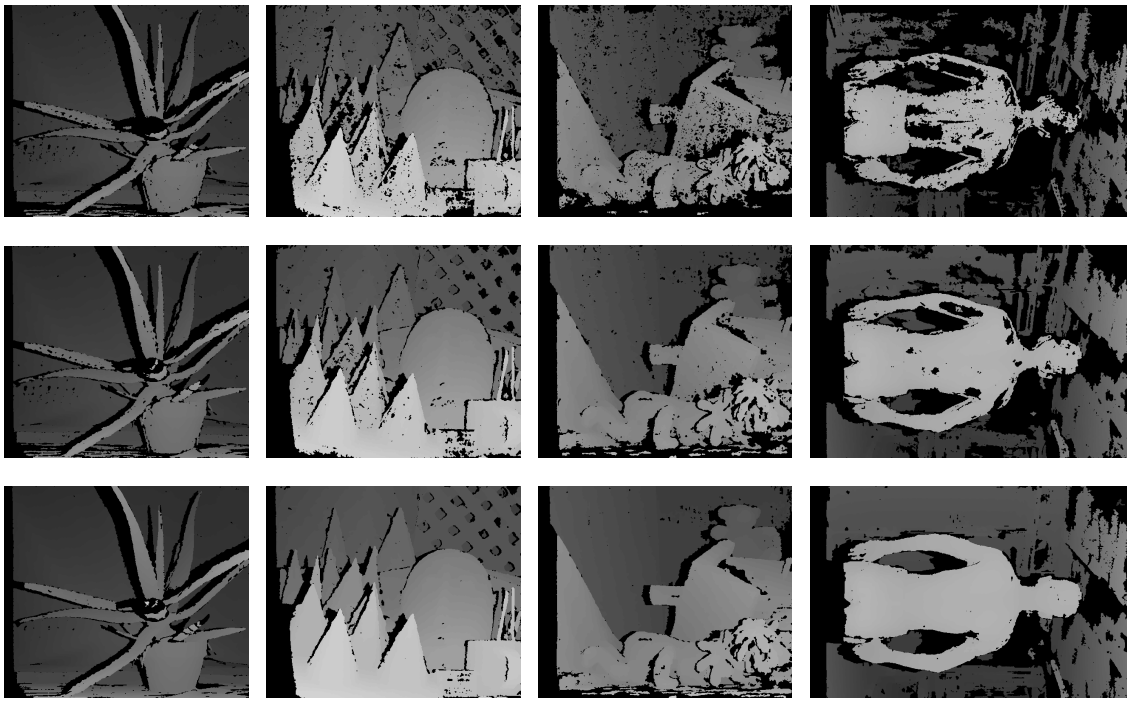


Figure 5.5: Cost aggregation with adaptive penalties at depth discontinuity. Top row: depth maps computed without cost aggregation; middle row: depth maps computed with static penalties in cost aggregation; bottom row: depth maps computed with adaptive penalties.

areas or mismatched areas. The middle row in Figure 5.5 shows the disparity map generated by cost aggregation with constant penalties, and the bottom row shows the disparity map generated with adaptive penalties. The images on the bottom row have less holes, and the edges of foreground objects are more accurate than the images on the middle row. It can also be observed that our adaptive cost aggregation algorithm has limited ability in picking up clear and sharp edges for the foreground object. This is because our aggregation is done by following 16 aggregation directions evenly distributed around a pixel. When an edge is encountered during the aggregation, about half of the directions are from background to the foreground, in which a large penalty is applied to the disparity change from background to foreground. As a result, the aggregation on an edge pixel is only about

half as effective as on a pixel within a foreground surface. But even with this reduced effectiveness in aggregation on edge pixels, the disparity map computed from aggregated cost volume is still more accurate than a disparity map computed without aggregation. Overall, our result indicates that mismatches have been greatly reduced, and our algorithms is capable of generating complete surfaces based on robust matching cost computation and aggregation.

5.4 Acceleration on Multi-core Processors

To achieve high performance in computing, we take advantage of the parallel computing power in modern multi-core processor, and implement the matching cost computation and cost aggregation in multiple threads to enhance computational speed. The computation parallelism is implemented in OpenMP [143]. OpenMP parallelizes a computation task by branching the master thread, which is a series of instructions executed consecutively, into a number of slave threads, through which a task is divided among them (Figure 5.6). The threads then run concurrently, with the runtime environment allocating threads to different cores and processors.

5.4.1 Parallelized Matching Cost Computation and Aggregation

In theory, the amount of speed gain that can be achieved by parallelizing a task depends solely on the fraction of serial code which could have been executed simultaneously. A helpful guide to discover the underlying parallelism while ensuring data integrity is to look for the same repeated computation performed on different data, because the operations that execute on the current data does not affect the results from the previous and subsequent data. In the matching cost computation stage, most of the operations can be parallelized because the same operation is performed on each image pixel at each disparity step. Our parallelized operations include adaptive support region computation, bilateral

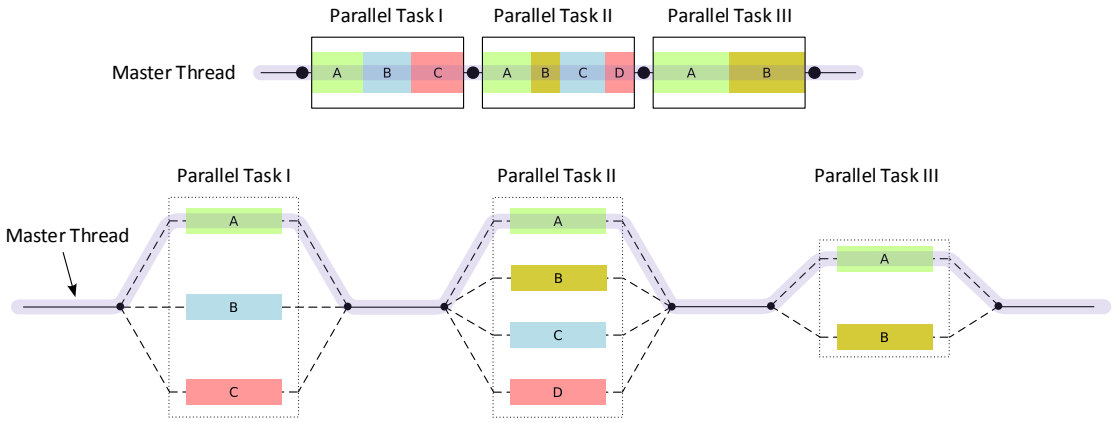


Figure 5.6: An illustration of OpenMP multi-threading where the master thread forks off a number of threads which execute blocks of code in parallel.

filtering, census transform, fast NCC computation, Hamming distance computation, and absolute color distance computation. The first three operations are done in *two* nested `for` loops which iterate through all pixels in an image, while the last three operations are done in *three* nested `for` loops with one more dimension in the disparity space. Figure 5.7 shows an example of the parallel execution of the adaptive support region computation on a quad-core processor with four threads running at the same time. The inner loop that is highlighted in red in Figure 5.7 is treated as a work unit and it handles an individual row of the image. All work units are evenly assigned to all work threads.

The result of matching cost computation is a three-dimensional volume $C(\mathbf{p}, d)$ of size $W \times H \times D$, with W and H being the width and height of the image, and D being the range of disparities. The cost aggregation is performed on this cost volume, following 16 path directions defined at each slice of W-H plane at every disparity step. Illustration of the parallelized cost aggregation is displayed in Figure 5.8. Three path directions are shown for horizontal (east), vertical (south), and diagonal (southeast). The rest of cost paths are similar. According to (5.18), the aggregated value at a voxel in the cost volume is dependent on its immediate neighbors preceding to it along the path at current disparity,

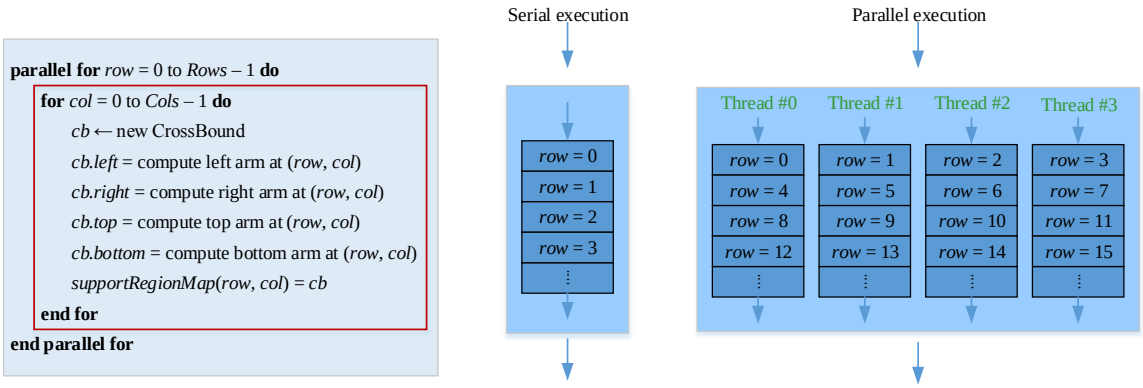


Figure 5.7: Example of the parallelization of adaptive support region computation through cross bound on a quad-core processor. The codes listed at left shows the nested loops that iterate through every pixel to compute the cross bounds. The codes highlighted in red are treated as a code block that is executed for an individual row of pixels. Each thread in the parallelized execution chain takes one forth of the total work load.

the disparity that is one step less, and the disparity that is one step greater, namely the $L_{\mathbf{g}}(\mathbf{p}-\mathbf{g}, d)$, $L_{\mathbf{g}}(\mathbf{p}-\mathbf{g}, d-1)$, and $L_{\mathbf{g}}(\mathbf{p}-\mathbf{g}, d+1)$ as is defined in (5.18). However, within a parallel computing framework, the order of execution of the same operation on multiple data cannot be predetermined during the programming stage, and it is handled by the operating system's task scheduling mechanism at the run time. This data dependency calls for careful design of our parallel algorithm to ensure that all the required data have been updated when following the path to compute the aggregated costs.

It is clear from (5.18) that the computation of aggregated cost at a new pixel location is dependent on its predecessor pixel location along the path. This suggests the use of a synchronization mechanism among all thread once they finish updating one pixel along the path. Before the start of cost aggregation at a new direction, all the header pixels that define the beginnings of each path along current direction are collected (edges highlighted in red in Figure 5.8). Each thread then take one row of voxels of the cost volume at a specified disparity step, and update the aggregated cost at the new location. Once updates are done, all threads synchronize and get ready to move on the next voxel along the path.

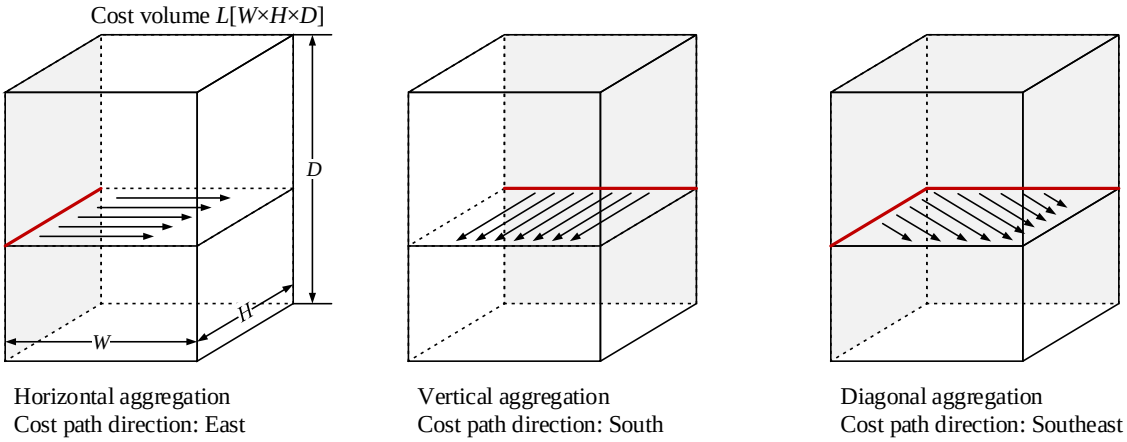


Figure 5.8: Parallelized cost aggregation. Edges highlighted in red on the cross-sectional slice indicate the header pixels for all paths. Three path directions are shown, others are similar. The shaded surfaces on the cost volume represent the voxels serve as path headers for each aggregation direction at each disparity step.

This procedure can be visualized as the shaded faces as indicated in Figure 5.8 shift along the path direction one layer at a time when updating the cost volume. The details of the parallel algorithm is illustrated in Algorithm 1.

5.4.2 Performance Evaluation

We tested our matching cost computation and aggregation algorithms with the Middlebury stereo images illustrated in Figure **??**. The test platforms are two desktop computers. One is with an Intel© quad-core 2.8 GHz CPU and dual-channel 8 GB system memory, and the other is with a high-end Intel© hex-core 3.2 GHz CPU and quad-channel 8 GB system memory. Both processors feature Intel's proprietary simultaneous multi-threading technology marketed as Hyper-Threading, through which the operating system addresses two virtual or logical cores for each processor core that is physically present. Workloads is shared between these logical cores when possible. This feature enables the number of threads to be run concurrently two times as many as the number of

Algorithm 1: Parallel implementation of matching cost aggregation

```
1  initialize a new cost volume L_g [W × H × D];
   /* header pixels are highlighted in red from Figure 5.8                  */
   /* they remain the same across all slices of the cost volume             */
2  ptHeaders ← CollectHeaderPixels(pathDirection);
3  pathLengthMax ← GetLongestPath(pathDirection);
   /* the aggregation at each pixel along the path has to be computed       */
   /* sequentially, from 0 → pathLengthMax                                  */
4  for i ← 0 to pathLengthMax do
       /* selecting a slice at depth d is done concurrently                 */
       /* entering parallel region                                          */
5      parallel for d ← dispMin to dispMax do
6          foreach p0 ∈ ptHeaders do
7              if i ≥ GetCurrentPathLength(p0, pathDirection) then
                   /* reached to the end of current path                    */
8                  break;
9              ptCurr ← GetCurrentPixelLocation(p0, pathDirection);
10             ptPrev ← GetPreviousPixelLocation(p0, pathDirection);
11             (x, y) ← ptCurr;
12             L_g [x, y, d] ← ComputeAggregatedCost(ptCurr, ptPrev, d);
       /* leaving parallel region                                           */
```

physical cores on the processor, that is eight threads on the quad-core system and twelve threads for the hex-core system. Our cost computation and aggregation algorithms are developed in C++ programming language, and the multi-threading is done through OpenMP parallelized loops.

The number of concurrent threads for a parallel code region can be controlled by OpenMP's API call `omp_set_num_threads()`. Thread numbers from 1 to 8 are tested on the quad-core system, and numbers from 1 to 12 are tested on the hex-core system. The multi-threading speedup is calculated by computing the ratio of the processing time in serial codes to the time in parallel codes. According the Amdahl's law [144] of theoretical maximum speedup using multiple processors, the speedup that can be achieved by executing

a given algorithm on a system capable of executing n threads of execution is

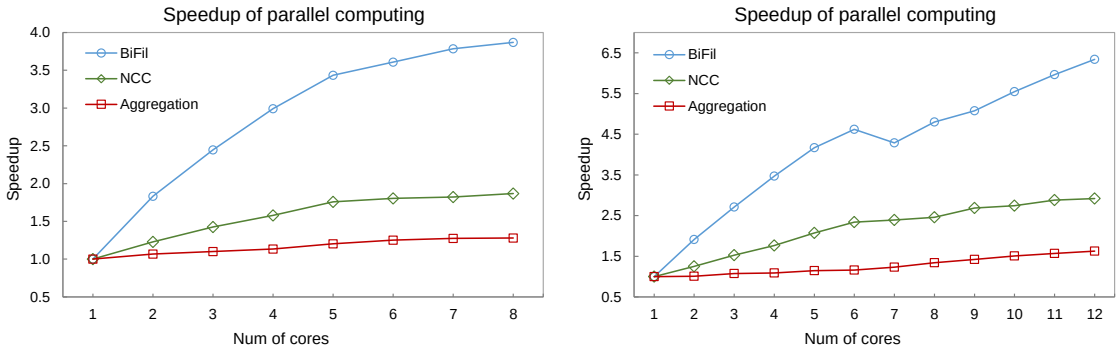
$$S(n) = \frac{T(1)}{T(n)} = \frac{T(1)}{T(1)\left[B + \frac{1}{n}(1-B)\right]} = \frac{1}{B + \frac{1}{n}(1-B)}, \tag{5.21}$$

in which $n \in \mathbb{N}$ is the number of threads of execution, $B \in [0,1]$ is the fraction of the algorithm that is strictly serial, and $T(n)$ is the time an algorithm takes to finish when being executed on n thread(s) of execution. $T(1)$ is the time the algorithm takes to run in strictly serial, and is taken as the reference to calculate the speedup.

The speedup calculated by Amdahl's law is the *theoretical* maximum. Actual values are usually lower than the theoretical values, because of various factors that affect the performance, such as system overheads to initiate parallelism, cache miss in fetching data for execution, memory bus bandwidth, etc. In our implemented algorithm, the processing time is recorded by invoking the `time()` system call immediately before and right after the code blocks of interest.

We execute the serial version and parallel version of the same algorithms eight times on the *Cones* and *Aloe* images. The average of speedups at each thread number setting are calculated. Figure 5.9 shows the graphs of the speedups measured on three sub-algorithms: the bilateral filtering of the input images, the accelerated computation of NCC on adaptive supports, and the cost aggregation. It is clear that performance gain was achieved by applying multiple threads in the computation, however, the amount of speedups varies among the three sub-algorithms on our two test systems.

The bilateral filtering (BiFil) shows the highest potential in speed gains when converting into multi-threading. The speedup is almost linear with respect to the thread numbers, especially when the thread number does not exceed the number of physical cores on the processors (thread count from 1 to 4 on the quad-core system, and thread count from 1 to 6 on the hex-core system). On our quad-core system, the speedup of the BiFil tends to



(a) Speedups on a quad-core system.

(b) Speedups on a hex-core system.

Figure 5.9: Performance analysis of parallel computation on multi-core desktop computers. Performance was evaluated on three sub algorithms: bilateral filtering (BiFil), fast computation of NCC, and cost aggregation.

level off after logical cores are engaged in the computation. And it reaches to 3.4x speed gain at our predefined maximum thread number. The speedup curve of the BiFil on our hex-core system shows an interesting feature that the trend of the curve breaks off when the number of threads reaches to 7, when one of the physical core has to run two threads concurrently while the others run one thread. There is performance loss at this point, but the speed gain picks up and the trend was remained. It finally reaches to 6.3x speedup at the maximum thread count, showing less level off than on the quad-core system.

The NCC computation and the cost aggregation achieve less speedup comparing to the BiFil. While NCC reached to the highest speedup at 1.8x on quad-core and 2.8x on hex-core, the cost aggregation only reached to 1.3x on quad-core and 1.7x on hex-core. This can be explained that the NCC and the aggregation are more data-intensive than the BiFil. The BiFil runs on a two dimensional image, while the NCC and the aggregation run on a three dimensional cost volume, which is close to a hundred of times larger then two dimensional image for our test data. This requires frequent access to the memory for read and write operations. Memory access is very slow compared to the arithmetic operation on the processor, and it results in hundreds of idle CPU cycles time waiting for the data to

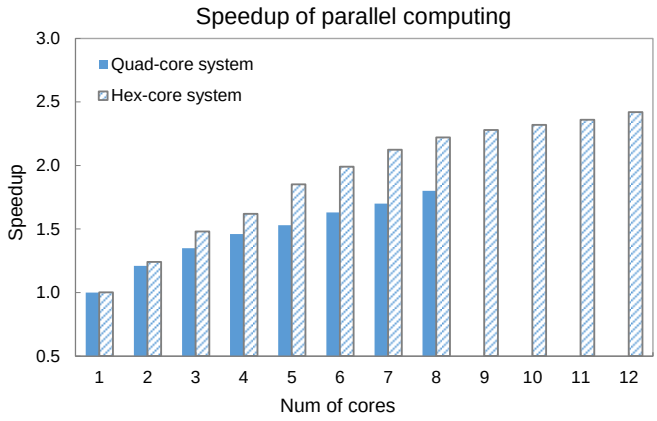


Figure 5.10: The total speedups of cost computation and aggregation on our quad-core and hex-core test systems.

be fetched from memory. In addition, BiFil is more cache friendly because several rows of image pixels can be fitted into the high-speed cache, so that fetching the data in the cache is almost as fast as the arithmetic operations on the core. This, in turn, reduces the chances of accessing the slow system memory for data. On the other hand, the cost aggregation not only need to access the cost volume slice at depth d (5.18), it also has to access the slices at depth $d-1$ and $d+1$ for one aggregated cost value for pixels at depth d. The slices at $d-1$ and $d+1$ are stored far away from slice at d in the memory. All these three slices are too large to be fit into the cache, thus more frequent memory accesses are required in the cost aggregation, and it has the lowest speed gain among all the three sub-algorithms.

For calculating the overall speedup of our parallelized cost computation and aggregation algorithms, we compare the *total* times they take to finish the computation at each different core-count setting. Amdahl's law tells us that the overall speedup is related to the fraction of code that cannot be parallelized. In our case, only the main computation is parallelized, the rest runs in serial, these include memory allocation, data initialization, and other operations that require a great effort to parallelize but does not account for too much speedup. In addition, the speedup of each individual sub-algorithm that has been

parallelized varies a lot, therefore the overall speedup also depends on the fraction of each sub-algorithm with respect to the whole program. For example, the execution time of BiFil sub-algorithm only takes up a small fraction of the total processing time. Even though it can achieve a relatively high speedup, its contribution to the overall speedup is very limited. On the other hand, the cost aggregation takes up about 50% of the total computation time, the speedup of the cost aggregation sub-algorithm has a significant impact to the overall speedup. We observed a maximum of speedup at 1.8x and 2.4x total speedups (Figure 5.10) on our quad-core (8-thread) and hex-core (12-thread) systems, respectively. Our hex-core system reached a higher speed-gain then the quad-core system as expected, due to 50% more physical cores available from the processor. Furthermore, the hex-core system features the quad-channel memories while the quad-core system is dual-channel in configuration. This boosts the memory bandwidth to twice as much as the dual-channel system, and effectively reduces data traffic confession between the processor and the main memory, allowing additional cores on the hex-core system to spend less time waiting for data and be more efficient in calculation.

5.5 Summary

In this chapter, we briefly reviewed popular methods for matching cost computation and aggregation, and then proposed a robust cost computation algorithm that combines three components of matching costs: the cost of color difference from background suppressed images, the cost of census, and the cost computed from normalized cross-correlation (NCC). A multi-path cost aggregation framework was also introduced.

The stereo images captured by our proposed imaging system exhibit features such as different brightness levels due to different gain and bias between cameras, surfaces with rich geometric changes, and textureless regions. Images with different brightness

levels call for a matching strategy that is insensitive to the absolute color values, and require the cost function to rely on texture information instead of color information. Since NCC computes the *correlation* of neighborhood pixels with respect to the kernel between an image pair, the matching cost computed from NCC is insensitive to different gain and bias settings. The cost from census transform is non-parametric because it generates the matching cost based on the relative ordering of color values between the kernel pixel and its neighbors, thus it is also insensitive to brightness level. However, NCC and census are all window-based cost function. A windows-based function usually degrades on non-frontal-parallel surfaces, because it assumes constant disparity within the window. This may introduce large error when computing the costs for feature points at depth discontinuity. A point-based cost function, the background suppressed absolute color difference, is included in our combined cost evaluation. The point-based cost term also works well for surfaces with rich geometrical changes, because it does not take into account any neighboring pixels, which may have different disparities on steep or curvy surfaces.

Our multi-path cost aggregation method simulates the concept of global energy minimization by finding a disparity path that yields the lowest total costs. It adds penalty to disparity changes along the path, awarding the disparity values that would result in a smooth surface. The cost aggregation suppresses noise and prevents incorrect matching that yields a cost value lower than the real match.

The cost computation and aggregation are both computation-intensive tasks, because they both work on a three-dimensional cost volume. To improve the algorithm performance, we applied parallel computing technique, allowing the computation tasks to be distributed among multiple cores on the processor and to be executed simultaneously. Our performance analysis show that the overall speedups of 1.8x and 2.4x were reached on a quad-core and a hex-core desktop computer.

Chapter 6

Disparity Computation and Refinement

A disparity map computed from the aggregated cost volume may contain outliers in occlusion regions and mismatched regions, especially at the depth discontinuity. This requires additional step to detect occlusion regions and correct mismatches. A final disparity refinement step is also needed to interpolate the disparity map to achieve subpixel accuracy. This chapter describes the third and the forth step in a stereo matching framework: the disparity computation/optimization, and disparity refinement.

6.1 Related Work

Disparity computation and optimization refers to the methods of assigning a correct disparity value to a pixel. Local method through Winner-Takes-All (WTA) strategy can be implemented to be very efficient and sometimes may meet the requirement of real-time application. But they are more prone to noise and local ambiguity in textureless regions and occluded areas, because only local information collected from a small neighborhood of pixels contributes to the selection of disparities.

In contrast, global method make explicit assumptions about the scene that the imaged surfaces are piecewise smooth. This assumption is generally true and the constraint used to enforce piecewise smooth is referred to as the *smoothness constraint* in the stereo vision literature. Global methods are usually formulated in an energy-minimization framework. The standard and classical global stereo formulation aims to find an optimal dispar-

ity assignment function $f(p)$ that minimizes the following energy function

$$E(f) = E_{\text{data}}(f) + \lambda \times E_{\text{smooth}}(f), \tag{6.1}$$

where the data term $E_{\text{data}}(f)$ comes from the matching cost and penalizes disparity assignments that are inconsistent within the pair of stereo images, whereas the second term, the smoothness term $E_{\text{smooth}}(f)$, imposes the spatial coherence of labeling the disparity within a defined neighborhood. It enforces piecewise smoothness by encouraging neighboring pixels to have similar disparities. λ is a weight that adjusts the contribution of the smoothness term. In general, global energy minimization involves more computation than local methods. To make the optimization computationally affordable, the smoothness energy is often defined with a small neighborhood, e.g., using the common Potts model [141] or the truncated linear model [125]. Once the global energy function has been formulated, the lowest energy corresponding to the optimal disparity assignment can be solved using the methods surveyed by Szeliski *el al.* [69].

The strategies for finding the minimum of the global energy function differ. Belief Propagation (BP) [125, 126] and Graph Cut (GC) [79, 127] are two popular choices among stereo researchers. By applying various smoothness constraint and selecting robust cost functions, BP- and GC-based stereo methods were reported to produce state-of-the-art results in terms of 3D scene depth accuracy [90, 145, 146]. In contrast to BP and GC which approximate the global minimum of the energy defined in (6.1) over the two-dimensional pixel grid, the Dynamic Programming (DP) [96, 123, 124] finds the global minimum for each image scan line independently. The DP approach reduces the amount of computation and results in polynomial time complexity. The main problem with DP is the difficulty of enforcing disparity consistency between scan lines and it commonly leads to streaking effects.

Global methods are less sensitive to noise and textureless regions and are in general more robust than local methods since prior constraints provide regularization for regions difficult to match. However, global methods are usually more computationally intensive than local methods.

Disparity refinement is usually done as a post-processing for removing peaks and isolated values, interpolating gaps, or increasing the accuracy by subpixel interpolation. Occluded regions are usually detected using left-right consistency check [91,147,148] and unmatched pixels can be filled via interpolation or depth completion algorithms [149,150]. Median filter can also be used to remove small isolated mismatches. Due to the low computational cost and the edge-preserving property, median filtering particularly favored in real-time stereo algorithms as a post-processing step. Most stereo algorithms generate disparity estimates in discretized integer space. While integer disparities may be sufficient for application such as segmentation and object tracing, for view synthesis, 3D reconstruction and measurement, integer disparity maps usually result in stepped surface and unappealing visual artifacts. To overcome this limitation and improve the resolution of the disparity map, many stereo algorithms utilize a subpixl refinement stage to generate subpixel-accurate disparity values. One of the standard method is to fit a parabolic or Gaussian curve [151] to the matching costs defined at discrete values. Symmetric refinement can also be done by fitting a parametric surface over a 2D neighborhood of the matching cost function [152].

6.2 Disparity Computation and Optimization

The result of our previous matching cost computation and cost aggregation is a three-dimensional cost volume $S(\mathbf{p}, d)$, in which each voxel represents the *aggregated* cost of assigning a disparity value d to a pixel $\mathbf{p}$ that is indexed in the reference image. Our cost

aggregation step minimizes a global energy function so that an optimal disparity map can be found by selecting a disparity value at each pixel that yields the minimal cost at this pixel location. This procedure implements the WTA strategy.

The disparity map D_b that corresponds to the base image I_b is determined by selecting for each pixel $\mathbf{p}$ the disparity d that corresponds to the minimum cost, that is

$$D_\mathrm{b}(\mathbf{p}) = \arg\min_d S(\mathbf{p}, d). \tag{6.2}$$

The disparity map D_m that corresponds to the match image I_m can be determined from the same costs by traversing the epipolar line that corresponds to the pixel $\mathbf{q}$ of the match image. The same procedure can be used to determine the d, that is d is selected with the minimum cost

$$D_\mathrm{m}(\mathbf{q}) = \arg\min_d S[e_\mathrm{mb}(\mathbf{q}, d), d], \tag{6.3}$$

where $e_\mathrm{mb}(\mathbf{q}, [\cdot])$ is the epipolar line in the base image that corresponds to the pixel $\mathbf{q}$ in the match image, and $e_\mathrm{mb}(\mathbf{q}, d)$ is the matched pixel in the base image. Since the cost aggregation relies on a reference image, which is the base image in our case, it does not treat the base and match images symmetrically. Slightly better results can be expected, if D_m is calculated separately, that is, by performing pixelwise matching and aggregation with I_m as the *reference* and I_b as the *match*.

The calculation of D_b and D_m permits the determination of occlusions and mismatches by performing a *left-right* consistency check. The left-right check ensures that the matching needs to be bijective: if $\mathbf{p}$ in the base image I_b matches to $\mathbf{q}$ in the match image I_m, then $\mathbf{q}$ must also matches to $\mathbf{p}$. Or, in mathematical form, $D_\mathrm{b}(\mathbf{p}) = -D_\mathrm{m}(\mathbf{q})$. To take different foreshortening into account, we tolerate a disparity mismatch of up to one disparity step in our implementation. A disparity is set to invalid ($D_\mathbf{inv} = 0$, which represents

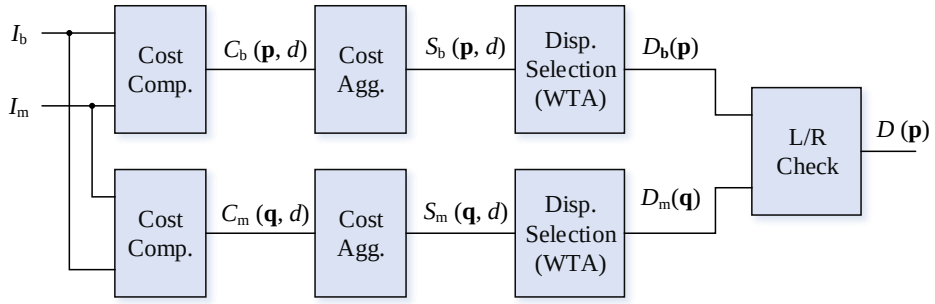


Figure 6.1: Summary of processing steps for matching cost computation, aggregation, and disparity computation.

infinite distance in the 3D space), if both differ by more than one:

$$D(\mathbf{p}) = \begin{cases} D_{\mathrm{b}}(\mathbf{p}), & \text{if } |D_{\mathrm{b}}(\mathbf{p}) + D_{\mathrm{m}}(\mathbf{q})| \geq 1 \\ D_{\mathrm{inv}}, & \text{otherwise} \end{cases} \tag{6.4}$$

The consistency check enforces the *uniqueness constraint*, by permitting one-to-one match only. The disparity computation and consistency check requires visiting each pixel at each disparity a constant number of times, thus is linear in complexity. The process of generating a validated disparity map is now complete, and a summary of all precessing steps is given in Figure 6.1.

6.3 Disparity Refinement

Even with the left-right consistency check, the disparity map computed from previous step can still contain certain kinds of errors. Furthermore, there are generally areas of invalid disparity values that need to be corrected. The post-processing procedures described in this section is designed to handle these issues.

6.3.1 Removal of Isolated Regions

Disparity map can contain small areas of wrong disparities, due to reflection, low texture, and noise. They usually show up as small patches of disparity that is very different

Figure 6.2: Errors in disparity map. Black regions on the disparity map: pixels with invalid disparity; highlighted region centered at the edge of a cone: untextured background; highlighted region on the box: isolated region;

from their neighboring pixels', as shown in Figure 6.2. Depending on the structure of the scene, a threshold value of the size of small disparity patches can be predefined such that smaller patches are unlikely to represent valid fine structure of the scene.

For identifying isolated regions, a segmentation method is applied by allowing 4-connected neighboring disparities within one segment to vary by one disparity step. The disparity patches of all segments below a certain size are set to invalid, and are to be either interpolated or extrapolated by the following steps.

6.3.2 Intensity Consistent Disparity Validation

6.3.2.1 Problem Definition

For most of indoor scene, it is common that foreground objects are in front of a low textured or textureless background. For example, the highlighted region in Figure 6.2. This is also the case for our body images that the background is mostly solid colored walls. Our energy function $E(D)$ for cost aggregation as shown in (5.16), however, does not exhibit a preference for the disparity value at different regions. It is unaware of the type (foreground or background) of current surface, thus it does not differentiate between placing a disparity step correctly just next to a foreground object, or a bit further away within a textureless

background. Section 5.3.3 suggests applying an adaptive penalty P_2 that is consistent with the intensity change. This helps placing a correct disparity step next to a foreground object, because this location coincides with only small intensity change.

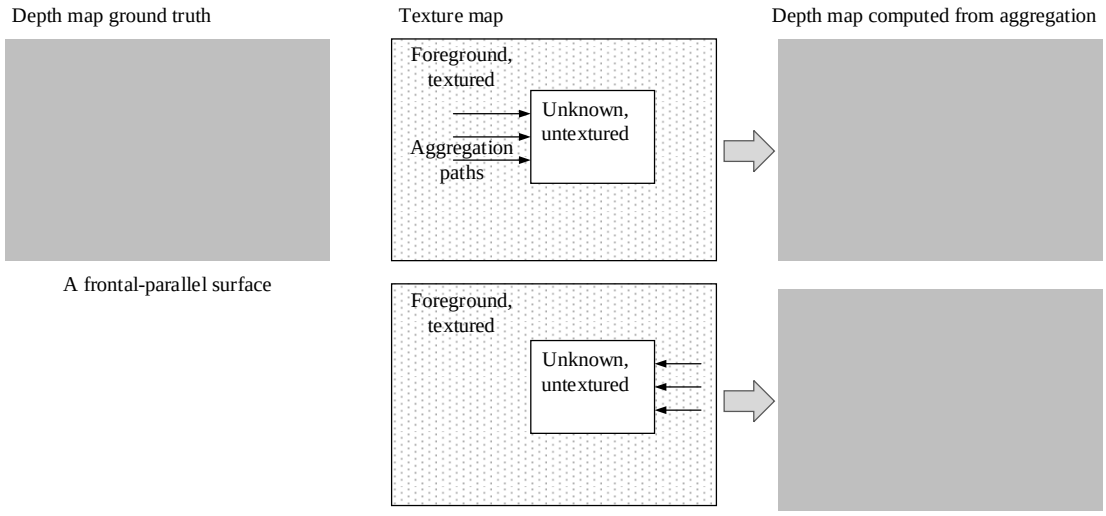
However, our adaptive cost aggregation applies the energy function not in 2D over the whole image but along multiple individual 1D paths from all directions and then computes the summed energy. Depending on the location and direction of 1D aggregation paths, they may encounter textured foreground or background objects around an textureless region (Figure 6.3), in which case two different approaches are needed to determine disparity values within the textureless region. If the disparity value stays constant along the aggregation direction as shown in Figure 6.3a, the same disparity value should be assigned to the textureless region because the aggregation most likely pass through a frontal-parallel surface. Otherwise if the surface is slanted as shown in Figure 6.3b, interpolation may be needed to fill in disparity values within the textureless region.

Textureless areas may have different shapes and sizes and can extend beyond image borders. This is quite common for our body images with backgrounds are usually walls. For these cases, the 1D aggregation paths may also encounter either foreground or background texture, or leave the image with the textureless areas in which case no disparity values would be placed. Summing all those inconsistent paths may easily lead to fuzzy discontinuities around foreground objects in front of textureless background.

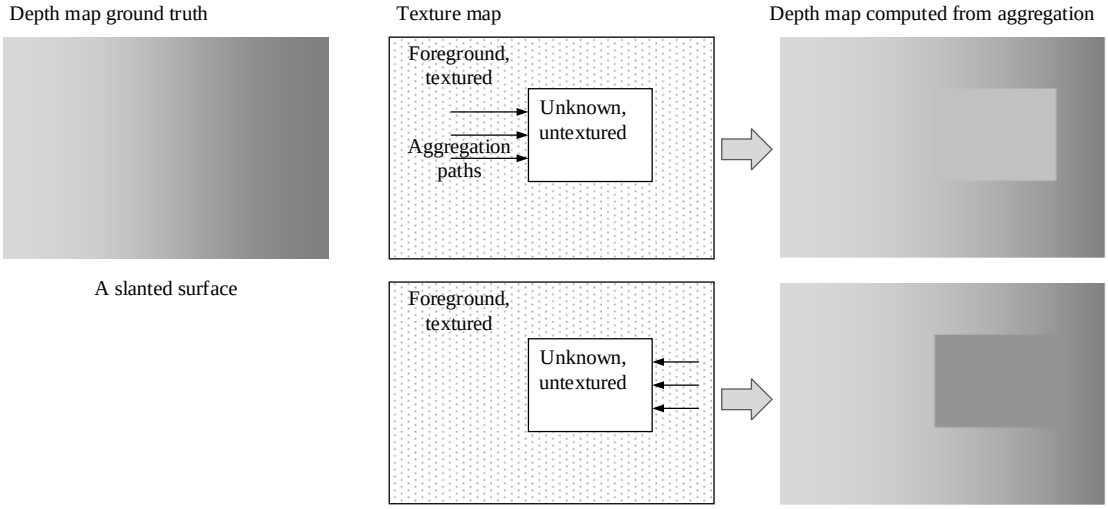
6.3.2.2 Assumptions

This feature of our multiple 1D aggreagtion paths based method calls for special care that only applies to certain scenes in structured environments. To present a solution to this, we may need to make some general assumptions:

1. Disparity discontinuity do not occur within textureless regions.



(a) An untextured region on a frontal-parallel surface results in a set of consistent disparity maps.



(b) An untextured region on a slanted surface (with continuously changing disparities) results in a set of in consistent disparity maps along two opposite aggregation directions, indicating the disparities should vary within the untextured region.

Figure 6.3: Examples of disparity selections along aggregation paths.

2. There should be some visible texture somewhere on an generally textureless surface.

3. The surface of a textureless area can be approximated by a plane.

The first assumption is mostly correct, since depth discontinuities usually cause at least some visible change in intensities. Otherwise, the discontinuity would be unde-

tectable. The second assumption is necessary, because the disparity of an absolutely untextured background surface would be impossible to be detected. The third assumption is the weakest among all three. Its justification is that a textureless surface with varying distance to the viewpoint usually appears with varying intensities. Thus, piecewise constant intensity can be treated as piecewise planar.

6.3.2.3 Solution

Untextured areas are identified by a fixed-parameter Mean Shift Segmentation [153] on the base image I_{b}. A small variance of radiometric difference σ_r is applied, so that intensity changes below this value is are treated as noise. The variance of spatial distances σ_s is also set to a low value for fast processing. Both σ_r and σ_s are empirically chosen, and we found $\sigma_r = 3$ and $\sigma_s = 5$ are sufficient for good segmentation. Furthermore, all segments that are smaller than a certain threshold, for example 20 pixels, are ignored, because small untextured areas are expected to be handled well by our adaptive cost aggregation.

After our cost aggregation and disparity computation, disparity discontinuities may occur in untextured areas. Thus, these areas are expected to contain incorrect disparities of the foreground object and correct disparities of the background, if the background surface contains at least some texture (Assumption 2). This leads to the state that some disparities within the i-th segment S_i are correct. Thus, several hypotheses for the correct disparity of S_i can be identified by segmenting the disparities within the S_i. This is done by simple segmentation with *smoothness constraint*, that is by allowing neighboring disparities within one segment to vary by one disparity step. This simple yet fast segmentation results in several segments S_{ik} with each S_i.

The next step is to create the surface hypotheses F_{ik} by calculating the best fitting planes (Assumption 3) through the disparities of Sik. Very small segments, for example,

less than 12 pixels, are ignored, as it is unlikely that such small patches belong to the correct hypothesis. Then each hypothesis is evaluated within the patch S_i by replacing all pixels of S_i by the surface hypothesis and calculating E_{ik} as defined in (5.16) for all non-occluded pixels of S_i. A pixel $\mathbf{p}$ is considered to be occluded if another pixel with higher disparity maps to the same pixel $\mathbf{q}$ in the match image. This detection is performed by first mapping $\mathbf{p}$ into the match image by $\mathbf{q} = e_{bm}(\mathbf{p}, D(\mathbf{p}))$. Then the epipolar line of $\mathbf{q}$ in the base image $e_{mb}(\mathbf{q}, d)$ is followed for $d > D(\mathbf{p})$. The pixel $\mathbf{p}$ is occluded if the epipolar line passes a pixel with a disparity larger than d. More details of determining whether a pixel is an occluded pixel or a mismatched pixel can be found in Section 6.3.3.

For each segmented patch S_i, the surface hypothesis F_{ik} with the minimum cost E_{ik} is chosen, that is

$$F_i = F_{ik'} \text{with} k' = \arg\min_k E_{ik}. \tag{6.5}$$

All disparities within S_i are replaced by values on the chosen surface for making the disparity selection consistent to the intensities of the base image, fulfilling Assumption 1:

$$D'(\mathbf{p}) = \begin{cases} F_i(\mathbf{p}), & \text{if } \mathbf{p} \in S_i \\ D(\mathbf{p}), & \text{otherwise} \end{cases} \tag{6.6}$$

The above approach is similar to some other methods [120, 147, 154] as it refines an initial disparity map through image segmentation and plane fitting. Compared to other methods, the initial disparity map generated by our multi-path adaptive cost aggregation is quite accurate already so that only untextured areas above a certain size are modified. Another difference is that disparities of the considered area are selected by considering a small number of hypotheses that are inherent in the initial disparity map. There is no time-consuming iteration involved.

(a) The mean-shift segmented image defines surfaces with constant disparities.

(b) Before the disparity replacement.

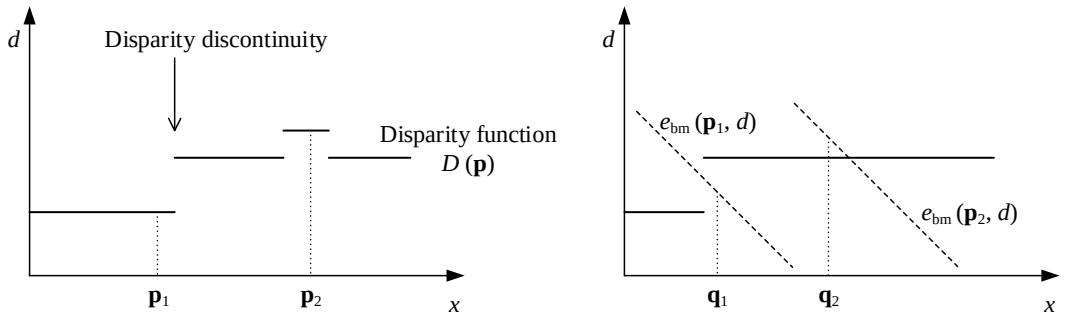
(c) After the disparity replacement.

Figure 6.4: Result of the intensity consistent disparity selection. Ambiguous disparities within a texture less region is replaced by disparities matched with high confidence within the same region.

6.3.3 Discontinuity-preserving Interpolation and Extrapolation

The left-right disparity consistency check of Section 6.2, as well as isolated region filtering of Section 6.3.1 may invalidate some disparities. There are also outliers in occluded regions and depth discontinuities. These lead to holes in the disparity map, which need to be properly fixed for a dense stereo matching result. After detecting these outliers, the simplest strategy is to fill them with reliable disparities [67], which is only useful for small occluded regions.

Invalid disparities can be classified into occlusions and mismatches. The treatments for both case must be conducted differently. Occlusions must not be interpolated from the foreground, but only from the background to avoid the extension of the foreground surface into the occluded regions. Thus, an extrapolation of the background into occluded regions is desired. In contrast, holes caused by mismatches can be smoothly interpolated from all neighboring pixels.



(a) Disparity map with base image as the reference.

(b) Disparity map with match image as the reference.

Figure 6.5: Differentiating between occluded pixels and mismatched pixels.

6.3.3.1 Occlusion Detection

Occlusions and mismatches can be distinguished as part of the left-right consistency check. Figure 6.5 shows the concept of differentiating between occluded pixels and mismatched pixels with the assistance of epilines. In Figure 6.5a, $\mathbf{p}_1$ and $\mathbf{p}_2$ are two pixels on the same row in a disparity map generated with base image as the reference, while $\mathbf{q}_1$ and $\mathbf{q}_2$ are determined by $\mathbf{q}_1 = \mathbf{p}_1 + D_{\mathrm{b}}(\mathbf{p}_1)$ and $\mathbf{q}_2 = \mathbf{p}_1 + D_{\mathrm{b}}(\mathbf{p}_1)$. The occluded pixel $\mathbf{p}_1$ in the base image goes through discontinuity that causes the occlusion. Its epiline in the match image does not intersect the disparity function $D_{\mathbf{m}}$, indicating that there is no pixel in the match image that matches to $\mathbf{p}_1$. Thus, $\mathbf{p}_1$ is an occluded pixel. In contrast, the epiline of $\mathbf{p}_2$ in the match image intersects with D_{m}. However, the intersection point does not coincide with $\mathbf{q}_2$, indicating that $\mathbf{p}_2$ and $\mathbf{q}_2$ are a pair of mismatched pixels. Therefore, for each invalidated pixel, an intersection of the corresponding epiline with D_{m} is sought, for marking it as either occluded or mismatched.

6.3.3.2 Iterative Region Voting

The detected outliers, either occluded pixels or mismatched pixels, should be filled with reliable neighboring disparities. We process these outliers with the constructed cross

based support regions and a robust voting scheme. For an outlier pixel $\mathbf{p}$, all the reliable disparities in its support region are collected to build a histogram $H_{\mathbf{p}}$ with $d_{\max}+1$ bins. The disparity with the highest bin (most voted) is denoted as $d^*_{\mathbf{p}}$, and the total number of the reliable pixels is denoted as $Q_{\mathbf{p}} = \sum_{d=0}^{d=d_{\max}} H_{\mathbf{p}}(d)$. The new disparity of $\mathbf{p}$ is then updated with $d^*_{\mathbf{p}}$ if enough reliable pixels and votes are found in the support region, that is

$$Q_{\mathbf{p}} > \tau_Q, \tag{6.7}$$

and

$$\frac{H_{\mathbf{p}}(d^*_{\mathbf{p}})}{Q_{\mathbf{p}}} > \tau_H, \tag{6.8}$$

where τ_Q and τ_H are two threshold values. To precess as many outliers as possible, the voting process runs for 6 iterations. The filled outliers are marked as *reliable* pixels and used in the next iteration, such that valid disparity values can gradually propagate into occluded regions.

This strategy works well for filling outliers in textureless regions, because the support of a pixel within this region is usually quite large. Thus it has a good chance to include more disparities from reliable pixels for the voting procedure, yielding a more accurate estimation of the disparity value within the occluded region.

6.3.3.3 Depth Consistent Extrapolation

Unlike the iterative region voting, the rest of outliers are filled with an extrapolation strategy that treats occlusion and mismatch differently. For an outlier pixel $\mathbf{p}$, we find the nearest reliable pixel in 16 directions around $\mathbf{p}$. If $\mathbf{p}$ is an occluded pixel, the pixel with the lowest disparity value (furthest distance in depth) is selected for extrapolation, since $\mathbf{p}$ most likely comes from the background; otherwise if $\mathbf{p}$ is a mismatched pixel, the pixel with the most similar color in its 8-neighbor region is selected for extrapolation. With

region voting and extrapolation, most outliers are effectively removed from the disparity results.

6.3.3.4 Depth Discontinuity Adjustment

In this step, the disparities at the depth discontinuities are further validated and refined with the information form neighboring pixels. All edges in the disparity map is first detected with two separable masks, which detect disparity change in both horizontal and vertical direction. Then, for each pixel $\mathbf{p}$ on the disparity edge, two pixels $\mathbf{p}_1$ and $\mathbf{p}_2$ from both sides of the edge are identified. The new disparity at pixel $\mathbf{p}$ is replaced either by $D(\mathbf{p}_1)$ or $D(\mathbf{p}_2)$ if one of the two pixels corresponds to a lower aggregated matching cost than $S(\mathbf{p}, D(\mathbf{p}))$. This procedure has to be run a few times to allow the new edge to converge to real depth discontinuity. However, this method can be made more efficient after the first iteration, because only the updated edge pixels requires further validation. An edge mask can be used to identify these edge pixels. This method helps to reduce the small errors around depth discontinuities.

6.4 Summary

This chapter covers the third and the forth step within a stereo matching framework, namely, the disparity computation and disparity refinement. Disparity computation focuses on assigning a correct disparity value to each pixel in the disparity map, based on the previously generated matching cost volume. Although a global method usually generate more reliable disparity map, for performance consideration, we proposed to use a winner-takes-all strategy and select the disparity value that corresponds to the lower cost at each pixel local. This method works well because our semi-global cost aggregation step has already take the neighbor information into account, thus the cost volume has relatively reliable matching cost values. Furthermore, errors in assigning an incorrect disparity value

to a pixel can have a change to be validated by the subsequent disparity refinement step.

The disparity refinement step serves as the last stage of the stereo matching pipeline to correct any errors in the computed disparity with various constraint, such as surface smoothness, color consistency. These constraints can be enforced by certain rules in validating a disparity value, for example, disparity discontinuity cannot occur within textureless regions, and the surface of a textureless area can be approximated by a plane on which some visible texture may be visible somewhere. The disparity refinement procedure first identify the occlusion regions on a disparity map, because the occluded pixels and mismatched pixels need to be handled differently. Then the mismatched areas and the occluded areas undergo an iterative region voting and depth consistent extrapolating, which allows reliable disparity values from a neighborhood region to propagate into the problematic areas. Finally, the disparity edges are checked and made consistent to the texture map.

Chapter 7

3D Body Model Generation

The result of stereo matching is 2D disparity map from which a dense 3D point cloud can be recovered with the known stereo geometry. The raw 3D point data are usually comprised of hundred thousands of scattered 3D points, and is hard to handle efficiently for measurement and rendering. This chapter presents the technique to effectively reduce the density of the data through 3D surface reconstruction, a method that converts dense 3D points into triangle mesh with proper surface approximation. A highly accurate, sub-pixel refinement procedure is performed on the discrete disparity map before the surface reconstruction is applied. The sub-pixel refinement recovers fine geometrical details, and it suppresses noise and enforces smoothness, which significantly improves the quality of the dense point cloud and provides an accurate input for surface reconstruction.

7.1 Sub-pixel Disparity Refinement

So far, we have presented detailed steps in Chapter 5 and Chapter 6 to compute the disparity map which corresponds to the 3D surface geometries in the scene. However, this disparity map takes discrete values and is not sufficient to recover fine geometric details. A dedicated disparity refinement process is needed to achieve sub-pixel accuracy.

Disparity refinement is often performed in an iterative fashion. A simple and straightforward way to implement sub-pixel refinement is to interpolate the matching cost at the previous, current and next disparity step, i.e., the cost at $d-1$, d, and $d+1$ to find the local minimum of the matching cost at sub-pixel disparity step. This strategy has been

adopted in the work of [103] for face stereo imaging with added surface smoothness constraint. Although less computation is involved within each iteration, the interpolation method does not account for texture match when updating the disparity value, thus is less effective in converging into the ideal disparity value and it may take more than a hundred of iterations to produce a smooth surface [103].

Refinement can also be done through curve fitting [90, 155]. But curve fitting usually suffers from systematic error called "pixel-locking" effect in which disparity values are pulled towards discrete values [90]. Research efforts have been made to address this problem. For example, Nehab *et al.* [152] suggested symmetric refinement by fitting a parametric surface over a 2D neighborhood of the matching cost function. Stein *et al.* [156] proposed an iterative refinement method that is essentially based on Lucas-Kanade algorithm [157]. These aforementioned improvements are all focused on reducing the "pixel-locking" effect and make disparity refinement on each individual pixel independently. However, in practice, like all other local methods, the result is prone to be noisy. Thus, it is a good practice to take the spatial coherence into account during disparity update.

7.1.1 Local Sub-pixel Estimation

The sub-pixel refinement method developed in this study adopts the iterative refinement framework proposed in [14], and here we show that its performance can be improved by introducing the bilateral filter for fine geometric detail enhancement. The iterative refinement works at a global level within a regularization framework. To begin with, the amount of update is estimated locally for each pixel. The estimation can be made by minimizing the hybrid matching function defined in (5.1) by

$$\Delta d = \arg\min_{\Delta d} C(x, y, d + \Delta d) = \arg\max_{\Delta d} \rho(x, y, d + \Delta d), \tag{7.1}$$

where d is the current disparity value, and Δd is the amount to be updated. Equation (7.1) is difficult to solve since the correlation function ρ is highly nonlinear. Although it is possible to perform linearization of ρ with first-order approximation, the computation is still intensive. So instead, we replace the correlation function ρ with the sum of squared differences (SSD) as the matching cost as in Lucas-Kanade's algorithm [157]. Now for the disparity refinement purpose, the matching cost is redefined as

$$C_{\text{SSD}}(x, y, d) = \sum_{(u,v)\in W(x,y)} \left[I_r(u + d, v) - a(I_l(u, v) + b)\right]^2, \tag{7.2}$$

where a and b are the gain and bias factors, respectively. Here we assume the disparity is constant within the matching window W. But this assumption is generally not true except for frontal-parallel surfaces. To allow the disparity to vary within the window, we first warp the right image based on the current disparity map,

$$\hat{I}_r(x, y) = I_r\left[x + d(x, y), y\right]. \tag{7.3}$$

To estimate Δd, a and b, we define an error function with $\hat{I}_r$ based on the SSD,

$$\text{Err}^2(\Delta d, a, b; x, y) = \sum_{(u,v)\in W(x,y)} \left[\hat{I}_r(u + \Delta d, v) - (aI_l(u, v) + b)\right]^2. \tag{7.4}$$

With a first-order approximation, we get

$$\text{Err}^2(\Delta d, a, b; x, y) = \sum_{(u,v)\in W(x,y)} \left[\hat{I}_r(u, v) + \hat{I}_{rx}(u, v)\Delta d - (aI_l(u, v) + b)\right]^2, \tag{7.5}$$

where $\hat{I}_{rx} = \frac{\partial \hat{I}_r}{\partial x}$ is the intensity gradient of the warped right image.

Let $\mathbf{p} = \begin{bmatrix}\Delta d & a & b\end{bmatrix}^T$, $\mathbf{a} = \begin{bmatrix}I_{rx} & -I_l & -1\end{bmatrix}^T$, then a concise form of (7.5) can be written as

$$\text{Err}^2(\mathbf{p}) = \sum \left(\mathbf{a}^T\mathbf{p} + I_r\right)^2. \tag{7.6}$$

This is a classic least squares problem. To minimize $\text{Err}^2(\mathbf{p})$ is equivalent to solve the normal equations,

$$\mathbf{A}\mathbf{p} = \mathbf{b}, \tag{7.7}$$

where $\mathbf{A} = \sum \mathbf{a}^T \mathbf{a}$, and $\mathbf{b} = -\sum I_r \mathbf{a}$.

7.1.2 Global Refinement

The previous sub-section describes the method to estimate Δd at each pixel, here we show how to update the disparity map at a global level. The global refinement minimizes a global energy function which takes the same form of (6.1), and is defined by

$$E(d) = \iint \left[d(x,y) - \tilde{d}(x,y)\right]^2 \mathrm{d}x\,\mathrm{d}y + \lambda \times \iint \left(d_x^2 + d_y^2\right) \mathrm{d}x\,\mathrm{d}y, \tag{7.8}$$

where $\tilde{d}$ is the local estimate of the disparity, and d_x, d_y are the disparity gradients. The first term in (7.8) measures the coherence with the local estimation, and the second term imposes smoothness constraint on the solution. λ is called the regularization parameter that weights the smoothness term.

For the n-th iteration, we set $\tilde{d}^{(n)} = \tilde{d}^{(n-1)} + \Delta d^{(n)}$. Then the discrete form of (7.8) can be expressed as

$$\begin{aligned} E(D) = \sum_{(i,j)\in I} \Bigg\{ & \left[d^{(n)}(i,j) - \left(d^{(n-1)}(i,j) + \Delta d^{(n)}(i,j)\right)\right]^2 + \\ & \lambda \times \left[\left(d^{(n)}(i+1,j) - d^{(n)}(i,j)\right)^2 + \left(d^{(n)}(i,j+1) - d^{(n)}(i,j)\right)^2\right] \Bigg\}, \end{aligned} \tag{7.9}$$

where (i,j) is the discrete coordinates of a pixel in the image plane I, and the discrete gradients are computed using the forward difference. Minimizing the energy function yields

$$(1 + \lambda k_{\mathbf{p}}) \times d_{\mathbf{p}}^{(n)} - \lambda \times \sum_{\mathbf{q}\in N(\mathbf{p})} d_{\mathbf{q}}^{(n)} = d_{\mathbf{p}}^{(n+1)} + \Delta d_{\mathbf{p}}^{(n)} \tag{7.10}$$

for each pixel $\mathbf{p}$ whose number of neighboring pixels is $k_{\mathbf{p}} = |N(\mathbf{p})|$. Then we can establish a linear system

$$\mathbf{Pd} = \mathbf{h} \tag{7.11}$$

where the main diagonal of $\mathbf{P}$ is $1+\lambda k_{\mathbf{p}}$, i.e., $[\mathbf{P}]_{\mathbf{p},\mathbf{p}} = 1+\lambda k_{\mathbf{p}}$, $[\mathbf{P}]_{\mathbf{p},\mathbf{q}(\mathbf{p}\neq\mathbf{q})} = \begin{cases} -\lambda, & \mathbf{q} \in N(\mathbf{p}) \\ 0, & \text{otherwise} \end{cases}$, $[\mathbf{d}]_{\mathbf{p}} = d_{\mathbf{p}}^{(n)}$, and $[\mathbf{h}]_{\mathbf{p}} = d_{\mathbf{p}}^{(n-1)} + \Delta d_{\mathbf{p}}^{(n)}$. Since $\mathbf{P}$ is a sparse, positive, and symmetric matrix, the solution can be searched efficiently using the conjugate gradient method [158].

7.1.3 Geometric Detail Enhancement

The smoothness term that is regulated by a coefficient λ from (7.8) apply constraints to prevent rapid disparity change within a small neighborhood, which is usually the result of noise or mismatch. However, the side effect of applying a smoothness term in the global energy function is that it smooths out the surface where fine geometric variation exists. To recover these fine surface details, a bilateral filter is applied to the refined disparity map. Due to the computational cost of the filtering, it is only applied as a post-processing after the sub-pixel refinement.

The concept of bilateral filtering for geometric enhancement is similar to the one that is used for background suppression in computing the cost of color difference for matching (Section 5.2.4). Instead of working on an image with pixel data stored in RGB channels, the input of the filtering is the disparity map with single channel pixel data. The support weight $w(\mathbf{p},\mathbf{k})$ of a kernel pixel $\mathbf{p}$'s neighbor $\mathbf{k}$ has the same form as the weight that is defined in (5.13), but takes in the sub-pixel disparity difference and the spatial distance between $\mathbf{p}$ and $\mathbf{k}$, respectively. The operation of bilateral filtering on the disparity map is defined by

$$BF[I(\mathbf{p})] = \frac{\sum_{\mathbf{k}\in\Omega_{\mathbf{p}}} [w(\mathbf{p},\mathbf{k})d(\mathbf{p})]}{\sum_{\mathbf{k}\in\Omega_{\mathbf{p}}} w(\mathbf{p},\mathbf{k})}, \tag{7.12}$$

where $\Omega_{\mathbf{p}}$ denotes the set of all pixels in the support region and the normalization factor $\sum_{\mathbf{k}\in\Omega_{\mathbf{p}}} w(\mathbf{p},\mathbf{k})$ ensures the support weights sum to one.

7.1.4 Point Cloud Generation

Once a refined disparity map is computed, the 3D surface point cloud in the world coordinate system can be computed in two steps. First, each pixel in the disparity map is back-projected into a 3D point in the camera's coordinate system. It is then transformed into the world coordinate system. For a pixel (u, v) in a disparity map with its disparity value being d, the 2D to 3D back-projection $(u, v, d) \rightarrow \begin{bmatrix} x & y & z \end{bmatrix}^T$ is defined as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = -\frac{b}{d} \times \left(\begin{bmatrix} u \\ v \\ f \end{bmatrix} - \begin{bmatrix} u_0 \\ v_0 \\ 0 \end{bmatrix} \right), \tag{7.13}$$

where b is the baseline between two calibrated stereo cameras, (u_0, v_0) is the camera center, and $\begin{bmatrix} x & y & z \end{bmatrix}^T$ is the back-projected 3D point in camera's coordinate system. Its coordinate in the world coordinate system $\begin{bmatrix} X & Y & Z \end{bmatrix}^T$ is then computed as

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{R}^* \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \mathbf{t}^*, \tag{7.14}$$

in which $\mathbf{R}^*$ and $\mathbf{t}^*$ are the relative rotation and translation of the reference camera (in our case, the left camera) with respect to the world coordinate system. They were calculated from our global registration procedure (Section 4.2.2) for each stereo unit.

7.1.5 Refinement Results

Table 7.1 lists the parameters that are used in our sub-pixel disparity refinement. The refinement was done in an iterative fashion, a total of 15 iterations are performed for each disparity map for a balanced computational cost and surface smoothness. We plot the total disparity value changes at each iteration, and the graph is shown in Figure 7.1. The convergence rate is close to exponential and the disparity update does not change noticeably after 10 iterations. We thus stop the refinement at iteration 15. Compared to the bilateral filtering used for background suppression in matching cost computation (Section

Table 7.1: Parameters for our sub-pixel disparity refinement and geometric enhancement.

Parameters	**Values**	**Descriptions**
N_{Iter}	15	Number of iterations for sub-pixel disparity refinement
W_{SSD}	11×11	Window size of SSD in disparity refinement
λ	10.0	Regularization parameter in disparity refinement
σ_d	0.0784	Variance for disparity difference in bilateral filtering
σ_s	6	Variance for spatial distance in bilateral filtering
W_{BiFil}	21×21	Window size of bilateral filtering

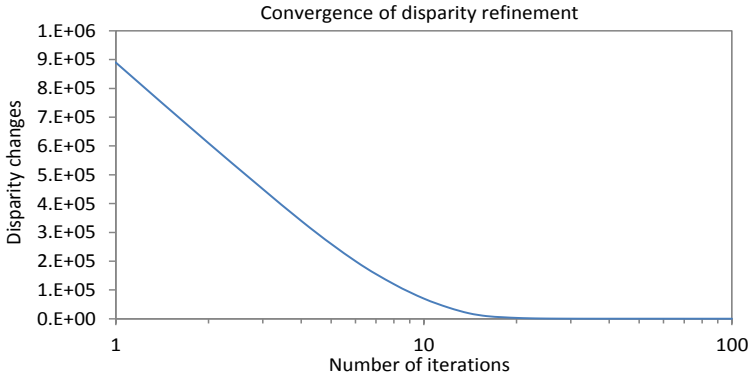
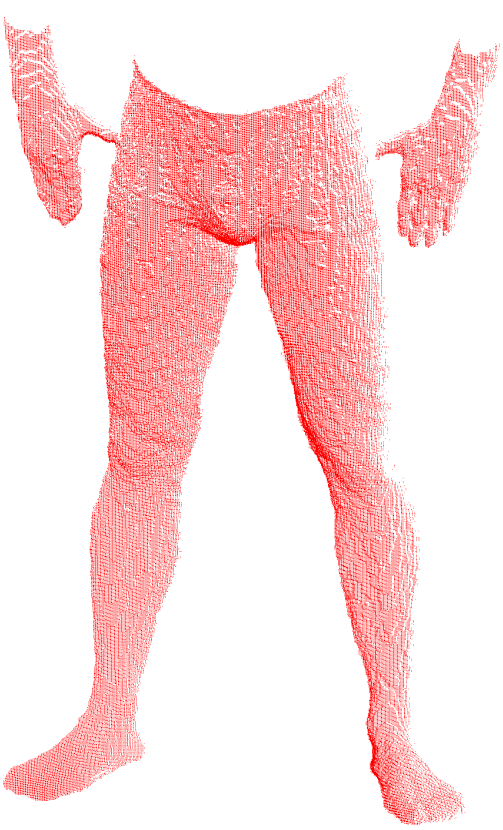


Figure 7.1: Convergence of the sub-pixel disparity refinement over the first 100 iteration. The initial convergence is close to exponential and the update of disparity does not change noticeably after 10 iterations.

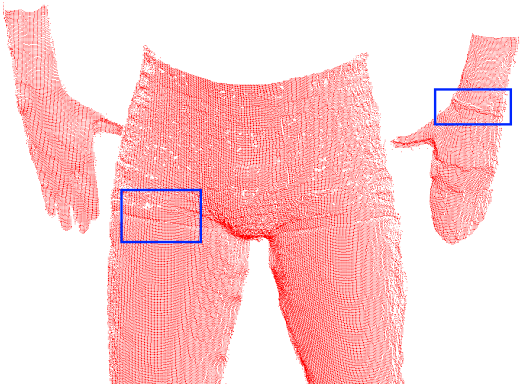
5.2.4), the size of filtering window (W_{BiFil}) and the variance of spatial distance σ_s are kept the same, however, a significantly smaller variance for the disparity difference (σ_d) is used here. This is due to the fact that on a smooth surface, disparity values within a neighborhood window are not expected to change by a large step. And for most of the pixels, the update to the disparity value after each iteration are at the scale of $\frac{1}{100}$.

Figure 7.2 shows an example of the reconstructed 3D point clouds before and after the sub-pixel refinement. The depth steps are very visible (Figure 7.2a) in the surface point cloud that is not refined. Figure 7.2b shows a refined surface though our global energy minimization approach. Depth steps get smoothed out, however, we also lose fine surface



(a) Surface point cloud reconstructed from discrete disparity map. Surface depth steps are visible.

(b) Surface reconstructed from sub-pixel refined disparity map. Depth steps are smoothed out.



(c) Surface detail enhanced by bilateral filtering. Fine details (high-lighted in the figure) are recovered.

Figure 7.2: The results of sub-pixel refinement.

details, such as the wrinkles and edges on the subject's underpants, and the wrist string on subject's left hand. Our bilateral filtering works as a post processing after the iterative refinement procedure, and is able to recover these fine surface details as is demonstrated in Figure 7.2c.

7.2 Surface Reconstruction

The raw surface point cloud data computed from stereo matching and depth conversion are usually comprised of hundred thousands of scattered 3D points, from which it is hard to read and to extract desired information directly. A body modeling process is

required to accurately fits the surface point data with a more manageable representation so that the data can be manipulated and interpreted more easily. In general, such a representation is in the form of 3D surface, the proess is also called body surface reconstruction.

We used the same software that was developed in [14] for surface reconstruction. It utilizes sub-division surface reconstruction algorithm. The basic idea of the method can be described in three steps. First, the original 3D data points is re-sampled on a predefined regular grid. The explicit neighborhood information of the re-sampled data is then used to create an initial dense mesh. Secondly, the initial dense mesh is simplified to produce an estimate of the control mesh. Finally, the control mesh is optimized by fitting its sub-division surface to the original data, and accordingly, the body model is reconstructed. In the surface reconstruction processing, the upper and lower mesh from the same side of the subject is blended together at the overlapped region between the waist and hip lines to smooth out the transition between the upper and lower surfaces captured by two different stereo units. The gaps along the side of the body model that are occluded to the stereo units are closed by stitching the edges of the surface point cloud. The final results of the surface reconstruction is a closed surface mesh comprised of triangles that approximate the original point cloud. The number of vertices on the surface mesh is greatly reduced from the original point cloud, and it represent a simplified form efficient feature characterization.

Figure 7.3 shows a collection of reconstructed body models for circumference and volume measurements. It can be observed that surfaces on the reconstructed body models are smooth due to the re-sampling and sub-division mesh simplification. A negative effect of the smoothed surface is that we may lose some fine geometrical details that are enhanced by our sub-pixel refinement step. However, this should not affect body volume measurement because the simplified surface mesh approximates the original surface

points in a least mean square error fashion.

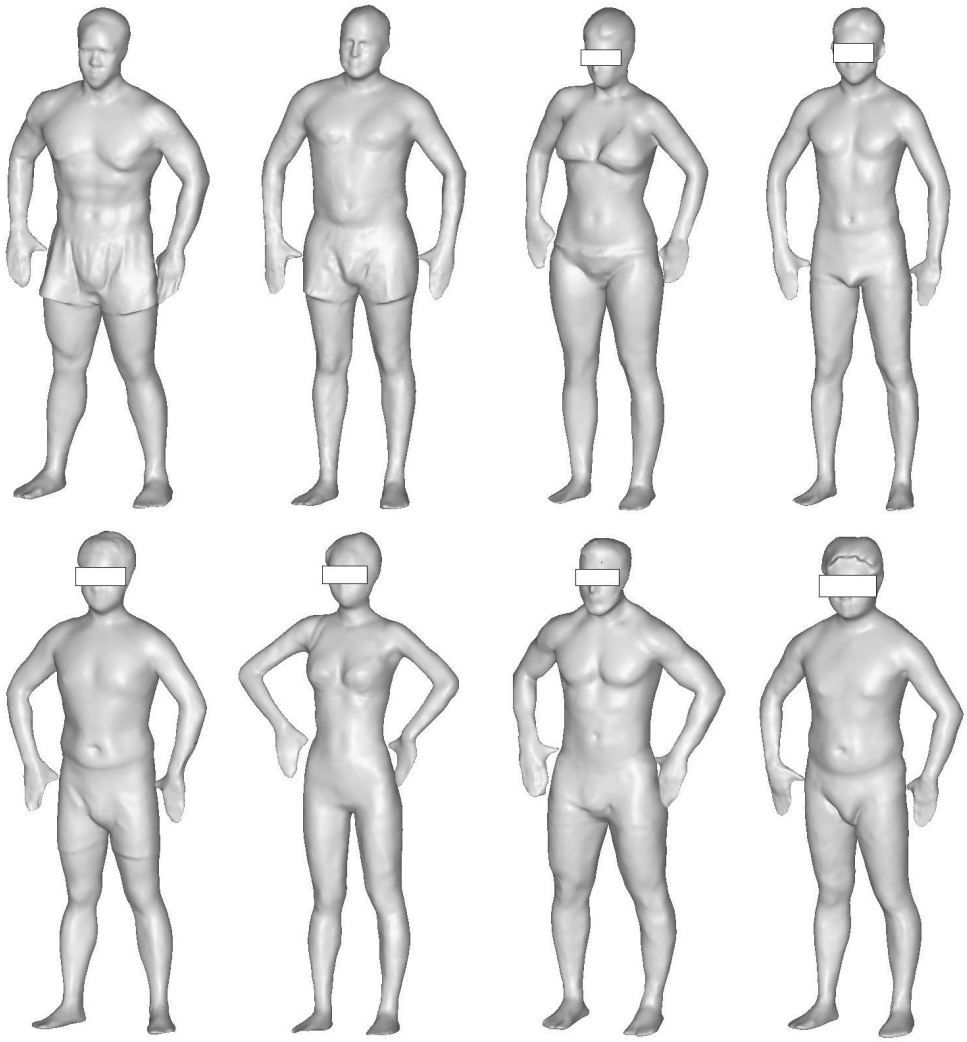


Figure 7.3: Reconstructed body models of subjects with various body shapes and sizes.

Chapter 8

Body Measurement and System Evaluation

This chapter is presented in two parts. In the first part, we describe how to perform body dimension measurement and percent body fat estimation on the reconstructed 3D body models. In the second part, we present methods and results on the evaluation of the developed stereo body imaging system. Measurement results were compared to physical tape measures, commercialized portable 3D scanner, and DEXA. The system was tested on mannequins and human subjects to evaluate its accuracy and repeatability.

8.1 Measurement Principles

8.1.1 Body Measurement on 3D Model

With the fast growing demand for full body imaging and the maturing of 3D capture technology, the segmentation and measurement on imaged 3D data have also received great attention form the research community. Early attempts for 3D measurements are mainly based on sliced scan data. A model-based approach was proposed by Dekker *et al.* [159] to aggregate sliced data into sectors, of which the centroid can be analyzed to automatically detect body surface landmarks. Body volume can be measured by integrating over the slices. Ju *et al.* [160] proposed a method in which the body is first segmented into head, torso, arms, and legs according to slice settings, and then the girth profiles of individual body parts are used to locate the neck, shoulders, waist, elbows, wrists, knees and ankles. Sliced data are usually captured by laser-based scanner, and a 3D surface is acquired by accumulating surface profiles as the laser bean scans across the body. Later on,

as other 3D surface acquisition methods became available and be more popular in terms of scanning speed, body measurement techniques based on point cloud and triangular surface mesh started to gain momentum. Xiao *et al.* [161] used geodesic distance to segment the body into primary parts. The advantage of this method is that geodesic distance is independent of body postures. Leong *et al.* [162] proposed an algorithm in which the torso data is transformed to cylindrical coordinates and then converted into a 2D depth map so that the problem is transformed into 2D scope in which image processing techniques can be used to extract features. A great review of segmentation and modeling of human body has been presented in [163].

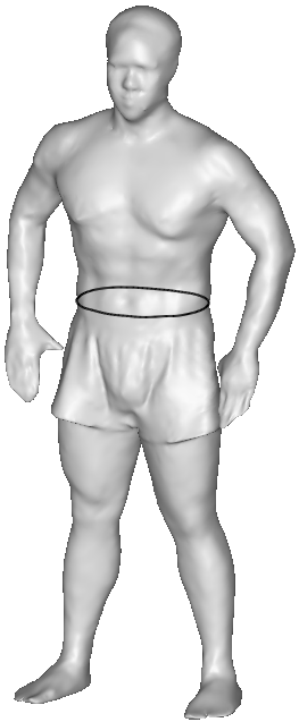
In the previous work of our research group, Zhong and Xu [164] reported a body segmentation and measurement system that works on triangular meshes with the goal for application in virtual apparel fitting. In this method, geometric landmarks are searched in their target zones that are predetermined based on the proportion relative to the stature. The armpits and neck are searched with the criterion of minimum inclination angle between neighboring triangles. The crotch is identified by detecting the transition of cusps along successive horizontal contours. Once key landmarks are located, the body is then segmented into head, torso, arms and legs. With segmented body parts, various measurements including circumferences and lengths can be extracted. Even with these useful functions, this system is not sufficient for body composition research. For example, functions for body volume measurement is very limited and the computation cost is very high. To accurately estimate whole body volume, it needs to section the body parts into dense slices and divide each slice into dense line segments, so that the body volume can be computed by integrating these slices. The procedure involves extensive computation of plane-to-plane and line-to-line intersection. To reduce the computational cost, this 3D measurement system was extended by Yu [14] to utilize the depth-buffer of the 3D scene available from

a computer graphics API to accelerate surface circumference and volume computation. Graphics API allows a higher level application to access 3D geometric functions that are implemented on hardware compute unit, thus is optimized for performance. We follow this practice for fast and efficient body geometric parameter calculation. The basic concept of using computer graphics API to accelerate the computation is briefly introduced in Section 8.1.2.

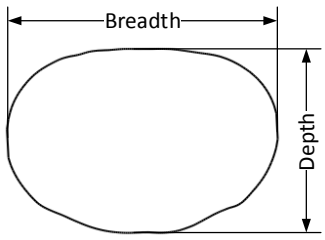
8.1.2 3D Measurements

8.1.2.1 Volume Measurement

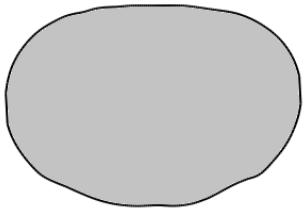
In computer graphics, the depth buffer (also called z-buffer) is a 2D map which records a depth value (distance to the viewport) for each rendered pixel of a 2D frame. With 3D graphics APIs such as OpenGL, we can switch the z-buffer to keep track of the minimum or maximum depth for each pixel on the frame. To measure the body volume, the 3D body model is rendered twice, one is for the front surface, and the other is for the back surface. During the two renderings, we choose the z-buffer to record the *minimum* and the *maximum* depth of each pixel, respectively. By taking the difference of the two z-buffer, we get a thickness map of the body. Finally, the body volume is calculated by integrating over the thickness map based on the known pixel's physical scale. It should be notices that, when generating the z-buffer for each rendering, orthographic projection is applied to reflect the actual size of the body. In principle, the z-buffer rendering method is equivalent to re-sample the surface data on a regular grid, thus the size of the viewport (in pixels) that determines the sampling interval may affect the measure accuracy. A moderate size of the viewport such as 500×500 yields a physical resolution of 3.6 mm/pixel. In practice, we found this is sufficient to reach high accuracy. It is reported in [14] that it takes about 50 ms to render a typical model presented in a triangular mesh with 15,000 vertices. Thus, this technique is extremely efficient time-wise compared to slice-based methods. For regional



(a) The circumference is marked on a 3D body model.



(b) Circumference, breadth and depth measurement.



(c) Cross-sectional area measurement.

Figure 8.1: Measurements extracted from a contour on a 3D body model.

volume measurement, instead of projecting the whole body on the z-buffer, we only need to project the individual segment and employ the same z-buffer difference method.

8.1.2.2 Circumference Measurement

Circumference measurement is especially convenient to evaluate the measurement accuracy with respect to manual measurement with a tape. With this function, a user has the freedom to take circumference measurement at any region on the body by marking a contour with a line-drawing tool. With the z-buffers of the front surface and back surface being readily available, the 3D data for the contour can be obtained instantaneously with sub-pixel accuracy. Then, the circumference as well as the breadth and depth of the contour can be calculated. It is worthy noting that the location of the front part of the contour does

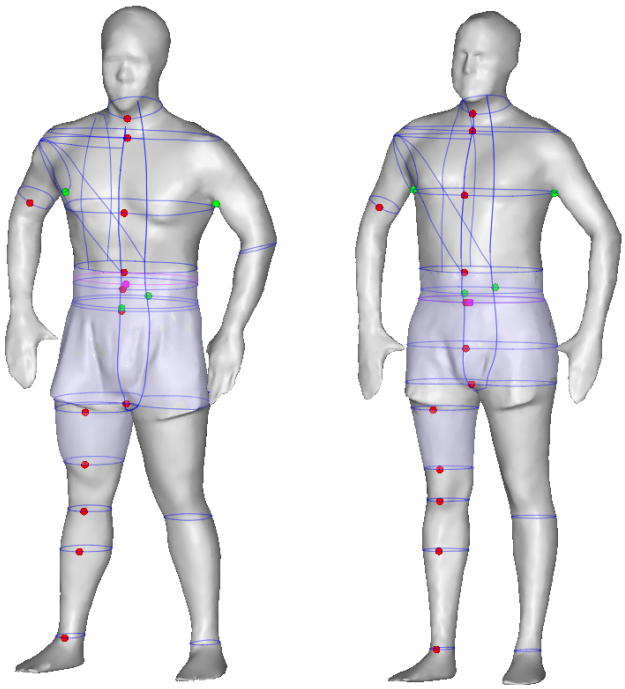
not need to match the location of the back contour. The contour can be slanted or tilted in any way to fit the location where needs to be measured. An example is demonstrated in Figure 8.1. A contour is marked on the body model as shown in Figure 8.1a, and then its circumference, breadth and depth are calculated as shown in Figure 8.1b.

8.1.2.3 Area Measurement

Once a contour is extracted in the circumference measurement, its cross-sectional area can be estimated by re-facing the contour with the normal of the cross-sectional plane coincident with the normal of the screen plane. Then the cross-section is projected onto the depth buffer, and the image of the projection is read for area calculation. The area of the cross-section is estimated by counting the pixels inside the contour. An example is shown in Figure 8.1c, where the shaded pixels are counted to get the cross-sectional area.

In addition to cross-sectional area, body surface area can also be estimated by summing up the areas of all triangles in the surface mesh. In our current system this computation is not graphics hardware accelerated, because the surface area computation does not involve user interaction and it is only computed once for each 3D body model. However, graphics acceleration can be made possible for modern graphics hardware that supports geometry shader. With geometric shader, the computation of the triangle area of the surface mesh can be programmed to be executed on the shader compute unit which is massively parallel.

To illustrate the output of the body measurement system, results of two subjects are shown in Figure 8.2. The measured body parameters include circumferences and cross-sectional areas of a number of body components (such as the chest, waist, abdomen, hip, upper thigh, etc), whole body volume, segmental volumes (such as the abdomen-hip volume and the upper thigh volume), and body surface area.



(a) Body measurement with various circumferences and landmarks highlighted.

(b) Body segments used for the abdomen-hip volume measurement.

Figure 8.2: Illustration of body measurement.

8.2 Subjects and Methods

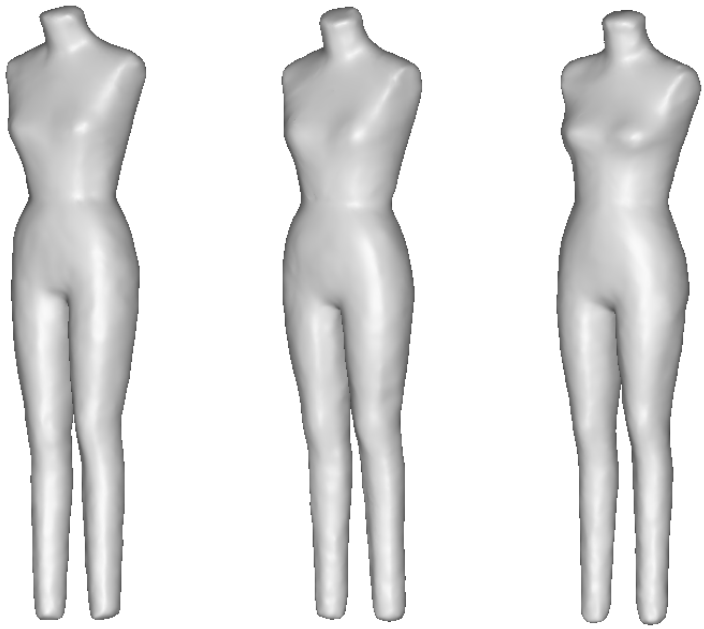
To evaluate the accuracy and repeatability of the developed prototype body imaging system, we have tested it on some mannequins whose body dimension can be measured manually. Tape measurement was conducted to acquire various circumferences from the mannequins. A commercialized handheld 3D scanner was used to scan our mannequins. Volume measurement from our stereo imaging system was compared to the results from the handheld scanner. The system was also tested on human subjects for the measurement of body circumference and volume. To validate its feasibility in body fat assessment, body densities of human subjects were calculated with body weight and volume, and their percent body fat (%BF) were estimated. The results were compared to DEXA.

8.2.1 Mannequins and Measurements

Three standard mannequins (Wolf Form Co., Englewood, NJ) for bridal dress fit were used to evaluate the accuracy and reliability of the system. The manufacturer-defined sizes of these mannequins are 8, 10 and 12. The reconstructed 3D models of these mannequins are shown in Figure 8.3. Models were zoomed at the same scale to show the size differences. A MyoTape body tape measure (AccuFitness LLC, Greenwood Village, CO) was used for circumferences measurement. Each mannequin was imaged five times with repositioning for each trial. Chest, waist and hip circumferences, and total body volume were measured on the 3D model automatically. The coefficient of variance (CV) was computed to estimate repeatability. To evaluate the accuracy on circumference and volume measurement, the results were compared to those obtained with physical tape measure and handheld scanner.

To estimate the longitudinal day-to-day repeatability of the system, the size-12 mannequin was imaged in five trails with no more than one trail was conducted on a single day. For each trial, the measurement were repeated three times.

A portable 3D scanner, Go!SCAN 3D (Creaform Inc., Quebec, Canada), was used to acquire the 3D models of our mannequins. These 3D body models were used as references for volume measurement. Go!SCAN 3D is an active lighting device. Its imaging principle is based on triangulating patterns projected onto the scanned surface. It has a working depth range of about 20–50 cm. The 3D reconstruction of a scan is performed at real-time. To initiate a 3D scan, the scanner has to be held steadily to image the same surface for a few seconds, so that sufficient frames can be captured to constructed an initial set of surface feature points. These feature points are used for new surface alignment. Once the initial surface is acquired and is stabilized, a user can move the scanner over the target to scan the complete surface. Whenever the scanner is moved to a new position to cover new



(a) Size 8. (b) Size 10. (c) Size 12.

Figure 8.3: Mannequins of three different sizes were used to verify the accuracy our developed body imaging system.

surfaces that have not been imaged before, it extracts feature points from the current frame, and matches these feature points to the collection of all feature points that it maintains. Parameters that are needed to align newly acquired 3D surfaces is calculated from feature matching.

8.2.2 Human Subjects and Measurements

Twenty adult subject (twelve males and eight females) were recruited in this study to help evaluate our developed body imaging and measurement system. The subjects were aged 24–41 years, with weights 38.6–101.0 kg, heights 153.7–182.4 cm, and BMI 16.33–30.37 km/m^2.

The study was approved by the Internal Review Board of the University of Texas at Austin. The data collection procedure together with the informed consent form were sent

to each subject before their visit. Signatures from the subjects on the consent form were collected when they were on site. The subjects were instructed to fast at least three hours, stay hydrated, and avoid excessive sweating, heavy exercise, and caffeine or alcohol use before all procedures were performed.

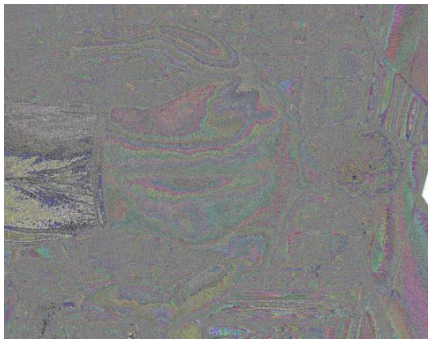
Subjects were asked to wear tight-fit underwear for body imaging. First, height, weight, chest circumference, waist circumference and hip circumference were measured with conventional anthropometric methods using the same tape that was used as in mannequin measurement. Then, the subjects were imaged by stereo cameras at maximum exhalation after normal breathing. Subject's body volume was corrected by subtracting the lung residual volume from the raw body volume. The residual volume was estimated by prediction equations that are functions of height and age [165],

$$\begin{cases} RV^{(\text{Men})} = 0.0216H + 0.0207A - 2.840 \\ RV^{(\text{Women})} = 0.0197H + 0.0201A - 2.421 \end{cases} \tag{8.1}$$

in which H is the height in cm, and A is the age in years. The estimated volume is in L.

Image capture took place in a burst mode, in which the computer software launched a batch of threads, and each of the thread was dedicated to send a shutter release command to their assigned camera and to wait for the image to be transferred in. Even though the picture taking was made to be multi-threaded and all threads were executed at the same time, the total time needed before all images were transferred back into computer may add up to two seconds, due to various software/hardware delays and the duration of exposure. It is crucial for a subject to remain steady, otherwise subject's motion will cause mis-alignment of captured 3D surfaces and eventually lead to measurement error. In order to minimize involuntary movements, the subjects were asked to stand still in a specific posture with the legs slightly spread and the arms abducted from the torso. Subjects were also asked to have their hands touch their hip with their thumbs and open their palms flat.

(a) The original image.



(b) The image with scrambled pixels.

Figure 8.4: To help protect privacy of our subjects, stereo pictures were scrambled to hide image contents before they were saved to our computer.

The imaging was repeated 10 times for each subject. The subjects were asked to relax for a few seconds and were repositioned between scans.

Since the body imaging system captures pictures of the volunteers and pictures are personally identifiable information, their privacy needs to be protected to prevent unauthorized use of these pictures. All pictures were encoded with a special procedure so that the color of each pixel is scrambled before they were saved onto our computer. Only the same software that was used for image capture and 3D reconstruction was able to decode these images. An example of the pixel scrambling is shown in Figure 8.4.

The subjects were also assessed for body fat by DEXA (Lunar iDXA, General Electronic, Fairfield, Connecticut). During the DEXA test, subject lied down on an open "table" for approximately eight minutes while the X-ray sensor scaned over their body. DEXA measures the whole body fat, lean, and bone mneral mass and fat percentage. In addition, the DEXA test determines the fat, lean, and bone mineral content and fat percentage for the arms, legs, and trunk. However, only the whole body fat percentage is compared to the stereo imaging system.

8.2.3 Statistical Analysis

The repeatability of the developed body imaging system was determined by computing the coefficient of variance (CV) and the intra-class correlation coefficient (ICC) from the results of one-way random effects ANOVA. Based on the between- and within-group mean errors available from ANOVA, CV is computed as the ratio of within-group standard deviation (SD_w) to the global mean, and is presented in percentage format, or

$$CV = \frac{SD_w}{Mean} \times 100\%. \tag{8.2}$$

The ICC is determined as

$$ICC = \frac{MS_b - MS_w}{MS_b + (n-1) \times MS_w}, \tag{8.3}$$

in which MS_b and MS_w are the between- and within-group mean square errors (MS), respectively. n is the number of samples per group. The comparisons of measurements using tape and stereo imaging were performed with *t*-tests and linear regression analysis.

Percent body fat was calculated from whole body volume measured by stereo imaging using Siri's equation (2.2). The paired-sample *t*-tests and linear regression were applied to compare the %BF estimates between stereo imaging and DEXA. In additional, Bland and Altman analysis was used to assess agreement of %BF between these two. A 95% agreement was estimated by the mean difference $\pm 1.96SD$. For all analyses, statistical significance was $P < 0.05$.

8.3 Results

8.3.1 Evaluation on Mannequins

All circumference measurements were automatically computed based on various landmark locations detected on the reconstructed 3D body models. The volume measurements were calculated through depth buffer integration by calling graphics APIs. The

Table 8.1: Repeatability test on mannequins of three different sizes.

Circumferences/Volume	Mean	MS_w	SD_w	CV (%)
Chest (mm)	902.8	5.2	2.3	0.25
Waist (mm)	672.2	1.9	1.4	0.20
Hip (mm)	941.6	2.0	1.4	0.15
Volume (L)	50.789	0.0022	0.047	0.09

MS_w, within-subject mean square error (MS); SD_w, within-subject standard deviation; CV, coefficient of variance.

results of repeatability test were computed from ANOVA in which the three mannequins were treated as three subject groups, and repeated scans of each mannequin were treated as multiple tests performed within one group. Each mannequin was imaged five times, thus five tests were available for each group. Table 8.1 shows the results from the ANOVA analysis, the within-subject standard deviation (SD_w) and CV. It should be noted that the between-subject mean square errors (MS_b) and the *P*-value reported by ANOVA were ignored for this evaluation, because we have already known that significant differences exist among the three groups.

The CVs for our multi-group analysis are presented as percentage values. The CVs were $\leq$ 0.2% for waist and hip circumferences, and was $<$ 0.1% for volume. A low CV value indicates small variation in measurements. The CV increased to 0.25% for chest circumference due to the rapid variation of the circumference sampled at different vertical locations that were above or below the true chest line. A small change at the vertical location could result in larger circumference change than in the waist and hip lines. This pattern has been observed on both mannequins and female human subjects, because their waist circumferences are significantly shorter than their chest circumferences.

The results of longitudinal repeatability test based on the volume of size-12 man-

Table 8.2: Longitudinal repeatability test on the size-12 mannequin's volume.

Source of Variation	SS	DF	MS	F	P-value	F_{crit}
Between Trials	0.001901	4	0.000475	0.924	**0.49**	3.478
Within Trials	0.005142	10	0.000514			
Total	0.007043	14				

The global mean of the multiple tests over five trials is 52.628 L.
SS, sum of squares; DF, degree of freedom; MS, mean square errors.

nequin are given in Table 8.2. A total of five trials were conducted on five days, and three scans were performed for each trial. The trials were treated as independent groups in ANOVA. Thus, the between-trial degree-of-freedom is 4 ($5 - 1 = 4$), and the within-trial degree-of-freedom is 10 ($5 \times 3 - 5 = 10$). The P-value of the variance analysis was $0.49 > 0.05$, which indicated that there was no significant difference in the body volume measurements over these five days.

Table 8.3 shows the comparison of circumference measurements between stereo vision and manual tape method on size-12 mannequin. The measurement data are presented in $(\text{Mean} \pm \text{SEM})$ format. The SEM (Standard Error of the Mean) is computed by dividing the standard deviation by the square root of the number of samples. The mannequin was imaged five times, and tape measured five times. The comparison results were generated by paired two sample t-tests, with one variable being the stereo measurement and the other being the tape measurement. With all P-values being > 0.05, the measurements between these two methods are not considered to be significantly different. However, since all P-values are still relatively low, noticeable differences could be expected. The "Difference" colume in Table 8.3 shows that the system may generate measurement results that are slightly higher than the tape measure.

Table 8.4 shows the whole body volumes of three mannequins measured by stereo imaging and the handheld scanner. The mannequins were imaged by stereo vision five

Table 8.3: Circumferences of the size-12 mannequin measured by stereo imaging and tape.

Circumferences	Stereo imaging	Tape	Difference	*P*-value
Chest (mm)	925.6 ± 0.9	923.6 ± 0.4	2.0 ± 0.8	**0.09**
Waist (mm)	702.2 ± 0.8	699.1 ± 0.5	3.0 ± 1.1	**0.06**
Hip (mm)	972.3 ± 0.5	965.2 ± 0.5	2.1 ± 1.0	**0.13**

Measurement data are presented in (Mean $\pm$ SEM) format. The mannequin was imaged and tape measured five times. The *P*-values were obtained from paired two sample *t*-tests.

Table 8.4: Whole body volumes of the three mannequins measured by stereo imaging and Go!SCAN.

Mannequins	Stereo imaging	Go!SCAN	Difference
Size 8 (L)	48.324 ± 0.018	48.077	0.247
Size 10 (L)	51.418 ± 0.026	51.138	0.280
Size 12 (L)	52.626 ± 0.009	52.349	0.277

Volumes measured by stereo imaging are presented in (Mean $\pm$ SEM) format. The mannequins were imaged by stereo system five times, but they were scanned by Go!SCAN only one time, due to the inability to close body surface mesh on small body parts, e.g., the end of legs.

times, but they were only scanned by Go!SCAN one time. We experienced a great amount of difficulty in get the complete surfaces of the mannequins from the handheld scanner. The software that the Go!SCAN uses to perform real-time surface fusion had trouble aligning small body parts. This typically happened when imaging the end of the legs. Possible reasons for the inability to align could be the small surface area and the lack of geometrical variation and features to identify a match. A third party software was used to close the scanned body meshes, and some manual editing was needed to fill up the missing regions. As a result, we could expect some error from Go!SCAN's measurements. The differences between these two methods were about 0.3 L for a 50 ± 2 L body. The ratio of the difference with respect to the measurement value was around 0.6% ($0.3/50 \times 100\%$). A *t*-test was not

performed on these two methods because of the limited number of measurements from the Go!SCAN.

8.3.2 Evaluation on Human Subjects

The overall age and physical body dimension measurements of the twenty human subjects are listed in Table 8.5. Out of these twenty subjects, fifteen have thin to regular body build with BMI within the range of 18.5–25 kg/m^2. One subject was professional athlete, and has a BMI value over 30 kg/m^2. Among the rest four subjects, two exercises regularly for muscle build, the other two have relatively more fat than average. Their BMIs are in the 25–29.9 kg/m^2 range. None of the subjects is obese.

Table 8.5: Human subject characteristics.

	Mean	SD	Range
Age (yr)	27.6	3.7	23–41
Height (cm)	168.1	9.6	153.7–182.9
Weight (kg)	64.2	16.0	38.6–101.0
BMI (kg/m^2)	22.3	3.79	16.33–30.37

The initial results of our system evaluation showed that the chest and waist circumferences measured by stereo imaging correlated very well with tape measurements, but the hip circumference and %BF were significantly different from tape measure and DEXA scan. The cause of these differences was a systematic, positive bias found in the reconstructed 3D models. After reviewing each of the 3D body models, we discovered that one of the major cause of over-estimated hip circumference and body volume was the loose-fit clothes the subjects wore during imaging. Although we advised tight-fit clothes, this requirement was hard to enforce. Some of the clothes were not tight enough to expose body shape. Another cause of the over-estimated body volume is the hair volume. An

Figure 8.5: A 3D body model was sculpted to reveal the actual body surface. Modified body surfaces include head (hair volume) and underwear. The unsculpted body model of the same subject can be found in Figure 8.1a.

over-estimated body volume reduces body density, and ultimately increase %BF.

To overcome the positive bias of body volume measurement, we modified the reconstructed 3D body models by manually sculpting the inflated regions to reflect the actual body profile. This procedure was performed in Meshmixer (Autodesk, Inc., San Rafael, CA) with editing tools. Figure 8.5 shows an example of a sculpted body. The unsculpted body model of the same subject can be found in Figure 8.1a. Subject's underwear was flattened to reveal the actual body surface, and the hair volume was also reduced. Our reported hip circumference and body volume measurements were computed from the manually modified body models. Table 8.6 shows parts of the measurements and statistics for all subjects. In the "Subject ID" column, a mark is placed next to a subject's ID if that subject's body models were sculpted to flatten out loose clothes. All body models were edited to remove hair volume. In addition to %BF estimation through Siri's equation, the estimation from Brozek's equation is also listed in Table 8.6. Compared to the DEXA results,

Table 8.6: Measurements and statistics of twenty human subjects.

Subject ID	BMI (kg/m^2)	Raw Volume (L)	Lung Residual (L)	DEXA %BF	Stereo %BF (Siri)	Stereo %BF (Brozek)
1†	18.5	53.756	1.542	11.9	8.5	9.1
2	21.6	48.781	1.107	30.6	30.1	29.1
3	20.7	51.444	1.409	20.2	22.0	21.6
4†	26.4	85.300	1.541	15.4	17.2	17.1
5	15.5	38.263	1.079	24.5	27.1	26.3
6	18.1	47.013	1.266	20.6	22.9	22.4
7	20.2	49.530	1.150	32.4	31.4	30.2
8	23.1	62.206	1.267	22.6	25.1	24.4
9†	31.0	96.561	1.470	15.3	16.0	16.0
10†	30.5	90.010	1.336	20.8	23.1	22.6
11	24.2	63.582	1.208	23.5	21.5	21.1
12	27.0	84.206	1.526	26.0	24.5	23.9
13	20.8	56.244	1.187	17.9	15.8	15.9
14	24.5	66.314	1.203	22.3	21.1	20.7
15	19.4	58.842	1.462	10.6	8.0	8.7
16†	21.1	59.637	1.317	16.3	15.4	15.5
17	22.2	67.090	1.431	20.8	17.2	17.1
18	21.8	52.192	1.152	31.4	28.8	27.9
19	23.4	55.083	1.077	33.9	31.4	30.3
20	26.3	62.123	1.072	40.6	37.6	36.0

All body models were edited to remove hair volume.
† Body sculpting was performed on subject's 3D body models to flatten out loose clothes.

the Brozek's equation has a higher chance to underestimate a person's %BF than the Siri's equation. However, for individuals with thin body build, the Brozek's equation actually performs more consistently with DEXA.

The repeatability of all circumference measurements and body volume measurements is shown in Table 8.7. All ICCs were > 0.99, and all CVs were $< 1.0\%$. The highest

Table 8.7: Repeatability test on 20 human subjects.

Circumference/Volume	Mean	MS_w	MS_b	SD_w	CV	ICC
Chest (mm)	914.3	31.0	27255.6	5.6	0.61	0.9966
Waist (mm)	767.2	12.6	26409.2	3.5	0.46	0.9986
Hip (mm)	956.6	20.9	18916.2	4.5	0.48	0.9967
Raw Volume (L)	62.408	0.047	718.290	0.218	0.35	0.9998

MS_w, within-subject mean square error (MSE); MS_b, between-subject MSE; SD_w, within-subject SD; CV, coefficient of variance; ICC, intra-class correlation coefficient.

precision was reached in body volume with the lowest CV value. This is mainly because there was no ambiguity to calculate whole body volume from a 3D model. However, it was difficult to locate precisely the chest, waist and hip lines. Compared to the repeatability tests of circumferences on mannequins, the CVs of human subject measurements were relatively higher. This is a sign that a higher variation exists in the measurements from multiple scans of the same human subject. There are several causes to this variation. Firstly, determining the location where a circumference should be taken is even more difficult than on mannequins. Secondly, the amount of air in a person's lung can be different at each time the person was imaged. This affects chest circumference measurement most (notice the CV for chest circumference is the highest in Table 8.7), and it may also affect waist measurement.

The measurement accuracy of stereo imaging with reference to tape for circumference measurements is shown in Table 8.8. The *P*-value is computed by paired sample *t*-tests. The *P*-value for chest circumference measurement is > 0.05, and it indicates there is no significant differences between stereo imaging and tape measure. However, the *P*-values for waist and hip circumference measurements are < 0.05.

The degree of agreement on circumference measurements were characterized by

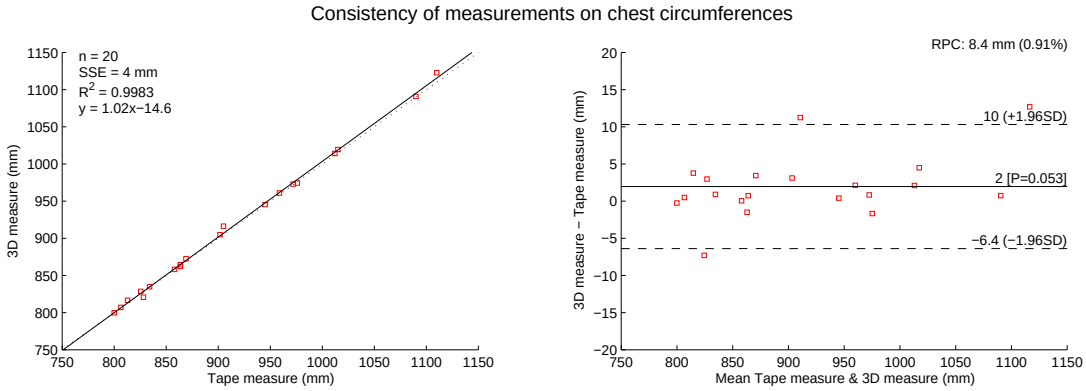
Table 8.8: Comparison of circumferences measured by stereo imaging and tape on human subjects.

Circumferences	Stereo imaging	Tape	Difference	P
Chest (mm)	914.3 ± 20.8	912.3 ± 20.4	2.0 ± 0.9	**0.053**
Waist (mm)	767.2 ± 20.4	764.6 ± 20.4	2.6 ± 0.9	0.012
Hip (mm)	934.4 ± 14.2	930.4 ± 13.8	4.0 ± 1.5	0.017

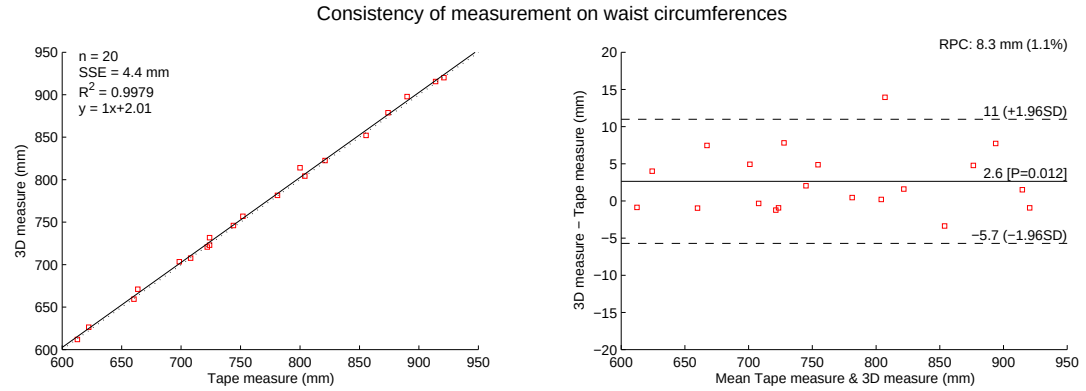
Measurement data are presented in (Mean $\pm$ SEM) format. The P-values were from paired-sample t-tests.

linear regression analysis, and the results are shown in Figure 8.6. The graphs of fitted linear equations are shown on the left and the Bland-Altman plots are on the right. A very high correlation was observed between stereo imaging and tape measure in chest and waist circumferences with $R^2 > 0.99$ and sum of squared errors (SSEs) being less than 5 mm. The correlation of hip measurement was slightly lower than chest and waist with $R^2 = 0.989$ and SSEs being 6.9 mm.

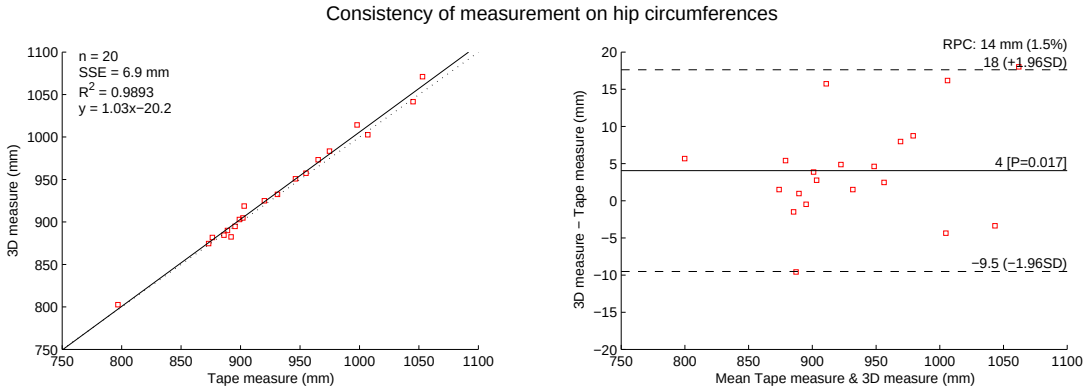
Siri's equation was used to predict %BF from stereo measurement. The average body density was computed by dividing the body weight by total body volume. The estimated %BF was compared to DEXA, and the prediction equation was obtained from linear regression with DEXA as the reference method. The results are shown in Figure 8.7. The predicted equation from linear regression was $y = 0.95x + 0.508$ with SSE $= 2.2$ %BF and $R^2 = 0.9231$. The bias and SD of difference is shown in Table 8.9. Paired sample t-test was performed to discover the difference between these two methods. With the P-value being 0.2, we infer that the %BF estimation through manually corrected 3D models is not significantly different from DEXA.



(a)



(b)



(c)

Figure 8.6: Agreement of tests of measurements on chest, waist, and hip circumferences. Left column: linear regression of the measurements between stereo imaging and tape measure. Right column: Bland-Altman plots of measurement agreement. n: sample size (20); SSE: sum of squared error; R^2: Pearson R-value squared; equation: slope and intercept equation; RPC(%): reproducibility coefficient (1.96 $\times$ SD) and % of mean value.

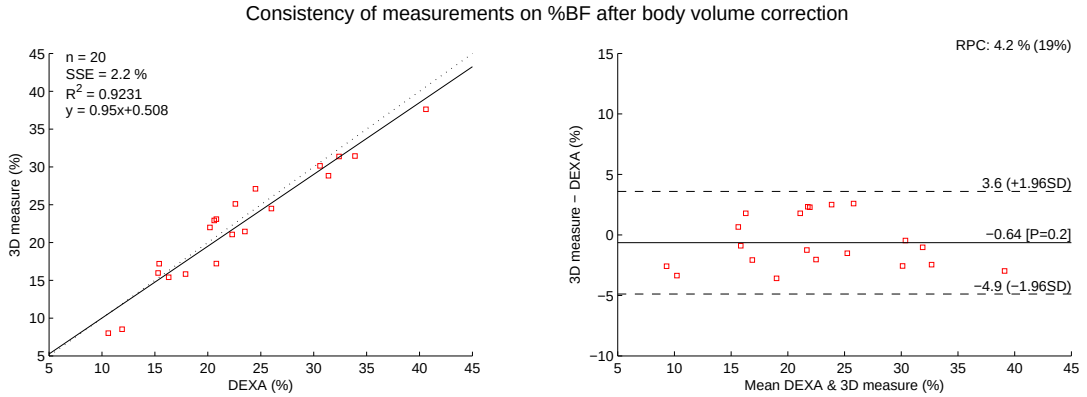


Figure 8.7: Agreement of tests of %BF after body volume correction. Left: linear regression of the measurements between stereo imaging and DEXA. Right: Bland-Altman plot of measurement agreement. n: sample size (20); SSE: sum of squared error; R^2: Pearson R-value squared; equation: slope and intercept equation; RPC(%): reproducibility coefficient ($1.96 \times$ SD) and % of mean value.

Table 8.9: Bland-Altman analysis on percent body fat through corrected body volumes.

	Bias	SD	Reproducibility coef.	P
Stereo imaging — DEXA	-0.64	2.15	4.20	**0.20**

Reproducibility coefficient is defined as $1.96 \times$ SD. The P-value was from paired sample t-test.

8.4 Discussion

8.4.1 Analysis of Results

To evaluate the measurement consistency of circumference and %BF among stereo imaging, tape measure and DEXA, we first looked at the correlation of the measurements. The circumference measurements between stereo imaging and tape showed high correlation ($R^2 > 0.99$). However, good correlation alone is not sufficient to imply that two methods generate similar data. Thus we applied Bland-Altman method to measure the agreement between any pair of comparison. A Bland-Altman plot is primarily used to compare two clinical measurements that each provides some errors in their measure. The mean difference of measurements is the estimated mean bias, and the SD of the differences

measures the random fluctuations around this mean. It is common to look at the 95% limits of agreement for each comparison (bias $\pm$ 1.96 $\times$ SD of difference), which tell us how far apart measurements by 2 methods were more likely to be for most individuals. It is also commonly accepted in clinical measurements that if the difference within the limits of agreement are not clinically important, the two methods may be used interchangeably.

For system evaluation on human subject, circumference measurements from stereo imaging were compared to tape measures, and %BF estimations from stereo imaging were compared to DEXA. We observed that circumference measurements received a higher degree of agreement than the %BF estimation. The biases in circumference measurements were within 2–4 mm range (Table 8.8). Compared to the average circumference measurements of our human subjects on chest, waist and hip, the biases are only 0.2%, 0.3% and 0.4%, respectively. The limits of agreement in all three circumference measurements were compact, with the chest measurement being 8.4 mm, waist being 8.3 mm, and hip being 14 mm (Figure 8.6). As we experienced that 5–10 mm measurement differences were common in tape measurements for circumference on our human subjects, we would accept that the measurement method on circumferences using stereo imaging and tape are similar.

8.4.2 Sources of Errors

It should also be noticed that the biases on circumference measurements were all positive values. This is actually inline with the measurement principles which are slightly different on 3D body models and on human subjects. On a 3D body model, the measurement is taken by fitting a curve that is precisely on the surface of the 3D mesh, and it traces the exact geometrical changes at the measurement site. While on the human subject with tape measure, the tape usually cannot follow surface areas that are concave and it only fits to the "envelope" of these areas. Thus, tape measurement could be shorter than the fitted curve on the 3D body model, if the circumference does not exhibit monotonic curvature

on its loop.

The measurement agreement on %BF between stereo imaging and DEXA was relatively low. Even though the bias was only 0.42% of body fat, the limits of agreement span a wide range. The reproducibility coefficient ($1.96 \times$ SD) was 4.2% of body fat, and it accounted for 18% of mean value. This low measurement agreement on %BF lies in the nature of Siri's equation in estimating %BF from a two-component body composition model. The %BF is very sensitive to the accuracy of body volume measurement. For example, Siri's equation yields

$$\Delta(\%\text{BF}) = \frac{495 \times \Delta V}{W}, \tag{8.4}$$

where W is the body weight in kg, and ΔV is the error of body volume measurement in L. If we assume $W = 60$ kg, then an error of 0.5 L in ΔV would lead to an over 4% difference in %BF. A small error in body volume measurement can readily result from inaccuracy of lung volume estimate or a slight body movement during imaging.

In our study, subjects' residual lung volume was estimated through empirical equations (8.1) that are functions of height and age. These equations were predicated from 245 healthy nonsmokers with a variety of body builds and a wide range of ages [165]. The 95% confidence intervals of the estimation were all below 0.8 L for both men and women. As pointed out in another study [166], using a predicted rather than a measured lung residual would not significantly affect %BF estimates in adults. While this statement might be considered acceptable in group comparison, the difference between true and predicted residual volume may have an impact on derived %BF at an individual level. Since these equations do not account for body weight and body shape, it is very likely that they may overestimate the residual volume on tall and skinny persons, and underestimate on muscular persons who exercise daily and thus have an active lung function. Note from Table 8.6 that subject 1 and 15 were both tall and with sub-twenty BMIs. Their actual lung

residual may be well below the average of persons of their heights, but were estimated to be larger than two muscular subjects (9 and 10) who exercises regularly. This could explain the tendency that the stereo vision system underestimates %BF on tall and thin persons, and overestimate on muscular persons due to inaccurate estimation of lung residual. It would be prudent to measure lung residual whenever possible; however, for a body imaging system that is designed to simplify the measurement steps, it may not be practical to implement.

Even if a person's body volume is measured with good accuracy, the person's actual densities of body components can still contribute to errors in %BF estimation from the two-component body model. The two-component model assumes the fat mass density to be 0.9 kg/L and fat-free mass density to be 1.09 kg/L. For example, a 60 kg person with 20% of body fat should have a body volume of 57.37 L according the two-component body model. However, if the person actually has a lower fat-free mass density of 1.05 km/L (0.04 kg/L less than the average), the actual %BF should be 23.8% with 14.3 kg of fat mass and 45.7 kg of fat-free mass (the total body weight and volume stay unchanged). In other words, the two-component body model may underestimate %BF for persons with his or her fat-free mass density lower than average. Among our volunteers for system evaluation, subject 11 has a bone density 0.1 kg/L lower than average. This brings down his fat-free mass density, and the stereo vision gave a lower %BF estimation than DEXA.

8.4.3 Comparison of Results

Compared to the active stereo body imaging system developed in Yu's work [14], this reported passive stereo body imaging system reached better measurement agreement in %BF estimation. In [14], three %BF estimation methods were evaluated, and the methods between active stereo and air displacement plethysmography (ADP) were found to have the highest measurement agreement (-0.789% bias and 8.189% reproducibility coef-

ficient). ADP computes whole body volume through air pressure change within a sealed chamber, and its %BF estimation is based on the same Siri's equation, while the other method for %BF estimation is based on bioelectrical impedance. This may explain that active stereo vision and ADP had the highest measurement agreement in Yu's findings. Our achieved agreement was also better than the work in [166], in which ADP was compared to underwater weighing (UWW). Their reported 95% confidence interval was -7 to $+9$ %BF, and the SD was approximately 4 %BF. More comparative study results are listed in Table 8.10. Recent %BF studies [167–169] have compared the DEXA to the four-component (4-C) body model, which is believed to be more accurate in %BF estimation. It was suggested in [170] that in order to verify two %BF estimation methods to be similar, an SD between 2% and 3% and a systematic bias less than 2% are both essential. For the %BF estimation through our passive stereo vision system, both the bias and standard error were within the advised criterion, and were among the lowest ones in Table 8.10. The improved measurement agreement report for this developed passive stereo imaging system benefits most from the high resolution imaging and the robustness of the stereo computation.

In general, this study confirms that the methodological error associated with body imaging appears to be comparable, and perhaps favorable to that of ADP and UWW, which all derive %BF from body density. Nevertheless, despite good reliability shown in our findings and other studies, there still remains certain methodological issues which may affect accuracy and precision. For example, the impact of body hair, gender difference, and the inflation of body surface by clothes. In addition, it is important to note that, to date, accuracy of body composition by body imaging compared with other methods has yet to be confirmed. The difference seen between UWW, ADP and DEXA were reviewed in [174] and demonstrated limits of agreement between ADP and DEXA or UWW of approximately ± 4 %BF. This potentially large difference could be due to the combined

Table 8.10: Comparison of %BF estimation from multiple studies.

	Paired comparison		Sample size		Mean diff.	Individual diff.	
References	Method 1	Method 2	M	F	(%BF)	(%BF)	*P*-value
This dissertation	Passive stereo	DEXA	12	8	-0.64 ± 2.15	-3.6–2.6	0.20
Yu, 2008 [14]	Active stereo	ADP	10	10	-0.79 ± 4.18	–	0.41
Prior *et al.*, 1999 [171]	DEXA	UWW	91	81	-0.8 ± 3.0	-8.4–9.3	–
Fields *et al.*, 2001 [172]	ADP	DEXA	0	43	0.6 ± 3.4	–	–
Millard-Stafford *et al.*, 2001 [173]	ADP	DEXA	50 (total)		-2.5 ± 3.7	–	–
Demerath *et al.*, 2002 [166]	ADP	UWW	41	46	1.6 ± 4.0	–	< 0.05
Mahon *et al.*, 2007 [167]	DEXA	4-C	0	29	0.6 ± 4.5	–	–
Minderico *et al.*, 2008 [168]	DEXA	4-C	48	0	-1.7 ± 2.0	–	–
Santos *et al.*, 2010 [169]	DEXA	4-C	7	0	0.81 ± 2.3	–	–

Mean difference is presented in $(Mean \pm SD)$ format; ADP, air displacement plythsmography; DEXA, dual-energy X-ray absorptimetry; UWW, underwater weighting; 4-C, four-component body model; M, male; F, female.

imprecision of the methodologies within each, but also due to true methodological differences. It is technically difficult to state better accuracy of one method over another as differences between methods could be due to errors in the underlying assumptions they rely on, such as the assumed constancy of FFM in densitometry, or tissue hydration assumptions made in DEXA. To fully elucidate this issue more comparative studies of body imaging with other body composition methods need to be undertaken, methods that do not rely as heavily on physiological or chemical assumptions such as MRI, or methods based on body shape analysis.

8.5 Summary

We have evaluated the automatic body measurement system based on passive stereo vision, which is an extension to its earlier form (active stereo) dedicated to the needs of body composition assessment. Various body circumference measurements were automatically detected and calculated on reconstructed 3D body models. The computation

of both circumferences and volume are accelerated on graphics hardware. %BF is estimated by Siri's equation using the average body density with the available body weight and volume. The measurements were highly repeatably both in mannequins and in human subjects. The %BF estimates were found to be consistent with DEXA ($P = 0.20$). The agreement of %BF estimation between stereo imaging and DEXA were tested with Bland-Altman plot. The bias was found to be -0.64 of %BF, and the limits of agreement is ± 4.2 of %BF. Compared to the active stereo imaging system, the results were improved in the form of the compactness of the limits of agreement. Our results meet the advised criterion justifying two %BF estimation methods, and reached lower limits of agreement than most of other similar studies. The limitation of using a two-component body model through an empirical equation for %BF estimation were also discussed.

Chapter 9

Conclusions and Future Work

9.1 Summary of the Dissertation

The prevalence of obesity has made it necessary to develop a reliable and safe tool for timely assessing and monitoring obesity in public health. After reviewing various 3D imaging techniques and their application in body imaging for body composition analysis, we suggested that 3D anthropometry based on passive stereo vision can provide convenient and accommodating means for the body composition assessment purpose.

This dissertation reports our efforts on developing such a system with the goal to make it more affordable, reliable and easy to deploy. A total of eight cameras were used to capture stereo images for 3D reconstruction. The system is configured to a two-stance design that is the minimum configuration required for whole body imaging. The system is calibrated in two stages: camera calibration and 3D registration. The camera calibration procedure involves relatively more work than the 3D registration, because the poses of a calibration target needs to be changed several times for better results. However, camera calibration does not need to be repeated frequently as long as the relative positions of cameras in a stereo pair stay unchanged. This can be readily achieved by fixing the cameras and locking the lenses. Therefore, only 3D registration needs to be redone when the system is deployed to a new location. This property meets the portability requirement of the system, and it effectively reduces cost of maintenance.

The hardware requirements of a passive stereo vision system are relatively low when comparing it with other active imaging techniques, such as laser scanner and struc-

tured light. Active lighting has been the most popular technique for 3D imaging, mainly due to its robustness under various lighting conditions and surface properties. But the complexity of the hardware that is involved in an active light system prevents it from being widely accepted. A passive stereo vision system has the benefits of fast image acquisition and simple hardware configuration, but its 3D computation is more complex and intensive, and it still poses great challenge in the related research community. In this dissertation, we proposed a multi-scale stereo matching strategy to meet the robustness and efficiency requirements.

Within a multi-scale stereo matching framework, an image pyramid is constructed by successively Gaussian filtering and down-sampling the original images. Stereo matching starts from the top level of the pyramid by matching large scale features. It generates a low resolution disparity map, which can be used as the disparity estimates for the next pyramid level. The matching done at the top level requires a full disparity range search. However, it only has to be done on the smallest image on the top of the pyramid. Successive matching takes the disparity map computed from a previous level as an estimate, and only searches a narrow range for the correct disparity value at current scale. This effectively reduces the amount of computation by avoiding the unnecessary searches out of the range, which is an expensive operation in the dense stereo matching algorithm. A full-sized disparity map is generated when the matching on the lowest level of the pyramid is done.

A stereo matching is sensitive to lighting conditions and surface textures. Passive stereo system usually suffer from poor image quality. To develop a robust stereo matching algorithm that is appropriate for body imaging, we took design inspirations from various sources and implemented the algorithm within a classic four-step matching framework. Pixel-wise matching cost was computed first. We used a hybrid matching cost function

which includes three cost terms. These cost terms take into account both local and regional color information, and generate a combined matching cost that is less sensitive to noise. Following cost computation, cost aggregation applies constraints to the computed cost volume to support smoothness by penalizing changes within a neighborhood. Our cost aggregation follows the multiple linear paths around a pixel for aggregation. This method simulates global optimization and can be performed in parallel because every aggregation path is independent. Our disparity computation step implements the winner-takes-all strategy. It converts a 3D cost volume into a 2D disparity map. With relative reliable results from the previous aggregation step, the winner-takes-all is extremely fast. The final disparity refinement step corrects any errors in the computed disparity map with various constraints. It first identify the occlusion regions on a disparity map, because the occluded pixels and mismatched pixels need to be handled differently. An iterative region voting method is then applied on mismatched and occluded areas, allowing reliable disparity values from a neighborhood region to propagate into the problematic areas. Once region voting finishes, disparity edges are checked and made consistent to the texture map. The result of the stereo matching is a complete disparity map free of holes and smooth on continuous surfaces.

With the known camera parameters, a disparity map can be converted into a 3D point cloud in each stereo units' coordinate system. Points clouds from multiple stereo units are then merged into the global, common world coordinate system with the 3D registration results. The merged point cloud can be converted to a body surface model in triangle mesh to be more interpretable and manageable. The surface mesh generation is performed in the software that was previously developed for our active stereo system. To make the 3D anthropometry system ready for practical use, automatic body measurement is indispensable. A body measurement system dedicated to body composition assessment was

developed based on an earlier system. The function of 3D measurement were enhanced by taking advantage of graphics hardware APIs. The parameters that are made available from the hardware accelerated methods include circumference, whole body volume, segmental volumes, cross-sectional areas, and body surface areas.

The overall performance of the presented system was evaluated. The measurements were highly repeatable. The chest, waist, and hip circumference measurements were found to be accurate and reliable. The %BF estimate based on 3D body models shows no significant difference compared to DEXA. The limits of agreement of %BF estimation between stereo and DEXA were found to be comparable with other studies. Despite good repeatability shown in our findings, there still remains certain methodological issues which affect accuracy and precision. In general, %BF estimate derived from body density is sensitive to body volume measurement, and an accurate estimation of lung residual volume is usually difficult to achieve with simple steps. The potentially disagreement between stereo imaging and DEXA could be due to the combined imprecision of the two methodologies, but also due to two methodological differences, i.e., the densitometry assumes constancy of FFM while DEXA assumes the constancy in tissue hydration.

9.2 Suggestions on Future Work

In our current hardware configuration, eight cameras were used to form four stereo units. Each stereo unit covers half of body on one side. However, in our multi-scale matching framework, the disparity map computed at half-size images has already shown good quality. This indicates that there is potential to reduce the number of cameras by half, so that each stereo unit covers the whole body on one side. The resolution of stereo images in a four-camera system is at the same level as the images that are reduced by half in the eight-camera system. A system with less cameras are easier to deploy and requires less

work for cameras calibration. The stereo matching also costs less time due to fewer images to process. Another improvement to the imaging system in terms of hardware configuration is to place cameras in a multi-view setup, in which cameras are placed on a circle with equal distances between two adjacent cameras. In the binocular stereo setup, two cameras are placed close to each other and only one 3D observation is made from the stereo unit. While in the circular setup, the number of 3D observations is the same as to the number of cameras in use. A body imaging system that is configured in the multi-view setup can also reduce occlusion.

As is common in almost all passive stereo imaging, lighting always plays an important role for a successful matching. Specularity on oily skins is a problem when doing imaging under direct lighting. This usually occurs on faces. Specular areas are textureless, and they typically causes mismatch and distort the reconstruct mesh. Ways to deal with this including preventing it from happening in the first place by using indirect lighting. The desirable solution is a studio setup for photography, in which light intensity is sufficient and specular reflection is minimized. When professional studio setup is not available or limited in space, cross-polarization lighting may be an alternative.

There is still room for improvement on the algorithm for 3D surface mesh generation. In the current implementation, the 3D point cloud is first re-sampled on a regular grid and then triangulated to form the initial mesh. Re-sampling on a regular grid simplifies the triangulation, but it loses rapid geometrical changes on the surface. An example of this is that the fine surface details achieved through sub-pixel refinement with bilateral filtering are no longer visible in the reconstructed 3D surface mesh. It is preferred to make the algorithm more adaptive to local geometrical details when doing re-sampling. Improvements can also be made in the fusion of front and back surfaces to create the whole body model. Edges of the meshes are currently stitched by connecting front and back edge

points to form triangles. Direct connection of edge points takes short cut on body curves, and it may result in reduced body circumference and volume. A better way to avoid this is to extrapolate the side point, or to perform a curve fitting on the circumference to recover a few side points.

The capability of the developed body imaging system is beyond the requirement for whole body imaging. Localized body parts, such as fingers and face, are also very visible in the reconstructed body models. This system can be easily reconfigured to image regional body parts with greater resolution. In an unreported study, we have successfully imaged faces without modifying the stereo matching and surface reconstruction algorithms. However, to achieve even better results for high resolution 3D imaing, some optimizations may be need to handle the rich geometrical features at a fine scale.

The potential of the applications of 3D body imaging in public health is enormous. For example, it may be of great value if new indices can be developed for estimating the distribution of body fat in localized regions or more directly predicting health risks. The accuracy of %BF estimation may by improved by using a more reliable model that does not rely on the average body density but on body shape and dimensions. A body imaging system that is conveniently accessible to its users is an ideal tool for tracing changes in body size and shape and monitoring related health conditions.

Bibliography

[1] R. S. Ahima and M. A. Lazar, "The health risk of obesitybetter metrics imperative," *Science*, vol. 341, no. 6148, pp. 856–858, 2013.

[2] F. Sassi, *Obesity and the economics of prevention: fit not fat*. OECD Publishing, 2010.

[3] A. Must, J. Spadano, E. H. Coakley, A. E. Field, G. Colditz, and W. H. Dietz, "The disease burden associated with overweight and obesity," *JAMA: the journal of the American Medical Association*, vol. 282, no. 16, pp. 1523–1529, 1999.

[4] D. W. Haslam and W. P. T. James, "Obesity," *Lancet*, vol. 366, no. 9492, pp. 1197–1209, 2005.

[5] WHO, "Obesity: preventing and managing the global epidemic," WHO, Tech. Rep. WHO Technical Report Series 894, 2000.

[6] T. Rankinen, S.-Y. Kim, L. Perusse, J.-P. Després, and C. Bouchard, "The prediction of abdominal visceral fat level from body composition and anthropometry: Roc analysis," *International journal of obesity*, vol. 23, no. 8, pp. 801–809, 1999.

[7] P. Dempster, S. Aitkens *et al.*, "A new air displacement method for the determination of human body composition," *Medicine and Science in Sports and Exercise*, vol. 27, no. 12, pp. 1692–1697, 1995.

[8] WHO, "Physical status: The use and interpretation of anthropometry," WHO, Tech. Rep. WHO Technical Report Series 854:9, 1995.

[9] P. Björntorp *et al.*, "The regulation of adipose tissue distribution in humans." *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity*, vol. 20, no. 4, p. 291, 1996.

[10] P. Björntorp, "Centralization of body fat," in *International Textbook of Obesity*. New York: Wiley, 2001, ch. 16, pp. 213–224.

[11] M. Lean, T. Han, and C. Morrison, "Waist circumference as a measure for indicating need for weight management," *BMJ: British Medical Journal*, vol. 311, no. 6998, p. 158, 1995.

[12] C. L. Istook and S.-J. Hwang, "3d body scanning systems with application to the apparel industry," *Journal of Fashion Marketing and Management*, vol. 5, no. 2, pp. 120–132, 2001.

[13] J.-M. Lu, M.-J. J. Wang, C.-W. Chen, and J.-H. Wu, "The development of an intelligent system for customized clothing making," *Expert Systems with Applications*, vol. 37, no. 1, pp. 799–803, 2010.

[14] W. Yu, "Development of a three-dimensional anthropometry system for human body composition assessment," Ph.D. dissertation, University of Texas at Austin, 2008.

[15] A. G. Renehan, M. Tyson, M. Egger, R. F. Heller, and M. Zwahlen, "Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies," *The Lancet*, vol. 371, no. 9612, pp. 569–578, 2008.

[16] D. P. Guh, W. Zhang, N. Bansback, Z. Amarsi, C. L. Birmingham, and A. H. Anis, "The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis," *BMC public health*, vol. 9, no. 1, p. 88, 2009.

[17] M. Nichols, A. de Silva-Sanigorski, J. Cleary, S. Goldfeld, A. Colahan, and B. Swinburn, "Decreasing trends in overweight and obesity among an australian population of preschool children," *International Journal of Obesity*, vol. 35, no. 7, pp. 916–924, 2011.

[18] D. Withrow and D. Alter, "The economic burden of obesity worldwide: a systematic review of the direct costs of obesity," *Obesity Reviews*, vol. 12, no. 2, pp. 131–141, 2011.

[19] Y. Wang, M. A. Beydoun, L. Liang, B. Caballero, and S. K. Kumanyika, "Will all americans become overweight or obese? estimating the progression and cost of the us obesity epidemic," *Obesity*, vol. 16, no. 10, pp. 2323–2330, 2008.

[20] K. E. Thorpe, C. S. Florence, D. H. Howard, and P. Joski, "The impact of obesity on rising medical spending," *HEALTH AFFAIRS-MILLWOOD VA THEN BETHESDA MA-*, vol. 23, pp. 283–283, 2004.

[21] S. J. Olshansky, D. J. Passaro, R. C. Hershow, J. Layden, B. A. Carnes, J. Brody, L. Hayflick, R. N. Butler, D. B. Allison, and D. S. Ludwig, "A potential decline in life expectancy in the united states in the 21st century," *New England Journal of Medicine*, vol. 352, no. 11, pp. 1138–1145, 2005.

[22] M. Finucane, G. Stevens, M. Cowan, G. Danaei, J. Lin, C. Paciorek, G. Singh, H. Gutierrez, Y. Lu, A. Bahalim *et al.*, "Global burden of metabolic risk factors of chronic diseases collaborating group (body mass index). national, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants," *Lancet*, vol. 377, no. 9765, pp. 557–567, 2011.

[23] F. Sassi, M. Devaux, M. Cecchini, and E. Rusticelli, "The obesity epidemic: analysis of past and projected future trends in selected oecd countries," OECD Publishing, Tech. Rep., 2009.

[24] M. M. Finucane, G. A. Stevens, M. J. Cowan, G. Danaei, J. K. Lin, C. J. Paciorek, G. M. Singh, H. R. Gutierrez, Y. Lu, A. N. Bahalim *et al.*, "National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9· 1 million participants," *The Lancet*, vol. 377, no. 9765, pp. 557–567, 2011.

[25] T. Lobstein, L. Baur, and R. Uauy, "Obesity in children and young people: a crisis in public health," *Obesity reviews*, vol. 5, no. s1, pp. 4–85, 2004.

[26] B. Rokholm, J. Baker, and T. Sørensen, "The levelling off of the obesity epidemic since the year 1999: a review of evidence and perspectives," *Obesity Reviews*, vol. 11, no. 12, pp. 835–846, 2010.

[27] E. A. Finkelstein, J. G. Trogdon, J. W. Cohen, and W. Dietz, "Annual medical spending attributable to obesity: payer-and service-specific estimates," *Health affairs*, vol. 28, no. 5, pp. w822–w831, 2009.

[28] E. A. Finkelstein, I. C. Fiebelkorn, and G. Wang, "State-level estimates of annual medical expenditures attributable to obesity*," *Obesity research*, vol. 12, no. 1, pp. 18–24, 2004.

[29] E. A. Finkelstein, I. C. Fiebelkorn, G. Wang *et al.*, "National medical spending attributable to overweight and obesity: how much, and who's paying?" *HEALTH AFFAIRS-MILLWOOD VA THEN BETHESDA MA-*, vol. 22, no. 3; SUPP, pp. W3–219, 2003.

[30] J. Trogdon, E. Finkelstein, T. Hylands, P. Dellea, and S. Kamal-Bahl, "Indirect costs of obesity: a review of the current literature," *Obesity Reviews*, vol. 9, no. 5, pp. 489–500, 2008.

[31] B. Popkin, S. Kim, E. Rusev, S. Du, and C. Zizza, "Measuring the full economic costs of diet, physical activity and obesity-related chronic diseases," *obesity reviews*, vol. 7, no. 3, pp. 271–293, 2006.

[32] E. A. Finkelstein, M. daCosta DiBonaventura, S. M. Burgess, B. C. Hale *et al.*, "The costs of obesity in the workplace," *Journal of Occupational and Environmental Medicine*, vol. 52, no. 10, pp. 971–976, 2010.

[33] R. Roubenoff, G. E. Dallal, and P. Wilson, "Predicting body fatness: the body mass index vs estimation by bioelectrical impedance." *American journal of public health*, vol. 85, no. 5, pp. 726–728, 1995.

[34] D. C. Frankenfield, W. A. Rowe, R. N. Cooney, J. S. Smith, and D. Becker, "Limits of body mass index to detect obesity and predict body composition," *Nutrition*, vol. 17, no. 1, pp. 26–30, 2001.

[35] R. V. Burkhauser and J. Cawley, "Beyond bmi: the value of more accurate measures of fatness and obesity in social science research," *Journal of health economics*, vol. 27, no. 2, pp. 519–529, 2008.

[36] K. J. Smalley, A. N. Knerr, Z. V. Kendrick, J. A. Colliver, and O. E. Owen, "Reassessment of body mass indices." *The American journal of clinical nutrition*, vol. 52, no. 3, pp. 405–408, 1990.

[37] T. VanItallie, M.-U. Yang, S. B. Heymsfield, R. C. Funk, and R. A. Boileau, "Height-normalized indices of the body's fat-free mass and fat mass: potentially useful indi-

cators of nutritional status." *The American journal of clinical nutrition*, vol. 52, no. 6, pp. 953–959, 1990.

[38] Z.-M. Wang, R. Pierson, and S. B. Heymsfield, "The five-level model: a new approach to organizing body-composition research." *The American journal of clinical nutrition*, vol. 56, no. 1, pp. 19–28, 1992.

[39] W. E. Siri, "Body composition from fluid spaces and density: analysis of methods," *Techniques for measuring body composition*, vol. 61, pp. 223–44, 1961.

[40] J. Brožek, F. Grande, J. T. Anderson, and A. Keys, "Densitometric analysis of body composition: revision of some quantitative assumptions*," *Annals of the New York Academy of Sciences*, vol. 110, no. 1, pp. 113–140, 1963.

[41] S. B. Going, *Human body composition*, 2nd ed. Champaign, IL: Human Kinetics, 2005, ch. 2 Hydrodensitometry and air displacement plethysmography.

[42] E. E. Noreen and P. Lemon, "Reliability of air displacement plethysmography in a large, heterogeneous sample." *Medicine and science in sports and exercise*, vol. 38, no. 8, pp. 1505–1509, 2006.

[43] D. E. Anderson, "Reliability of air displacement plethysmography," *The Journal of Strength & Conditioning Research*, vol. 21, no. 1, pp. 169–172, 2007.

[44] E. Rush, V. Chandu, and L. Plank, "Prediction of fat-free mass by bioimpedance analysis in migrant asian indian men and women: a cross validation study," *International journal of obesity*, vol. 30, no. 7, pp. 1125–1131, 2006.

[45] P. Deurenberg, "Limitations of the bioelectrical impedance method for the assessment of body fat in severe obesity." *The American journal of clinical nutrition*, vol. 64, no. 3, pp. 449S–452S, 1996.

[46] R. B. Mazess, H. S. Barden, J. P. Bisek, and J. Hanson, "Dual-energy x-ray absorptiometry for total-body and regional bone-mineral and soft-tissue composition." *The American journal of clinical nutrition*, vol. 51, no. 6, pp. 1106–1112, 1990.

[47] A. Pietrobelli, C. Formica, Z. Wang, S. B. Heymsfield *et al.*, "Dual-energy x-ray absorptiometry body composition model: review of physical concepts," *American Journal of Physiology-Endocrinology And Metabolism*, vol. 34, no. 6, p. E941, 1996.

[48] R. Patel, G. Blake, S. Batchelor, and I. Fogelman, "Occupational dose to the radiographer in dual x-ray absorptiometry: a comparison of pencil-beam and fan-beam systems," *The British journal of radiology*, vol. 69, no. 822, pp. 539–543, 1996.

[49] C. F. Njeh, T. Fuerst, D. Hans, G. M. Blake, and H. K. Genant, "Radiation exposure in bone mineral density assessment," *Applied radiation and isotopes*, vol. 50, no. 1, pp. 215–236, 1999.

[50] J. Cleary, S. Daniells, A. D. Okely, M. Batterham, and J. Nicholls, "Predictive validity of four bioelectrical impedance equations in determining percent fat mass in overweight and obese children," *Journal of the American Dietetic Association*, vol. 108, no. 1, pp. 136–139, 2008.

[51] S. B. Going, M. P. Massett, M. C. Hall, L. A. Bare, P. A. Root, D. P. Williams, and T. G. Lohman, "Detection of small changes in body composition by dual-energy x-ray absorptiometry." *The American journal of clinical nutrition*, vol. 57, no. 6, pp. 845–850, 1993.

[52] W. M. Kohrt, "Preliminary evidence that dexa provides an accurate assessment of body composition," *Journal of applied physiology*, vol. 84, no. 1, pp. 372–377, 1998.

[53] T. G. Lohman and Z. Chen, *Human body composition*, 2nd ed. Champaign, IL: Human Kinetics, 2005, ch. 5 Dual-energy X-ray Absorptiometry.

[54] J. E. Williams, J. C. Wells, C. M. Wilson, D. Haroun, A. Lucas, and M. S. Fewtrell, "Evaluation of lunar prodigy dual-energy x-ray absorptiometry for assessing body composition in healthy persons and patients by comparison with the criterion 4-component model," *The American journal of clinical nutrition*, vol. 83, no. 5, pp. 1047–1054, 2006.

[55] D. Gallagher, P. Kuznia, S. Heshka, J. Albu, S. B. Heymsfield, B. Goodpaster, M. Visser, and T. B. Harris, "Adipose tissue in muscle: a novel depot similar in size to visceral adipose tissue," *The American journal of clinical nutrition*, vol. 81, no. 4, pp. 903–910, 2005.

[56] J. Yim, S. Heshka, J. Albu, S. Heymsfield, P. Kuznia, T. Harris, and D. Gallagher, "Intermuscular adipose tissue rivals visceral adipose tissue in independent associations with cardiovascular risk," *International journal of obesity*, vol. 31, no. 9, pp. 1400–1405, 2007.

[57] D. Gallagher, J. Albu, Q. He, S. Heshka, L. Boxt, N. Krasnow, and M. Elia, "Small organs with a high metabolic rate explain lower resting energy expenditure in african american than in white adults," *The American journal of clinical nutrition*, vol. 83, no. 5, pp. 1062–1067, 2006.

[58] J. Wang, D. Gallagher, J. C. Thornton, W. Yu, M. Horlick, and F. X. Pi-Sunyer, "Validation of a 3-dimensional photonic scanner for the measurement of body volumes, dimensions, and percentage body fat," *The American journal of clinical nutrition*, vol. 83, no. 4, pp. 809–816, 2006.

[59] F. Blais, M. Picard, and G. Godin, "Accurate 3d acquisition of freely moving objects," in *Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, 2004, pp. 422–429.

[60] S. Goel and B. Lohani, "A motion correction technique for laser scanning of moving objects," *Geoscience and Remote Sensing Letters, IEEE*, vol. 11, no. 1, pp. 225–228, Jan 2014.

[61] J. Batlle, E. Mouaddib, and J. Salvi, "Recent progress in coded structured light as a technique to solve the correspondence problem: a survey," *Pattern recognition*, vol. 31, no. 7, pp. 963–982, 1998.

[62] H. Sagan, *Space-filling curves*. Springer-Verlag New York, 1994, vol. 18.

[63] I. Ishii, K. Yamamoto, T. Tsuji *et al.*, "High-speed 3d image acquisition using coded structured light projection," in *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*. IEEE, 2007, pp. 925–930.

[64] S. Inokuchi, K. Sato, and F. Matsuda, "Range imaging system for 3-d object recognition," in *Proceedings of the International Conference on Pattern Recognition*, 1984, pp. 806–808.

[65] H. Zhao, W. Chen, and Y. Tan, "Phase-unwrapping algorithm for the measurement of three-dimensional object shapes," *Applied Optics*, vol. 33, no. 20, pp. 4497–4500, 1994.

[66] D. C. Ghiglia and M. D. Pritt, *Two-dimensional phase unwrapping: theory, algorithms, and software*. Wiley New York:, 1998.

[67] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[68] N. Lazaros, G. C. Sirakoulis, and A. Gasteratos, "Review of stereo vision algorithms: from software to hardware," *International Journal of Optomechatronics*, vol. 2, no. 4,

pp. 435–462, 2008.

[69] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 6, pp. 1068–1080, 2008.

[70] P. Treleaven, "Sizing us up," *IEEE Spectrum*, vol. 41, no. 4, pp. 28–31, 2004.

[71] N. D Apuzzo, "3d body scanning technology for fashion and apparel industry," in *Proc. SPIE 6491, Videometrics IX*, San Jose, 2007, pp. 64 910O–64 910O–12.

[72] D. Addleman and L. Addleman, "Rapid 3d digitizing," *Computer Graphics World*, vol. 8, no. 11, p. 41, 1985.

[73] P. R. Jones, G. M. West, D. H. Harris, and J. B. Read, "The loughborough anthropometric shadow scanner (lass)," *Endeavour*, vol. 13, no. 4, pp. 162–168, 1989.

[74] K. Konolige and P. Mihelich. (2012) Technical description of kinect calibration. Accessed: 2014-02-16. [Online]. Available: http://wiki.ros.org/kinect_calibration/technical

[75] D. Herrera, J. Kannala, and J. Heikkilä, "Accurate and practical calibration of a depth and color camera pair," in *Computer Analysis of Images and Patterns*. Springer, 2011, pp. 437–445.

[76] C. Zhang and Z. Zhang, "Calibration between depth and color sensors for commodity depth cameras," in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–6.

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

[78] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 519–528.

[79] H. Tao, H. S. Sawhney, and R. Kumar, "A global matching framework for stereo computation," in *ICCV 2001. Eighth IEEE International Conference on Computer Vision, 2001.*, vol. 1. IEEE, 2001, pp. 532–539.

[80] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, 2nd ed. Cambridge: Cambridge university press, 2003.

[81] D. A. Forsyth and J. Ponce, *Computer vision: a modern approach*, 2nd ed. Prentice Hall Professional Technical Reference, 2011.

[82] V. Papadimitriou and T. J. Dennis, "Epipolar line estimation and rectification for stereo image pairs," *Image Processing, IEEE Transactions on*, vol. 5, no. 4, pp. 672–676, 1996.

[83] C. Loop and Z. Zhang, "Computing rectifying homographies for stereo vision," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 1. IEEE, 1999.

[84] M. Pollefeys, R. Koch, and L. Van Gool, "A simple and efficient rectification method for general motion," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1. IEEE, 1999, pp. 496–501.

[85] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 593–600.

[86] R. Šára, "Finding the largest unambiguous component of stereo matching," in *Computer VisionECCV 2002*. Springer, 2002, pp. 900–914.

[87] J. Cech and R. Šára, "Efficient sampling of disparity space for fast and accurate matching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[88] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[89] S. Birchfield and C. Tomasi, "Depth discontinuities by pixel-to-pixel stereo," *International Journal of Computer Vision*, vol. 35, no. 3, pp. 269–293, 1999.

[90] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3. IEEE, 2006, pp. 15–18.

[91] H. Hirschmüller, P. R. Innocent, and J. Garibaldi, "Real-time correlation-based stereo vision with reduced border errors," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 229–246, 2002.

[92] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Computer VisionECCV'94*. Springer, 1994, pp. 151–158.

[93] H. Hirschmüller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[94] R. Yang and M. Pollefeys, "Multi-resolution real-time stereo on commodity graphics hardware," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1. IEEE, 2003, pp. I–211.

[95] 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 *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 1075–1082.

[96] A. F. Bobick and S. S. Intille, "Large occlusion stereo," *International Journal of Computer Vision*, vol. 33, no. 3, pp. 181–200, 1999.

[97] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window: Theory and experiment," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, no. 9, pp. 920–932, 1994.

[98] O. Veksler, "Stereo matching by compact windows via minimum ratio cycle," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 540–547.

[99] K.-J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 650–656, 2006.

[100] S. Mattoccia, S. Giardino, and A. Gambini, "Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering," in *Computer Vision–ACCV 2009*. Springer, 2010, pp. 371–380.

[101] B. Xu and Y. Huang, "3d technology for apparel mass customization, part i: rotary body scanning," *Journal of Textile Institute*, vol. 94, no. 1, pp. 72–80, 2003.

[102] B. Xu, M. R. Pepper, J. H. Freeland-Graves, W. Yu, and M. Yao, "Three-dimensional surface imaging system for assessing human obesity," *Optical Engineering*, vol. 48, no. 10, pp. 107 204–107 204, 2009.

[103] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross, "High-quality single-shot capture of facial geometry," *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, p. 40, 2010.

[104] Z. Zhang, "A flexible new technique for camera calibration," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 11, pp. 1330–1334, 2000.

[105] "Opencv library," http://opencv.org/, accessed: 2014-02-27.

[106] R. M. Haralock and L. G. Shapiro, *Computer and robot vision*. Addison-Wesley Longman Publishing Co., Inc., 1991.

[107] B. K. Horn, "Closed-form solution of absolute orientation using unit quaternions," *JOSA A*, vol. 4, no. 4, pp. 629–642, 1987.

[108] R. Micheals and T. Boult, "A new closed-form approach to the absolute orientation problem," VAST Lab, Lehigh University, Bethlehem, PA, Tech. Rep., 1999.

[109] X. Mei, X. Sun, M. Zhou, H. Wang, X. Zhang *et al.*, "On building an accurate stereo matching system on graphics hardware," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 467–474.

[110] D. Scharstein and R. Szeliski, "Middlebury stereo," http://vision.middlebury.edu/stereo/, accessed: 2014-02-25.

[111] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 4, pp. 401–406, 1998.

[112] E. P. Baltsavias and D. Stallmann, *SPOT stereo matching for Digital Terrain Model generation*. Swiss Federal Institute of Technology, Institute of Geodesy and Photogrammetry, 1993.

[113] K. Konolige, "Small vision systems: Hardware and implementation," in *Proc. Eighth Intl Symp. Robotics Research*. Springer, 1998, pp. 203–212.

[114] D. N. Bhat and S. K. Nayar, "Ordinal measures for image correspondence," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 4, pp. 415–423, 1998.

[115] R. Sara and R. Bajcsy, "On occluding contour artifacts in stereo vision," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. IEEE, 1997, pp. 852–857.

[116] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International journal of computer vision*, vol. 24, no. 2, pp. 137–154, 1997.

[117] G. Egnal, "Mutual information as a stereo correspondence measure," Computer and Information Science, Univ. of Pennsylvania, Technical Report MS-CIS-00-20, 2000.

[118] C. Fookes, M. Bennamoun, and A. Lamanna, "Improved stereo image matching using mutual information and hierarchical prior probabilities," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 2. IEEE, 2002, pp. 937–940.

[119] I. Sarkar and M. Bansal, "A wavelet-based multiresolution approach to solve the stereo correspondence problem using mutual information," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 37, no. 4, pp. 1009–1014, 2007.

[120] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in *ACM Transactions on*

Graphics (TOG), vol. 23, no. 3. ACM, 2004, pp. 600–608.

[121] J. Kim, V. Kolmogorov, and R. Zabih, "Visual correspondence using energy minimization and mutual information," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 1033–1040.

[122] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 2, pp. 328–341, 2008.

[123] Y. Ohta and T. Kanade, "Stereo by intra-and inter-scanline search using dynamic programming," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 2, pp. 139–154, 1985.

[124] I. J. Cox, S. L. Hingorani, S. B. Rao, and B. M. Maggs, "A maximum likelihood stereo algorithm," *Computer vision and image understanding*, vol. 63, no. 3, pp. 542–567, 1996.

[125] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *International journal of computer vision*, vol. 70, no. 1, pp. 41–54, 2006.

[126] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 7, pp. 787–800, 2003.

[127] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1124–1137, 2004.

[128] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[129] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Computer Vision, 1998. Sixth International Conference on*. IEEE, 1998, pp. 839–846.

[130] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[131] S. Paris and F. Durand, "A fast approximation of the bilateral filter using a signal processing approach," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 568–580.

[132] F. Porikli, "Constant time o (1) bilateral filtering," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[133] Q. Yang, K.-H. Tan, and N. Ahuja, "Real-time o (1) bilateral filtering," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 557–564.

[134] Q. Yang, S. Wang, and N. Ahuja, "Svm for edge-preserving filtering," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1775–1782.

[135] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 510–523.

[136] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 1–14.

[137] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3017–3024.

[138] K. Zhang, J. Lu, and G. Lafruit, "Cross-based local stereo matching using orthogonal integral images," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 7, pp. 1073–1079, 2009.

[139] S. Paris, P. Kornprobst, J. Tumblin, and F. Durand, "A gentle introduction to bilateral filtering and its applications," in *ACM SIGGRAPH 2007 courses*. ACM, 2007, p. 1.

[140] J.-P. Pons, R. Keriven, and O. Faugeras, "Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 179–193, 2007.

[141] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, 2001.

[142] G. Van Meerbergen, M. Vergauwen, M. Pollefeys, and L. Van Gool, "A hierarchical symmetric stereo algorithm using dynamic programming," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 275–285, 2002.

[143] B. Chapman, G. Jost, and R. Van Der Pas, *Using OpenMP: portable shared memory parallel programming*. MIT press, 2008, vol. 10.

[144] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," in *Proceedings of the April 18-20, 1967, spring joint computer conference*. ACM, 1967, pp. 483–485.

[145] B. M. Smith, L. Zhang, and H. Jin, "Stereo matching with nonparametric smoothness priors in feature space," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 485–492.

[146] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon, "Global stereo reconstruction under second-order smoothness priors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 12, pp. 2115–2128, 2009.

[147] Q. Yang, L. Wang, R. Yang, H. Stewénius, and D. Nistér, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 3, pp. 492–504, 2009.

[148] C. Lei, J. Selzer, and Y.-H. Yang, "Region-tree based stereo using dynamic programming optimization," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2378–2385.

[149] H. Hirschmüller, "Stereo vision based mapping and immediate virtual walkthroughs," Ph.D. dissertation, School of Computing, De Montfort University, 2003.

[150] L. Wang, H. Jin, R. Yang, and M. Gong, "Stereoscopic inpainting: Joint color and depth completion from stereo images," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[151] C. Sun, "Fast stereo matching using rectangular subregioning and 3d maximum-surface techniques," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 99–117, 2002.

[152] D. Nehab, S. Rusinkiewiez, and J. Davis, "Improved sub-pixel stereo correspondences through symmetric refinement," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 557–563.

[153] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[154] M. Bleyer and M. Gelautz, "A layered stereo matching algorithm using image segmentation and global visibility constraints," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 59, no. 3, pp. 128–150, 2005.

[155] J. Westerweel, *Digital particle image velocimetry: theory and application*. TU Delft, Delft University of Technology, 1993.

[156] A. N. Stein, A. Huertas, and L. Matthies, "Attenuating stereo pixel-locking via affine window adaptation," in *Robotics and Automation. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE, 2006, pp. 914–921.

[157] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," in *Proceedings of International Joint Conference on Artificial Intelligence*, vol. 81, 1981, pp. 674–679.

[158] J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," Pittsburgh, PA, 1994.

[159] L. Dekker, I. Douros, B. Buxton, and P. Treleaven, "Building symbolic information for 3d human body modeling from range data," in *Proceedings of the Second Internaional Conference on 3D Digital Imaging and Modeling*. Ottawa, ON, Canada: IEEE Computer Society, 1999, pp. 388–397.

[160] X. Ju, N. Weighi, and J. Siebert, "Automatic segmentation of 3d human body scans," in *Proceedings of IASTED Int. Conf. on Computer Graphics and Imaging*, Lad Vegas, NV, 2000.

[161] Y. Xiao, P. Siebert, and N. Werghi, "Topological segmentation of discrete human body shape in various posture based on geodesic distance," in *Proceedings of the 17th International Conference on Pattern Recognition*, 2004, pp. 131–135.

[162] L.-F. Leong, J.-J. Fang, and M.-J. Tsai, "Automatic body feature extraction from a maker-less scanned human body," *Computer-Aided Design*, pp. 568–582, 2007.

[163] N. Werghi, "Segmentation and modeling of full human body shape from 3-d scan data: A survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 6, pp. 1122–1136, 2007.

[164] Y. Zhong and B. Xu, "Automatic segmenting and measurement on scanned human body," *International Journal of Clothing Science and Technology*, vol. 18, no. 1, pp. 19–30, 2006.

[165] R. Crapo, A. Morris, P. Clayton, and C. Nixon, "Lung volumes in healthy nonsmoking adults." *Bulletin europeen de physiopathologie respiratoire*, vol. 18, no. 3, pp. 419–425, 1981.

[166] E. Demerath, S. Guo, W. Chumlea, B. Towne, A. Roche, and R. Siervogel, "Comparison of percent body fat estimates using air displacement plethysmography and hydrodensitometry in adults and children." *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity*, vol. 26, no. 3, pp. 389–397, 2002.

[167] A. Mahon, M. Flynn, H. Iglay, L. Stewart, C. Johnson, B. McFarlin, and W. Campbell, "Measurement of body composition changes with weight loss in postmenopausal women: comparison of methods," *The journal of nutrition, health & aging*, vol. 11, no. 3, p. 203, 2007.

[168] C. S. Minderico, A. M. Silva, K. Keller, T. L. Branco, S. S. Martins, A. L. Palmeira, J. T. Barata, E. A. Carnero, P. M. Rocha, P. J. Teixeira *et al.*, "Usefulness of different techniques for measuring body composition changes during weight loss in overweight and obese women," *British Journal of Nutrition*, vol. 99, no. 02, pp. 432–441, 2008.

[169] D. A. Santos, A. M. Silva, C. N. Matias, D. A. Fields, S. B. Heymsfield, and L. B. Sardinha, "Accuracy of dxa in estimating body composition changes in elite athletes using a four compartment model as the reference method," *Nutr Metab (Lond)*, vol. 7, no. 22, pp. 7075–7, 2010.

[170] S. Heymsfield, *Human body composition*. Human kinetics, 2005, vol. 918.

[171] B. M. Prior, K. J. Cureton, C. M. Modlesky, E. M. Evans, M. A. Sloniger, M. Saunders, and R. D. Lewis, "In vivo validation of whole body composition estimates from dual-energy x-ray absorptiometry," *Journal of applied physiology*, vol. 83, no. 2, pp. 623–630, 1997.

[172] D. A. Fields, D. G. Wilson, B. L. Gladden, G. R. Hunter, D. D. Pascoe, and M. I. Goran, "Comparison of the bodpod with the four-component model in adult females," *Medicine & Science in Sports & Exercise*, vol. 33, no. 9, pp. 1605–1610, 2001.

[173] M. L. Millard-Stafford, M. A. Collins, E. M. Evans, T. K. Snow, K. J. Cureton, and L. B. Rosskopf, "Use of air displacement plethysmography for estimating body fat in a four-component model." *Medicine & Science in Sports & Exercise*, vol. 33, no. 8, pp. 1311–1317, 2001.

[174] D. A. Fields, M. I. Goran, and M. A. McCrory, "Body-composition assessment via air-displacement plethysmography in adults and children: a review," *The American journal of clinical nutrition*, vol. 75, no. 3, pp. 453–467, 2002.

Vita

Ming Yao received the Bachelor of Science degree in Communication Engineering, and the Master of Science degree in Pattern Recognition and Intelligent Systems from Donghua University, Shanghai, China, in 2003 and 2006, respectively. He was a visiting researcher in the School of Human Ecology, University of Texas at Austin, from 2006 to 2008. He was then enrolled in the Ph.D. program in Biomedical Engineering in the University of Texas at Austin, under the supervision of Dr. Bugao Xu. Throughout most of time at the University of Texas at Austin, he worked on various 2D and 3D surface imaging projects with applications in highway pavement distress detection, fabric surface characterization, and body imaging. His current research interests include image processing, computer vision, computer graphics, and high-performance computing.

E-mail address: mingyao@utexas.edu

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

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