UNIVERSITY OF CALIFORNIA, SAN DIEGO

Disparity Estimation and Enhancement for Stereo Panoramic and Multi-array Image/Video

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Communication Theory and Systems)

by

Zucheul Lee

Committee in charge:

Professor Truong Q. Nguyen, Chair
Professor Pamela C. Cosman
Professor William S. Hodgkiss
Professor David J. Kriegman
Professor Jurgen P. Schulze

2014

The dissertation of Zucheul Lee is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2014

DEDICATION

To my family.

EPIGRAPH

“With age comes wisdom, but sometimes age comes alone.”

— Oscar Wilde

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my advisor, Professor Truong Nguyen. He has always encouraged my research and given me a great deal of freedom in my work. He has been a great advisor, giving me insightful comments and constructive suggestions. I received generous support from him, and I am deeply grateful to him.

I would also like to thank my committee members, Professors Pamela Cosman, William Hodgkiss, David Kriegman, and Jurgen Schulze, for their time and help in my research. Professor Cosman helped me understand how to be a good researcher at the beginning of my Ph.D. studies. Professor Hodgkiss, Krieman, and Schulze provided me with research motivation and material as well as background knowledge. I have greatly benefited from them.

I would also like to thank other professors who helped me during my Ph.D. studies: Professors Laurence Milstein, Won Ha Kim (Kyunghee University), and Eung-Tae Kim (Korea Polytechnic University). Moreover, I have to give thanks to my colleagues in the Video Processing Lab: Kyoung-Rok Lee, Yeejin Lee, Yujia Wang, Lee-Kang Liu, Jason Juang, Enming Luo, Kris Gibson, Ramsin Khoshabeh, Stanley Chan, Ankit Jain, Byeong Keun Kang, Y. H Hsieh, Haleh Azartash, Can Bal, Menglin Zeng, Subarna Tripathi.

A special thanks goes to my parents. Words cannot express how grateful I am to my mother-in-law, mother, and father for all the sacrifices they have made on my behalf. Finally, I would like to show my greatest appreciation to my family whose prayers for me have sustained me so far. I love my daughter, Yunseo Lee, and my son, Chanyoung Lee, as they always make me happy. They are the meaning of my life. My wife, Minjung Park, has always been my best support, especially when I was having a hard time during my Ph.D. studies. I will be forever grateful for their unconditional love.

Chapter 3 and 4 is in part a reprint of a published paper in IEEE Transactions on Multimedia, 2013, and a reprint of a conference paper presented in European Signal Processing Conference, Aug 2013. Chapter 5 is in part a reprint of a conference paper presented in IEEE International Conference on Acoustics, Speech and Signal Processing, May 2014, and a reprint of a submitted paper to IEEE Transactions on Multimedia, 2014. Chapter 6 is in part a reprint of a submitted paper to IEEE Transactions on Mul-

timedia, 2014.

VITA

1997	B. S. in Electronic Engineering, Yonsei University, Seoul, Korea
2003	M. S. in Electronic Engineering, Yonsei University, Seoul, Korea
2014	Ph. D. in Electrical Computer Engineering (Communication Theory and Systems), University of California, San Diego

PUBLICATIONS

Zucheul Lee, Jason Juang, and Truong Q. Nguyen, "Local Disparity Estimation with Three-Moded Cross Census and Advanced Support Weight," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 1855 - 1864, 2013.

Zucheul Lee and Truong Q. Nguyen, "Multi-resolution Depth Processing and Fusion for Large Panoramic Image," *submitted to IEEE Transactions on Multimedia*, 2014.

Zucheul Lee and Truong Q. Nguyen, "Multi-array Camera Disparity Enhancement," *submitted to IEEE Transactions on Multimedia*, 2014.

Zucheul Lee, Ramsin Khoshabeh, Jason Juang, and Truong Q. Nguyen, "Local Stereo Matching using Motion Cue and Modified Census in Video Disparity Estimation," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pp. 1114-1118, Aug. 2012.

Zucheul Lee and Truong Q. Nguyen, "Hierarchical Depth Processing with Adaptive Search Range and Fusion," in *Proceedings of IEEE ICASSP*, pp. ,May, 2014.

ABSTRACT OF THE DISSERTATION

Disparity Estimation and Enhancement for Stereo Panoramic and Multi-array Image/Video

by

Zucheul Lee

Doctor of Philosophy in Electrical Engineering (Communication Theory and Systems)

University of California, San Diego, 2014

Professor Truong Q. Nguyen, Chair

Stereo matching involves finding corresponding pixels on other images from stereo or multiple views. Disparity estimation (stereo matching) is an essential step of 3D depth-based processing and application.

In this dissertation, we present an accurate and efficient method of one-pass local disparity estimation for stereo images. The proposed disparity method is extended to the video domain by using motion information as well as imposing spatio-temporal consistency. In the local disparity method, matching accuracy depends on precise cost measures and proper cost aggregation. To ensure the accuracy of cost measures, we propose a novel three-moded cross census transform with a noise buffer that is robust to image noises in homogeneous areas. It is revealed that the cross-square census and

combination metric of the three cost measures achieve more reliable cost measures in a variety of image regions. We further increase the aggregation accuracy by adopting the advanced support weight and incorporating motion flow.

Stereo panoramic images have advantages over regular images, such as wide field of view and high resolution. However, the large size makes the stereo matching problem challenging. In this dissertation, an effective method of multi-resolution depth processing for large panoramic images is presented. We propose an adaptive disparity search range based on the combined local structure. The adaptive range value can propagate the smoothness properties from the low-resolution level to the high-resolution level while preserving fine details and reducing undesirable errors. To reduce disparity quantization error in a hierarchical manner, we propose a reliable multiple fitting algorithm. The spatial-multi-resolution Total Variation (TV) method is employed to enforce consistency in both the spatial and scaling dimensions. The proposed algorithm is able to produce high-quality depth maps by effectively combining individual multi-scale disparities.

Multi-array camera systems have greater potential for 3D depth-based application development than stereo camera systems. However, few studies have been conducted on multi-array-based disparity estimation, due to a lack of data. In this dissertation, we propose an alternate use of local and global fusion of multi-array disparities to maximize disparity enhancement in array camera systems. We propose a new cascade regularization-based approach that can restore diagonal structures better than conventional approaches. The detailed analyses and simulation results demonstrate that the cascade approach better regularizes diagonal variations and in turn yields better image enhancement. We adapt a cascade TV regularization to the multi-array camera system in order to globally combine multiple disparities. A local multiple cross-filling algorithm is proposed to achieve cross consistency between array disparity estimates by effectively filling the mismatches. Experimental results show that the proposed multi-array disparity enhancement algorithm improves the accuracy of initial array disparity estimates up to 65% while alleviating memory limitations.

Chapter 1

Introduction

1.1 Motivation

Disparity estimation provides fundamental information for a wide range of depth-based applications, such as 3D-TV, multi-view synthesis, 3D surgery, automatic navigation, and 3D panorama. The goal of this dissertation is to propose an effective disparity-estimation and -enhancement algorithm for various stereo and multiple views using stereo and array cameras, respectively.

Disparity estimation methods are categorized as either local or global. Local stereo methods are structurally simple and very efficient in real-time processing even though they produce relatively poor-quality estimations in homogeneous areas and noisy environments. It would be desirable to obtain disparity quality comparable with that of global methods in such areas while preserving simplicity and efficiency. The accuracy of local disparity estimation depends on two main functions: similarity measures and support window selection. The associated challenges are as follows.

- Homogeneous areas do not always provide sufficient texture information for similarity measures. Moreover, noises in these areas make the matching problem difficult. On the contrary, densely textured areas often increase matching ambiguities, because complicated textures are likely to be regarded as repetitive patterns.

- Most local disparity methods use a support window where each matching cost is aggregated according to the assumption of the same disparity. We encounter a

dilemma when determining the support window size. The window size should be large enough to get more local information in homogeneous areas and small enough to avoid crossing disparity discontinuities. The ideal support window contains only pixels with the same disparity.

For the first issue, image texture characteristics are shift-variant for every support window. Therefore, each similarity measure might perform differently, depending on the different texture types. To obtain a reliable cost function in various regions, the similarity function must be selected optimally. In addition, the noisy environment scenario must be taken into account in similarity measures. For the second issue, it is practically impossible to build the ideal support window composed of only pixels with the same depth. However, the optimal segmentation for the support window can be achieved by using as many meaningful cues as possible.

As an additional issue, the direct extension of image disparity estimation to the video domain is undesirable, because it causes flickering artifacts. Therefore, some treatments enforcing temporal consistency are required. There is also the difficulty of dealing with the edges of fast-moving objects in video disparity estimation. In this case, optical flow can be an important cue in the support window construction, as color is a crucial cue in image disparity estimation.

Customers prefer large stereo panoramic images, as they provide a wide field of view as well as high resolution. However, the large size complicates the stereo matching problem, requiring high computational complexity. A multi-resolution (hierarchical) scheme is one possible solution to handling large panoramic views that are dozens of times larger than regular views. However, hierarchical processing leads to side issues, such as error propagation and blurring at object boundaries. In large panorama processing, it is challenging to obtain high quality and high-resolution panoramic disparity maps while suppressing error propagation as well as unexpected errors.

To reduce error propagation from coarse level to fine level, disparities at the next level need to be carefully determined by fully searching the maximum disparity range. On the other hand, to suppress the unexpected errors often resulting from high-scale estimation, disparity should be estimated within a minimized search range so that the desirable initial disparity can be preserved. Therefore, the optimal search range might

be position-variant. The way in which the optimal disparity search range is determined at each pixel point is central in multi-resolution depth processing.

The upsampling/downsampling processes are necessary in the pyramid scheme, but they cause scale dimensional inconsistency, as objects in the world appear differently, depending on the scale of observation. Therefore, the scaling consistency for multi-scale disparities can be enforced as temporal consistency is for video disparities. For instance, some features are not visible on a certain scale disparity map. These invisible features might be reconstructed by a process fusing the corresponding multi-scale disparities. In the pyramid scheme, disparity quantization artifacts are also likely to be observed at the low-scale estimation. The disparity quantization problem can be alleviated by using sub-pixel disparity estimation.

Multi-array camera systems might be able to yield accurate disparity estimates using redundant stereo pairs. However, there has been little research on multi-array-based disparity estimation. In addition, there are no ground-truth disparity maps of array images available for quantitative evaluation. If multi-array images and videos with associated ground-truth maps were created and shared, they might invigorate the research on multi-array disparity estimation.

In multi-array camera systems, there exist multiple stereo pairs: horizontal, vertical, narrow baseline, and wide baseline pairs. The disparity estimates from several stereo pairs show different properties according to the scan-line direction and baseline length. These properties can be exploited to enhance initial array disparities. In addition, local and global fusion of multiple disparities can be used alternately to maximize disparity enhancement.

Corresponding multiple disparity estimates are supposed to be the same along the array dimension. However, they show inconsistency. A cross-filling approach based on a Left-Right Consistency (LRC) check may be considered to locally replace crossly mismatched values with valid ones. However, the LRC using only two views would not be proper to use directly in multiple views.

Most image-restoration methods are used for image-denoising applications. It has been demonstrated that the method can be extended to disparity refinement by reducing disparity outliers. Typically, the image-restoration problem is an inverse and

ill-posed problem. To obtain a meaningful and stable solution to the inverse problem, regularization is required. In conventional regularization problems over multiple dimensions (2D, 3D), simultaneous regularization is always used. A new and different regularization approach can be sought to better restore complicated image structures while alleviating memory issues on the increased dimension.

1.2 Contributions

We propose an accurate and efficient local disparity algorithm for stereo images/videos and a multi-resolution disparity processing technique for large panoramic views. In addition, we propose a disparity enhancement algorithm for multi-array camera systems. Our contributions to these applications are as follows:

1. We propose a three-moded census transform with a noise buffer for reliable similarity measures. The adaptive noise buffer promotes the tolerance of image noise in homogeneous areas. We investigate the cross-square census to obtain more spatial information while reducing exposure to the occlusion area.

2. To model an advanced support weight window, we define conditional and correlated relations between Gestalt principles. For video disparity, we propose incorporating optical flow in order to reduce spatial ambiguities by utilizing temporally consistent information. The proposed support weight computation improves disparity quality near moving object boundaries.

3. We propose an effective method of multi-resolution depth processing and fusion for large panoramic images. The proposed method consists of three main functions: the adaptive pixel-wise disparity search range based on local structures of both the image and disparity map, reliable multiple fitting for sub-pixel disparity, and spatial-multi-resolution TV to enforce scaling consistency as well as spatial consistency.

4. We investigate the fusion effect of multi-scale disparity estimates. The adaptive disparity search range propagates desirable initial estimates into the high-scale

direction, and the spatial-multi-resolution TV fuses multi-scale disparity results by combining their complementary information.

5. For multi-array camera systems, we investigate the advantage of a new cascade regularization approach for image restoration, including disparity refinement. It is proved by showing both detailed analyses and simulation results. We also demonstrate that the proposed cascade regularization approach and multiple cross-filling method can be used for multi-array camera disparity enhancement as global and local optimization methods, respectively. They are used alternately to maximize disparity enhancement.

6. We create synthetic multi-array images and videos with associated ground-truth disparity maps so that other researchers can use them for performance comparisons. The dataset is available on the website - http://videoprocessing.ucsd.edu/~zucheul/multi-array.html.

1.3 Organization

The organization of this dissertation is as follows:

Chapter 2 introduces background materials on computational stereo matching. We start with fundamental matching potential for two view geometry, followed by similarity measures, Gestalt theory, support weight window, and disparity conversion from 3D rendering software.

Chapter 3 presents an efficient disparity algorithm that is used for stereo images and is able to be extended for stereo videos and large panoramic views. A novel three-moded cross census and combination metric of three similarity measures are presented. An advanced support weight and an occlusion filling algorithm are proposed. Finally, comparison results with other methods are provided.

Chapter 4 presents a video disparity algorithm that incorporates optical flow to alleviate the spatial ambiguity problem by using temporally consistent information. Spatial consistency and temporal consistency for video disparities are discussed, and quantitative and qualitative evaluations are performed.

Chapter 5 presents a multi-resolution depth processing and fusion algorithm that is effective for large panoramic views and promotes the fusion effect of multi-scale disparity estimates. The adaptive pixel-wise disparity search range and reliable sub-pixel algorithm are introduced. Moreover, the spatial-multi-resolution TV algorithm for scaling consistency is adapted. Quantitative and qualitative simulation results are then demonstrated.

Chapter 6 discusses the difference and advantage of a new cascade regularization approach against conventional simultaneous approaches. Detailed analyses and simulation results demonstrate that the cascade approach achieves better restoration of complicated image structures than conventional approaches. A local multiple cross-filling method extended from the traditional region voting technique is presented. It is demonstrated that the proposed twofold disparity enhancement algorithm for multi-array cameras achieves a performance gain of 65% compared to initial disparities.

Chapter 7 concludes with a brief summary and remarks, and future work is discussed.

Chapter 2

Background

This chapter provides the necessary background knowledge on stereo correspondence matching.

2.1 Computational Stereo Matching

A computational visual system is composed of two cameras and a computer, somewhat like the human visual system, which consists of two eyes and a brain. Slightly displaced cameras replace the human eyes, taking two views (left and right). We consider two view geometry with two calibrated cameras, as shown in Fig. 2.1. The left camera and the right camera are centered at o_1 and o_2, respectively. P_L and P_R are the left and right image planes respectively, where a 3D point $\boldsymbol{p}$ is projected. The exact matching point of $\boldsymbol{x_1}$ projected on the left plane becomes $\boldsymbol{x_2}$ on the right plane, and $\boldsymbol{x_i}$ is one of the correspondence candidates. The line passing through the two centers is called the baseline. Two vectors, $\boldsymbol{e_1}$ and $\boldsymbol{e_2}$, are the epipoles representing the points where the baseline passes through the image planes. The lines $\boldsymbol{l_1}$ and $\boldsymbol{l_2}$ are called the epipolar lines. The plane spanned by o_1, o_2, and p is the epipolar plane. The matrix R is the rotation matrix and the vector $\boldsymbol{T}$ is the translation vector. We find a relation between point vector $\boldsymbol{x_1}$ and its corresponding point vector $\boldsymbol{x_2}$ as

$$\lambda_2 \boldsymbol{x_2} = \lambda_1 R \boldsymbol{x_1} + \boldsymbol{T} \tag{2.1}$$

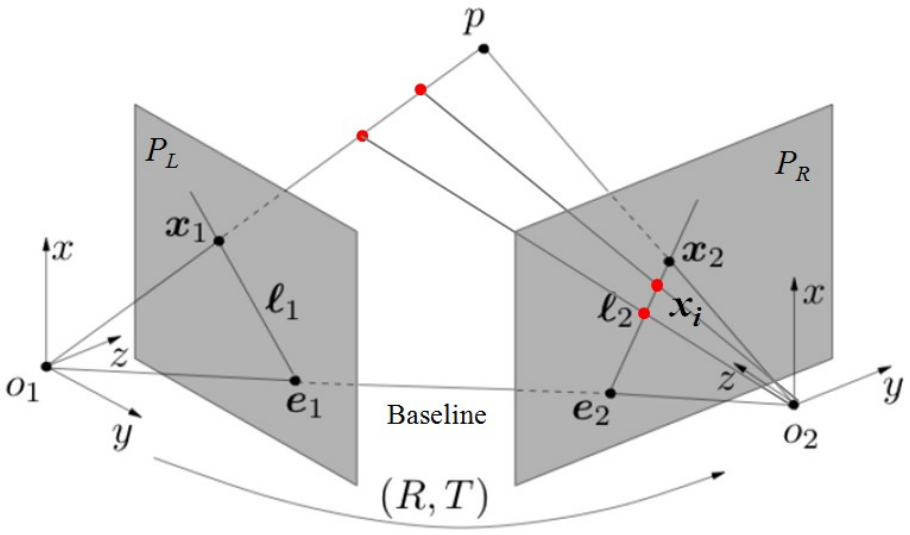

Figure 2.1: Two view geometry.

where λ_1 and λ_2 are unknown depth values [5]. We take the cross product of both sides with $\boldsymbol{T}$ and then take the inner product with $\boldsymbol{x_2}$. The relationship between two points is given as

$$\boldsymbol{x_2}^T \hat{\boldsymbol{T}} R \boldsymbol{x_1} = \boldsymbol{x_2}^T E \boldsymbol{x_1} = 0 \tag{2.2}$$

where $\boldsymbol{x}^T$ denotes the vector transpose of $\boldsymbol{x}$ and $\hat{\boldsymbol{T}}$ is the matrix representation of the cross product with $\boldsymbol{T}$ [5]. E is called the essential matrix, which encodes the relative camera pose. Equation (2.2) is the epipolar constraint, which indicates that if a point is observed on one image plane, its position on the other image plane is constrained to lie on the line defined as the epipolar line. This is very useful, because it can reduce the correspondence matching problem to a 1D search instead of a 2D search. To further simplify the matching problem, we rectify the stereo images so that the epipolar lines are horizontal and thus points in one image plane map to the horizontal scan line with the same y coordinate on the other image plane. To rectify two images, a transformation process to project them onto a common plane is necessary [6].

Stereo matching (correspondence matching or disparity estimation) refers to finding the pair ($\boldsymbol{x_1} = \boldsymbol{x_L}, \boldsymbol{x_2} = \boldsymbol{x_R}$) of projections of the same 3D point. Once the corresponding point is found, we can compute the disparity ($x_L - x_R$), which is the difference in the x-coordinates of the corresponding points. Disparity and depth are in-

versely proportional. However, disparity is used synonymously with depth since depth can be directly converted from disparity, given camera parameters. Disparity estimation methods are typically categorized as either local or global. Global methods compute all disparities of the image simultaneously by optimizing the global energy function. They produce accurate disparity maps, but they are usually complicated and computationally expensive. On the other hand, local methods calculate the disparity of each pixel based on window cost aggregation. They have a simpler structure and show better processing efficiency in terms of computational complexity compared to global methods. We will discuss how to improve disparity accuracy with local methods while preserving simplicity. Two essential parts of local methods are similarity (cost) measures and support weight (aggregation) window selection, which will be introduced in the following section.

2.2 Similarity Measures

Similarity measures also called cost functions calculate how correlated the corresponding pixels or windows are. The lowest cost implies the best match, and the highest cost implies the worst match. In general, window (block)-based similarity measures are used rather than pixel-based measures due to the robustness of the former. The well-known window similarity measures are as follows.

- Sum of Absolute Differences (SAD):

$$\sum_{p \in W_L, p_d \in W_R} |I_L(p) - I_R(p_d)| \tag{2.3}$$

  where W_L is the left window and $I_L(p)$ denotes the intensity at pixel p in the left image. p_d in the right window is the corresponding pixel of p in the left window.

- Sum of Squared Differences (SSD)

$$\sum_{p \in W_L, p_d \in W_R} (I_L(p) - I_R(p_d))^2 \tag{2.4}$$

- Normalized Cross Correlation (NCC):

$$\frac{\sum_{p \in W_L, p_d \in W_R} I_L(p) I_R(p_d)}{\sqrt{\sum_{p \in W_L} I_L(p)^2 \sum_{p_d \in W_R} I_R(p_d)^2}} \tag{2.5}$$

- The census transform [7] encodes the pixel intensity into a bitstream as

$$T = \bigotimes_{p \in W} \xi(I_c, I_p)$$
$$\xi(x, y) = \begin{cases} 1 & \text{if } y > x \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

  where $\bigotimes$ denotes concatenation and I_c represents the center pixel intensity in the window. W denotes the census window.

- The rank transform [7] is similar to the census transform, being defined as the number of pixels in the local window whose intensity is less than that of the center pixel.

The similarity (correlation or cost) measure is an important part of local stereo matching. Choosing an appropriate similarity measure leads to more reliable initial disparity estimates. We will discuss this in detail in Chapter 3.

2.3 Gestalt Theory

According to Gestalt principles, human observers are able to group visual objects that share certain common characteristics [8], as shown in Fig. 2.2. The best-known grouping laws are proximity (objects that are close to each other are grouped together), similarity (objects that have similar color are grouped together), and common fate (objects that move at the same speed in the same direction are grouped together) [9]. Common fate is closely related to motion flow, which can be denoted as "motion" for simplicity. Whenever objects have characteristics in common, they are grouped and formed into a larger visual object, known as a gestalt [8].

From these observations, we can assume that human observers can group pixels in a scene based on how close two pixels are spatially, how similar their colors are, and how similar their velocities are. Thus, we can define the strength of grouping, which should be proportional to the probability that two pixels have the same disparity: the closer two pixels are in proximity, color, and motion flow, the larger their strength of grouping. These three observations may be treated in an integrated manner to obtain

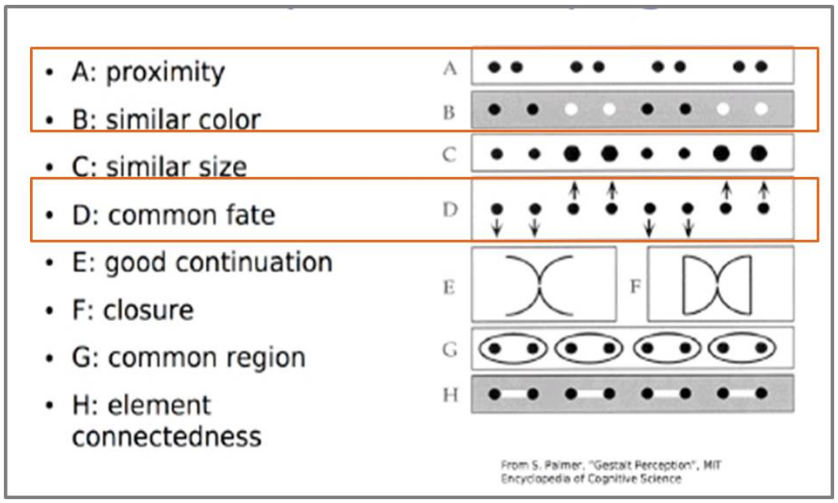

Figure 2.2: Gestalt principles.

better grouping. Each grouping law can compensate for the others when they fail in specific cases. For instance, the motion cue helps viewers distinguish figures when the object color or outlines are not clear. Therefore, it would be beneficial to model the human visual system and segment objects by using Gestalt principles in an integrated way. We analyze each principle and their relationships to find an effective integration method for stereo images and video.

2.4 Support Weight Window

Support window selection is the crux of local stereo matching. In local stereo matching, we assume that all pixels in the support window have the same disparity, which is known as the smoothness assumption. This assumption always fails at object boundaries, causing smearing artifacts. Optimal support windows are small enough to avoid crossing depth discontinuities and large enough to include sufficient intensity variation in homogeneous areas for reliable matching [10]. In addition, they have an arbitrary shape rather than a fixed shape, such as a rectangle. To satisfy these two requirements, Local Adaptive Support Weight (LASW) [3] assigns an adaptive weight to each pixel in the fixed support window based on the color difference and spatial distance with respect to the center pixel. The idea originates from the Gestalt principles

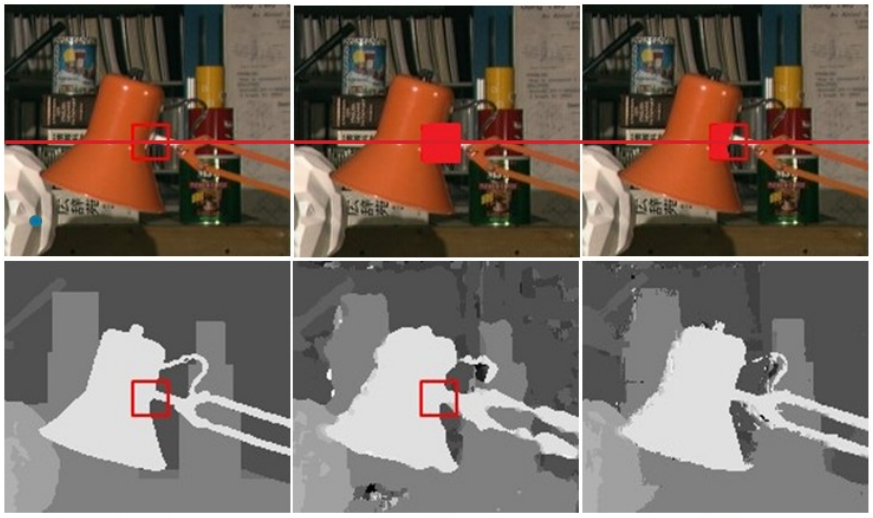

Figure 2.3: Comparison of two support windows. (a) Ground-truth disparity map. (b) Disparity map from the rectangular window. (c) Disparity map from the adaptive support weight window.

of color similarity and proximity. Assigning different (adaptive) weights to each pixel in the support window creates an effect similar to changing the support window size and shape adaptively. Fig. 2.3 illustrates a comparison of two support window types. The support weight window based on Gestalt principles has an arbitrary shape according to the object boundary, as shown in Fig. 2.3(c). It illustrates that adaptively changing the support weight window size and shape improves overall disparity estimation and reduces the smearing effect.

Once the similarity measure and support weight are calculated, the final matching cost is aggregated by taking the weighted average of similarity costs within a support window.

2.5 Disparity Conversion from 3D Rendering Software

Array image/video datasets are not available for simulation. Moreover, there are no available array disparity ground-truth maps, whereas various stereo disparity ground-truth maps are provided on the Middlebury benchmarking site [11]. 3DS MAX is a 3D modeling and rendering software program that is able to render a synthetic image in a

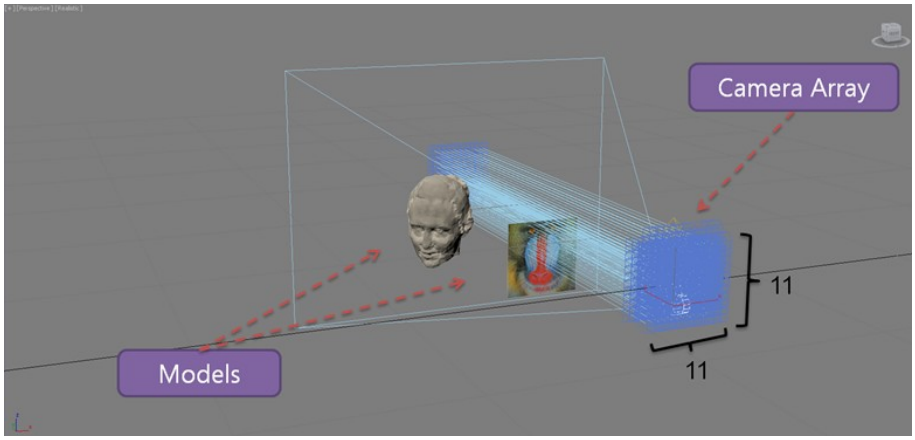

Figure 2.4: Synthesis environment of 3DS MAX.

virtual environment [12].

The synthesized scene is composed of 3D meshes with polygon modeling. Virtual array cameras can be placed in arbitrary positions, as shown in Fig. 2.4. A pair of array color images and Z-depth maps is produced by 3DS MAX. The Z-depth (Z_{depth}) map cannot be directly used as a ground-truth disparity map, since it is encoded by 3DS MAX as

$$Z_{depth} = \frac{Z_{max} - Z_{obj}}{Z_{max} - Z_{min}} \times 255 \tag{2.7}$$

where Z_{max} and Z_{min} are the maximum and minimum distances (mm) in the rendering process, respectively. Z_{obj} is the real distance between the object and the camera. 3DS MAX provides two parameter values: focal length (F_m) and aperture width (A_m) (in mm). However, to obtain a disparity value in pixel units from the real distance (Z_{obj}), the focal length (F_p) in pixel units is required. It can be calculated by

$$F_p = \frac{W_p F_m}{A_m} \tag{2.8}$$

where W_p is the image width in pixel units. The disparity value (d) is converted from the real distance (Z_{obj}) as

$$d = \frac{BF_p}{Z_{obj}} \tag{2.9}$$

where B is the baseline (in mm) between two cameras. Finally, the ground-truth disparity map can be obtained from the Z-depth map encoded by 3DS MAX.

Chapter 3

Disparity Estimation for Stereo Image

This chapter presents an efficient local disparity algorithm that can be directly applied to stereo images. It can be effectively extended for stereo videos and large panoramic views.

We propose a three-moded census transform with a noise buffer to increase tolerance of image noise in homogeneous areas and a cross-square census to increase the reliability of the census measure. We investigate the effect of a combination metric of three cost measures (census, color, and gradient) that have different characteristics on stereo matching. The combination metric is able to obtain a more accurate cost measure in a variety of image regions. These three new ideas can be utilized in both stereo images and videos in the same form. To further improve the original support weight, we define the conditional relation between similarity and proximity by analyzing the Gestalt principles. Simulation results show that the proposed local disparity method is the best-performing local method on the Middlebury stereo benchmark test [11].

This chapter is organized as follows. We discuss related works in Section 3.1. The proposed disparity algorithm is presented in Section 3.2 where we show the system structure, the new similarity measure, and details of the advanced support weight for stereo images. The disparity computation algorithm and occlusion-filling algorithm are then presented. Section 3.3 shows experimental results and discusses their significance. Section 3.4 summarizes the proposed local disparity method.

3.1 Related Work

Two main concerns associated with the local disparity method are the accuracy of the similarity measure and the proper choice of support window. Matching accuracy depends on these two factors.

Common similarity measures are SAD, SSD, NCC, and non-parametric transforms such as rank and census, as introduced in Section 2.2. The rank and census transforms are robust to radiometric distortion, because they yield relative ordering of the pixel intensity rather than the intensity values themselves. Therefore, for image regions with similar colors, non-parametric transforms may cope better with the chromatic matching ambiguities using structural information, while for image regions with similar local structures, the color differences (SAD and SSD) may cope better with the structural matching ambiguities. According to the evaluation of similarity measures [13], the census transform achieves the best overall performance throughout all experiments with simulated and real radiometric differences, except in the presence of strong image noise.

Another important research topic in the area of local methods is how to select the proper support window for each pixel. In early local approaches, a simple rectangular window with a fixed window size is used to find corresponding pixels in a pair of left and right images. However, this results in the foreground smearing problem near depth discontinuities due to the assumption that all pixels in the window have the same disparity. To solve this problem, the adaptive-window method [10] finds an optimal window based on the local variation of intensity and disparity. This method still uses a rectangular window, which is not suitable for arbitrarily shaped depth discontinuities. The multiple-window method [14] calculates the correlation with nine pre-defined windows and selects the disparity with the smallest matching cost. This method also has the limitation of window shape. To obtain more accurate results at depth discontinuities, the LASW approach [3] adjusts the support weights of the pixels in the window by using the photometric and geometric distance with respect to the center pixel. This method deals with the pixels near depth discontinuities more effectively than the two methods mentioned above. Segment support [15] improves the reliability of adaptive support aggregation by adding an additional segmentation process. Disparity calibration [16] increases the matching process to two steps by adding disparity calibration,

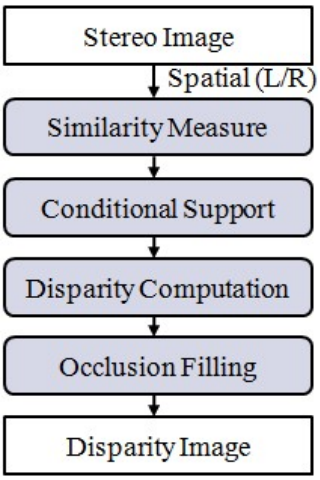

Figure 3.1: Block diagram of the proposed disparity estimation method.

while the traditional local methods use a one-step process. PatchMatch [17], the best local method of all the local methods on the Middlebury benchmark test [11], uses additional processes such as iteration, slanted plane, and propagation schemes to obtain better results. However, these three methods are computationally expensive. CostFilter [4], which is one of the best local methods, obtains consistent edge-preserving results by using a guided filter. It is worth noting that LASW and CostFilter do not use any iteration or additional step that could make the algorithm more complex. LASW and CostFilter are good edge-preserving methods, but they do not provide a reliable solution for disparity estimation in textureless areas that have different characteristics from the edges.

3.2 Proposed Method

The proposed method is an efficient one-pass local method applicable to both stereo images and videos with no iteration. The main goal is to improve the accuracy of the similarity measure and enhance the support weight function in order to achieve a high-quality disparity map. The block diagram of the overall system is shown in Fig. 3.1. It consists of four main components: a similarity measure, support weight, disparity computation, and occlusion filling. The core blocks for the accuracy of disparity estimation are the similarity measure and support weight block, which will be discussed

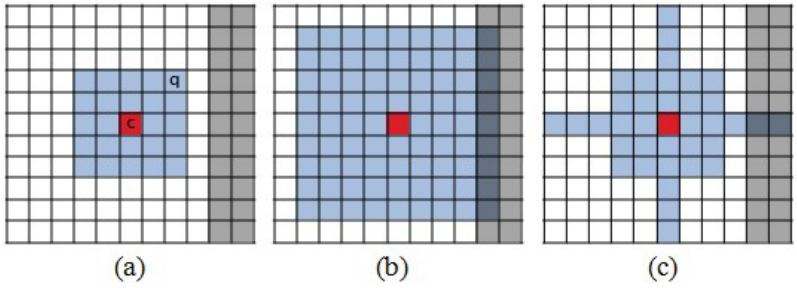

Figure 3.2: Three census windows. (a) 5×5. (b) 9×9. (c) Cross-square.

in detail.

3.2.1 Three-moded cross census and combination metric of similarity measures

The census transform encodes the pixel value into the bitstream representing the relative ordering of the neighboring pixels. To achieve a more precise census similarity measure, we need to obtain a larger spatial structure by increasing the size of the census window. However, the error probability might increase as the window size increases over a certain value. The larger the census window size is, the more likely occluded pixels are to be included in the transformed bitstream. There is a trade-off between the amount of spatial information and the accuracy of the estimate. Fig. 3.2 illustrates that the large square census window in Fig. 3.2(b) is more likely to be affected by the occlusion area (gray colored area) than the window in Fig. 3.2(a), and therefore its transformed information will be severely distorted. To alleviate this problem, we propose the cross-square shape census window, which can contain more spatial information while being less exposed to the occlusion area, as shown in Fig. 3.2(c).

The census transform is robust to radiometric distortions, and it achieves the best overall performance in both local and global methods. However, it experiences difficulties in finding the correct correspondences in homogeneous areas, as most methods do. This difficulty is due to the fact that the census matching cost is extremely sensitive to even small image noise in homogeneous areas, since all pixels have similar intensity values, and then the left and right census can be encoded differently due to the noises.

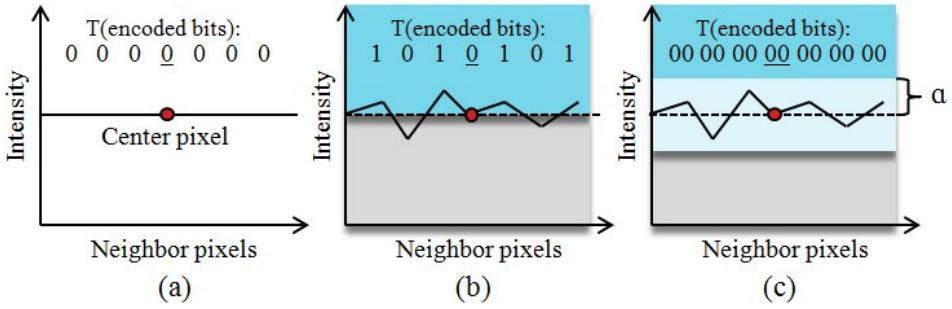

Figure 3.3: Comparison of the original census and three-moded census in homogeneous areas. (a) Original census without noise. (b) Original census with noise. (c) Three-moded census with noise.

In practice, most stereo images are distorted due to camera noises except for synthetic stereo images. To reduce the mismatch due to the distortion from the left and right cameras, we propose a three-moded census transform with a noise buffer. The original census has two modes where a bit is set to 1 if the neighboring pixel in the census window has a higher intensity than the center pixel and 0 otherwise. On the other hand, our three-moded census uses two bits to implement three modes, and it is defined as

$$T = \bigotimes_{q \in W} \xi(I_c, I_q)$$

$$\xi(x, y) = \begin{cases} 10 & \text{if } y > x + \alpha \\ 01 & \text{if } y < x - \alpha \\ 00 & \text{otherwise} \end{cases} \tag{3.1}$$

where $\bigotimes$ denotes concatenation and W represents the census window. I_c represents the intensity at the center pixel c, and α is the noise buffer threshold. Camera noise is intensity-dependent, and the noise variance is proportional to intensity [18, 19]. The noise buffer should be increased to get consistent results as the noise variance increases. Therefore, we can define α as a function of intensity:

$$\alpha = [\frac{I_c}{\beta}] \tag{3.2}$$

where $[\cdot]$ denotes the nearest integer operator, and empirically reasonable values for β are 500 and 50 for synthetic and real-world images, respectively.

Fig. 3.3 shows how the three-moded census works under a noisy environment. In homogeneous areas, the neighboring pixels show the same intensity as shown in

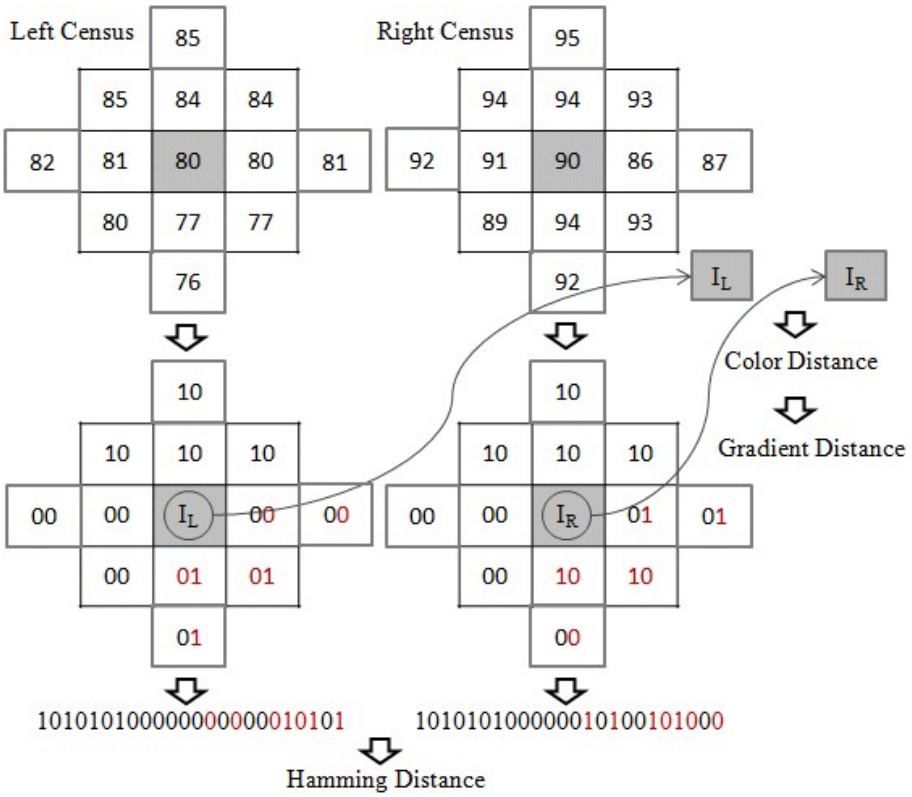

Figure 3.4: Example of three-moded census transform with $\alpha = 2$ and three similarity measures.

Fig. 3.3(a). Under a noisy environment, the original census transform yields a very different bitstream from the noiseless case, as shown in Fig. 3.3(b), while the three-moded census transform produces a consistent bitstream, as shown in Fig. 3.3(c). Note that we do not define the census transform at the center pixel because it is always 0.

Fig. 3.4 shows examples of the left and right bitstreams resulting from the three-moded cross census transform, which are used in the calculation of Hamming distance (ΔH). To further improve the matching accuracy, we incorporate the color distance (ΔI) and gradient distance (ΔG) between the two center pixels, as shown in Fig. 3.4. In other words, we use the census transform to compare the spatial structure of two census windows, while we use the color and gradient distance to compare the two center pixels. The Hamming distance of two census transforms is defined as

$$\Delta H = d(T_L, T_R) = T_L \oplus T_R \tag{3.3}$$

where T_L represents the left transformed bitstream and $\oplus$ denotes the bitwise XOR

Figure 3.5: Disparity maps on "Laundry" computed by different similarity measures. (a) Left image. (b) Color. (c) Combination of color and census. (d) Combination of color, census, and gradient.

operation. The color distance between $I_L = (I_L^r, I_L^g, I_L^b)$ and $I_R = (I_R^r, I_R^g, I_R^b)$ in the RGB vector space is defined as

$$\Delta I = d(I_L, I_R) = \sqrt{\sum_{j=r,g,b} (I_L^j - I_R^j)^2}. \tag{3.4}$$

The gradient $G = (G_x, G_y)$ is composed of two components, which are partial derivatives along the x-axis and y-axis, respectively. The partial derivative G_x can be expressed as (G_x^r, G_x^g, G_x^b) in the RGB space. The gradient distance between $G_L = (G_{Lx}, G_{Ly})$ and $G_R = (G_{Rx}, G_{Ry})$ is defined as

$$\begin{aligned} \Delta G &= d(G_L, G_R) = \sqrt{d(G_{Lx}, G_{Rx})^2 + d(G_{Ly}, G_{Ry})^2} \\ d(G_{Lx}, G_{Rx}) &= \sqrt{\sum_{j=r,g,b} (G_{Lx}^j - G_{Rx}^j)^2} \end{aligned} \tag{3.5}$$

where G_{Lx}^j represents the partial derivative along the x-axis in the j color domain of the left image.

We propose the combination of three distances, which is simple and very effective. It yields a more reliable similarity measure, as the three distances compensate for one another. Fig. 3.5 illustrates how each similarity measure improves the accuracy of disparity estimation. Fig. 3.5(b) is computed by using color distance, which is commonly used, and it shows many errors in the similar color area (green box). In Fig. 3.5(c), some errors are recovered by combining the Census Hamming distance. However, incorrect matches in densely textured regions with high-frequency conditions (red box) still exist. As shown in Fig. 3.5(d), combining three cost measures leads to

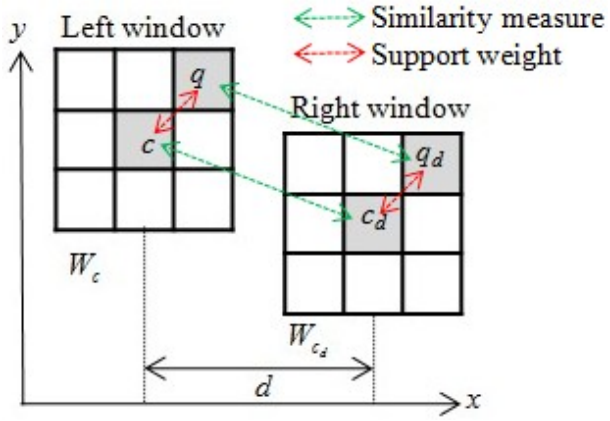

Figure 3.6: Left support window and right support window

the best overall disparity map, recovering different types of errors. Note that there are many similarity measures with different characteristics, and it is important to choose proper measures and integrate them properly to improve matching performance. For the combination metric of three similarity measures, we use a robust cost function including three distances:

$$C_0(q, q_d) = 3 - \exp\left(-\frac{\Delta H_{qq_d}}{\gamma_H}\right) - \exp\left(-\frac{\Delta I_{qq_d}}{\gamma_I}\right) - \exp\left(-\frac{\Delta G_{qq_d}}{\gamma_G}\right) \tag{3.6}$$

where ΔH_{qq_d}, ΔI_{qq_d} and ΔG_{qq_d} are the Hamming distance, color distance, and gradient distance, respectively, between pixel q and pixel q_d as shown in Fig. 3.6. γ_H, γ_I and γ_G are empirical parameters.

3.2.2 Conditional adaptive support weight

The ideal support window is an arbitrarily shaped window that consists of only pixels with the same depth. It is very difficult to accurately determine which pixels belong to the same object. We consider the adaptive support weight window based on two Gestalt grouping laws (color similarity and proximity) that can be used together to group objects as in [3]. To obtain an advanced support weight, we analyze color spaces and the relationship between similarity and proximity, which helps in deciding how to integrate the two properties. For the color space, previous works have used the Euclidean distance in the CIELab color space. The CIELab color space is perceptually uniform,

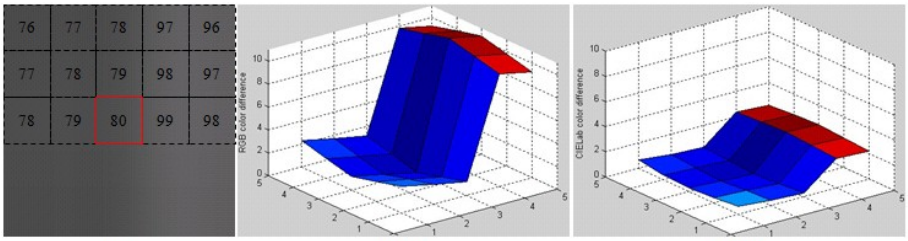

Figure 3.7: Comparison of RGB and CIELab color difference. (a) Support window. (b) RGB color difference. (c) CIELab color difference.

and its Euclidean distance corresponds to the perceptual color difference between two colors. However, the use of the CIELab color space makes the color distance less selective for the pixels, which are chromatically close. Fig. 3.7 illustrates a comparison of two color spaces in the area where each pixel has a chromatically similar color. In Fig. 3.7(a), the center pixel I_c =[80 80 80] and the neighboring pixel I_q =[79 79 79] in the RGB space are converted to [34.029 0.002 -0.004] and [33.603 0.002 -0.004] in the CIELab space, respectively. The RGB and CIELab color differences are 1.7321 and 0.4260, respectively. The ratio ($\frac{\Delta I_{RGB}}{\Delta I_{CIELab}}$) in the similar color region (I_c and I_q) is 4.0657. On the other hand, the ratio in the distinct color region (I_c and I_p=[200 80 80]) is 2.1321. Hence, there is a higher ratio in the similar color region than the distinct color region. Fig. 3.7(b) and (c) show the color difference at each pixel with respect to the center pixel in the RGB and CIELab spaces. The RGB space produces a more selective distance than the CIELab space in a similar color region. Additionally, the $L^*a^*b^*$ metrics are particularly sensitive to errors in low RGB signals [20]. The color space should provide a good distance metric for areas with similar colors as well as with distinct colors. To this end, we use the RGB space for color similarity. The RGB color difference (Δs_{cq}) between the center pixel and the neighboring pixel is calculated as in (3.4). The spatial distance (proximity) is calculated as the Euclidean distance.

The adaptive support weight is based on the strength of grouping by similarity and proximity. The strength of grouping by similarity is defined using a Laplacian kernel as

$$g_{\gamma_s}(\Delta s_{cq}) = \exp\left(-\frac{\Delta s_{cq}}{\gamma_s}\right) \tag{3.7}$$

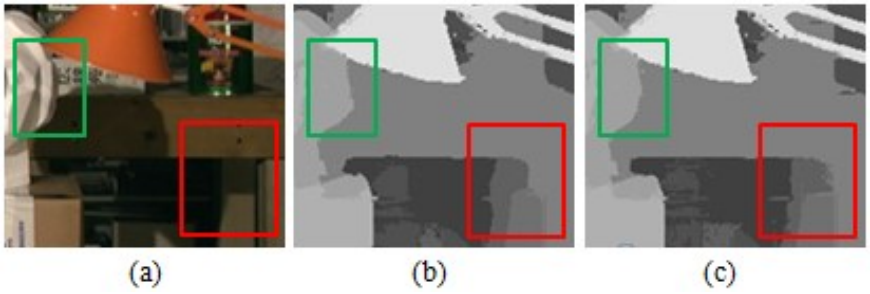

Figure 3.8: Comparison of the original support and proposed conditional support on "Tsukuba". (a) Left image. (b) Original support. (c) Conditional support.

with γ_s being an empirical similarity parameter. The strength of grouping by proximity is defined as it is in (3.7). The weights based on the spatial proximity with respect to the center pixel are constant for every shifted window, while the weights based on the color similarity vary for each shifted window. Hence, the spatially fixed kernel might yield negative consequences in the specific area, such as the disparity discontinuity area with similar colors, because it is blindly aggregated according to the distance and, thereby, it causes incorrect matches near the disparity discontinuity. To alleviate this problem, we suggest the conditional adaptive support weight as

$$w(c,q) = \begin{cases} g_{\gamma_s}(\Delta s_{cq}) & \text{if} \Delta s_{cq} \leq \eta \\ g_{\gamma_s}(\Delta s_{cq}) g_{\gamma_p}(\Delta p_{cq}) & \text{otherwise} \end{cases} \tag{3.8}$$

where Δp_{cq} is the spatial distance between pixel c and pixel q and η is a color difference threshold determining the similarity between two pixels.

Fig. 3.8 depicts the process where the conditional support weight improves the disparity map. Fig. 3.8 shows the left image and two disparity maps. Fig. 3.8(b) depicts the estimates using the original support always including proximity, while Fig. 3.8(c) shows the estimates using the conditional support measure. At the border of the disparity discontinuity area with a similar color in the foreground and background (red box), the spatial proximity kernel may produce many wrong disparities due to the blind aggregation by the close spatial distance, as shown in Fig. 3.8(b). In this case, we exclude the proximity term to avoid the blurring support at the edge of the disparity and exploit only the color similarity to determine the correct support according to even slight color differences. Therefore, our conditional support recovers many errors, as shown in Fig. 3.8(c).

This is precisely the goal of the conditional adaptive support weight in (3.8).

3.2.3 Disparity computation

Once the support weights are calculated, the aggregated cost is computed by aggregating the raw similarity measures, scaled by the support weights in the window. If we consider only the left support window, the cost computation may be erroneous, since the right support window may have pixels from different disparity levels. To reduce such errors, the aggregated cost is computed by combining the weights of both support windows as in [3]. The aggregated matching cost between pixel c and pixel c_d in Fig. 3.6 is given in the weighted mean form:

$$A(c, c_d) = \frac{\sum_{q \in W_c, q_d \in W_{c_d}} w(c, q) w(c_d, q_d) C_0(q, q_d)}{\sum_{q \in W_c, q_d \in W_{c_d}} w(c, q) w(c_d, q_d)} \tag{3.9}$$

where W_c and W_{c_d} represent the left and right support windows, respectively, and the function $w(c_d, q_d)$ is the support weight of pixel q_d in the right window.

After the aggregated matching costs have been computed within the disparity range, the disparity map is constructed by determining the disparity d_p of each pixel p through the Winner-Takes-All (WTA) algorithm:

$$d_p = \underset{d \in S}{\operatorname{argmin}} \; A(c, c_d) \tag{3.10}$$

where S represents the set of all possible disparities.

3.2.4 Occlusion filling

To ensure that both left and right disparities are spatially consistent, we perform a LRC check to detect unreliable pixels (i.e., those having different disparities on the left and right images). Fig. 3.9 illustrates an example of occlusion handling. In Fig. 3.9, for each unreliable pixel (x, y), the cross-based aggregation method [21] generates a neighborhood for $(x+s, y)$, as shown for the yellow region in Fig. 3.9, where $(x+s, y)$ is the left most reliable pixel. The white region indicates the unreliable (occluded) region, the dark gray region is the background, and the light gray region is the foreground. All reliable pixels within the neighborhood vote for the candidate disparity value at

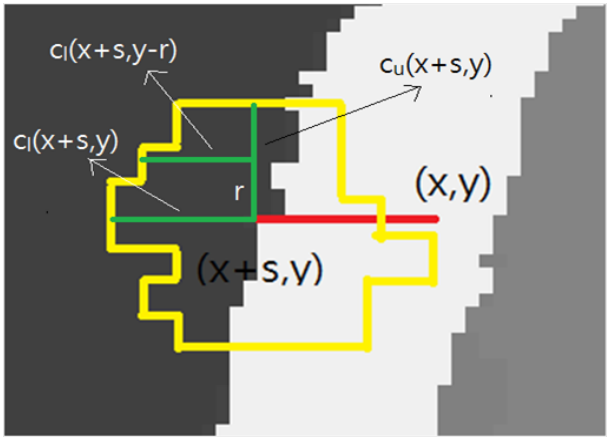

Figure 3.9: Illustration of the occlusion filling process

(x, y). The unreliable pixel at (x, y) is filled with the majority of the reliable pixels in the voting region. By this method, the center pixel is not automatically selected as the center pixel for occlusion handling. Instead, a first non-occluded pixel is selected to define the neighborhood. In Fig. 3.9, a left disparity map is used as an example, where occlusion pixels (white) appear at the right side of the background and the left side of the foreground if the disparity is positive. (In the right image, occlusion pixels would appear at the left side of the background and the right side of the foreground). Only the occlusion pixels are selected and need to be processed. For an arbitrary occlusion pixel (x, y), the method starts at its left neighboring pixel to determine whether it is a non-occluded pixel. If it is occluded, the process continues to the left. If it is non-occluded, the procedure stops. In Fig. 3.9, for the pixel at (x, y), the process goes to the left for s pixels. A neighborhood is constructed based on the cross-based aggregation method on pixel $(x + s, y)$. Every non-occluded pixel within that region votes. The majority of disparity values in that region are assigned to occlusion pixel (x, y). Fig. 3.9 presents an ideal situation where the majority is obviously the background, and consequently, the white region will be filled with the background.

Prior window-based voting methods [22] have been based on (x, y) instead of $(x + s, y)$. The number of non-occlusion pixels in the window constructed based on (x, y) will be significantly smaller than that in the window constructed based on $(x + s, y)$. Therefore, such methods are much more sensitive to outliers due to fewer votes,

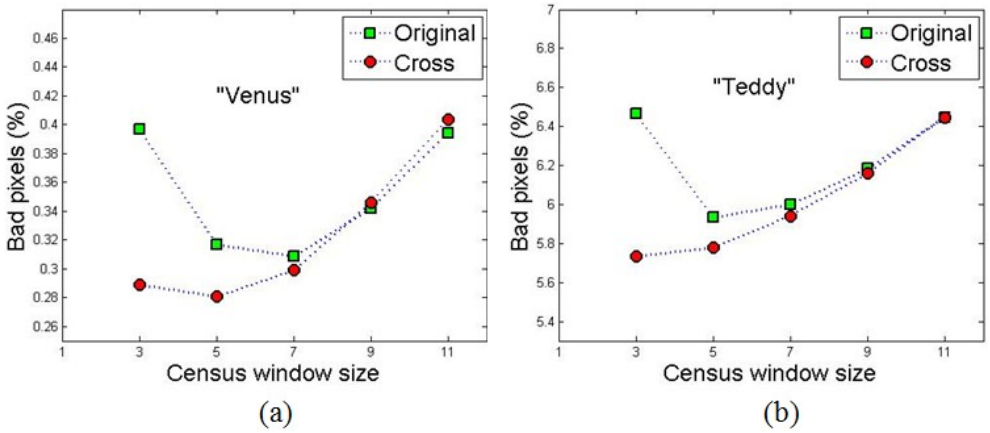

Figure 3.10: Errors (Bad pixels) rate versus census window size. (a) “Venus”. (b) “Teddy”.

and thus yield inaccurate result.

Other methods, such as plane fitting [23] for multiple disparity planes, are very computationally expensive. It is an iterative process that treats the occlusion pixel as outliers and finds the plane that minimizes the error for non-occlusion regions, and fills the occlusion pixel as if it is on the plane. On the other hand, the proposed occlusion method is non-iterative and thus more efficient.

3.3 Results

We perform quantitative and qualitative experiments on the Middlebury datasets in order to verify that the proposed method improves the quality and reliability of disparity estimates. In addition, sensitivity to the parameters is examined.

3.3.1 Cross-square census

To evaluate how the sizes of the original and proposed cross-square census window affect the disparity performance, we use two Middlebury datasets (“Venus” and “Teddy”). As shown in Fig. 3.10(a), the error rate of original census (green) decreases sharply as the window size increases from 3 to 7. That is when the census-transformed data contains more spatial structure information, and therefore the similarity measure is more accurate. However, the error rate increases as the window size increases from 7 to

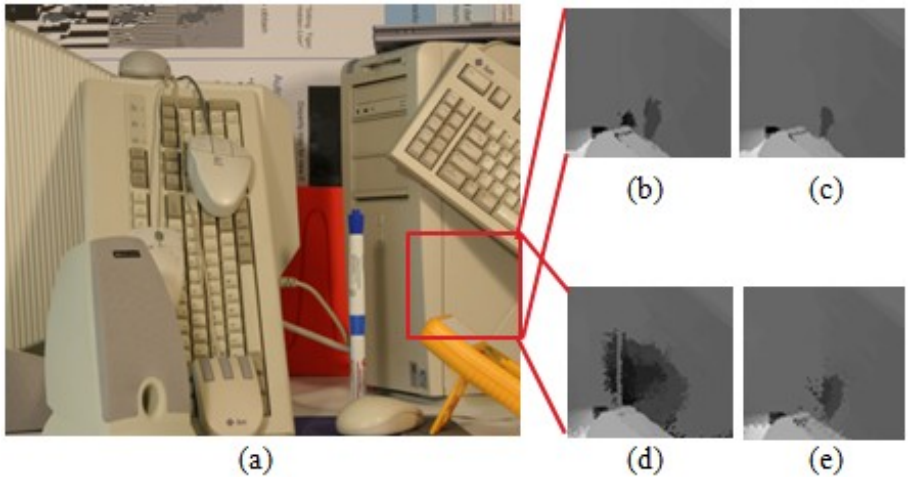

Figure 3.11: Comparison of the original census (2 mode) and the three-moded census with a noise buffer on "Computer". (a) Left image. (b) Original census. (c) Three-moded census. (d) Original census on noise added image. (e) Three-moded census on noise added image.

15. This is due to the fact that a larger census window would include more pixels from occlusion areas as well as more noises, which deteriorate the accuracy of the similarity measure. The cross census window has four wings (up, down, left, and right) as shown in Fig. 3.2, and each wing is composed of three pixels in the experiment. It is worth noting that the proposed cross-square window outperforms the original square one while using a smaller window size, as shown in Fig. 3.10.

3.3.2 Three-moded census transform

We implement the three-moded census transform with a noise buffer for robustness to image noise in homogeneous areas. Fig. 3.11 illustrates that the three-moded census with a noise buffer performs better than the original census in homogeneous areas. To simulate noise in homogeneous areas, we add Gaussian noise, distributed as $\mathcal{N}(0, 10^{-4})$, to the original image. For the noise buffer α, the parameter β is set to 50. First, we perform the experiment on the original image, where the three-moded census reduces some errors in the homogeneous area, as shown in Fig. 3.11(c), compared to the original census. Second, we perform the experiment on the noise-added image. In this case, it is verified that the proposed noise buffer works more effectively, as it reduces

Table 3.1: Performance evaluation of local methods on Middlebury (bad pixel percentage with threshold of 1)

Methods	Rank	Avg.Err.(%)	Err. non-occluded pixels(%)			
			Tsukuba	Venus	Teddy	Cones
Proposed	**13**	5.12	2.10	0.12	5.46	2.12
PatchMatch [17]	15	4.59	2.09	0.21	2.99	2.47
CostFilter [4]	20	5.55	1.51	0.20	6.16	2.71
InfoPermeable	21	5.51	1.06	0.32	5.60	2.65
GeoSup [24]	28	5.80	1.45	0.14	6.88	2.94
AdaptDisCalib [16]	37	6.10	1.19	0.23	7.80	3.62
SegmentSupport [15]	53	6.44	1.25	0.25	8.43	3.77
AdaptWeight [3]	67	6.67	1.38	0.71	7.88	3.97

errors much more, as shown in Fig. 3.11(e), compared to the original census. We also verify that the three-moded census shows better overall performance in terms of bad pixel rate than the original census does.

3.3.3 Quantitative and qualitative evaluation

The performance evaluation is made on the Middlebury datasets with ground-truth disparity maps provided by the benchmark site [11]. The parameters are set to constant values: $\gamma_s = 33$, $\gamma_p = 20$, $\gamma_H = 29$, $\gamma_I = 45$, $\gamma_G = 14$ and $\eta = 3$. The size of the support window is 35×35 (the same size as the LASW [3]), and the size of the cross-square census window is 5×5 for square with three pixels for a wing. Table 3.1 summarizes the quantitative results taken from the Middlebury database for local methods. The bad pixel (error) rate is expressed as

$$B(\%) = \frac{100}{|\Omega|} \sum_{p \in \Omega} I(|D_p - d_p| > \theta) \tag{3.11}$$

where $|\Omega|$ represents the number of pixel in whole image and I denotes the indicator function. D_p represents the true disparity at pixel p and θ represents the bad pixel threshold.

Our method achieves excellent results, ranking 13^{th} out of about 130 methods, and is the best performing local method at the time of the submission. Our method is an efficient one-pass method with no iteration or post-processing. It outperforms the

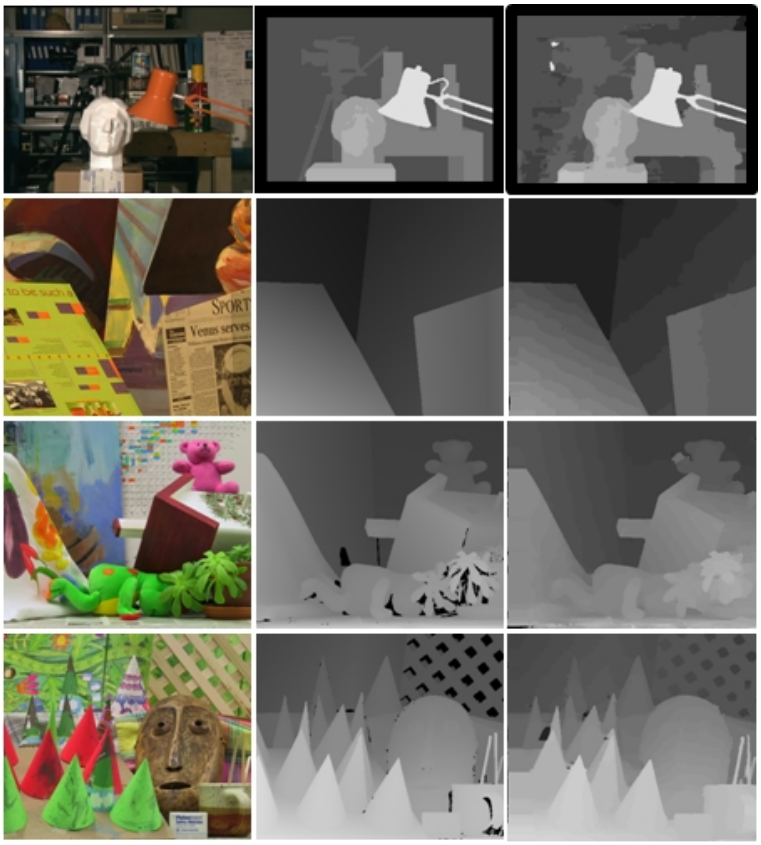

Figure 3.12: Disparity maps for “Tsukuba”, “Venus”, “Teddy” and “Cones”. Centered column shows ground-truth disparity map and right-most column shows the disparity map from the proposed algorithm.

original local method (LASW ranking 67^{th}), using efficient algorithms and structures. Fig. 3.12 shows left images (in the first column), ground-truth disparity maps (in the second column), and our disparity maps (in the third column). The proposed method produces accurate dense disparity maps, as shown in Fig. 3.12. Our method ranks 1^{st} on “Cones” in both non-occlusion and discontinuity areas.

In the proposed method, it takes about 12s to compute the disparity map on “Tsukuba” using a Central Processing Unit (CPU). It has been presented in [25] that the LASW [3] can be adopted into a real-time application by using a Graphics Processing Unit (GPU). Our initial implementation on GPU shows approximately 10 frames/sec on Quarter Video Graphics Array (QVGA)-size video frames.

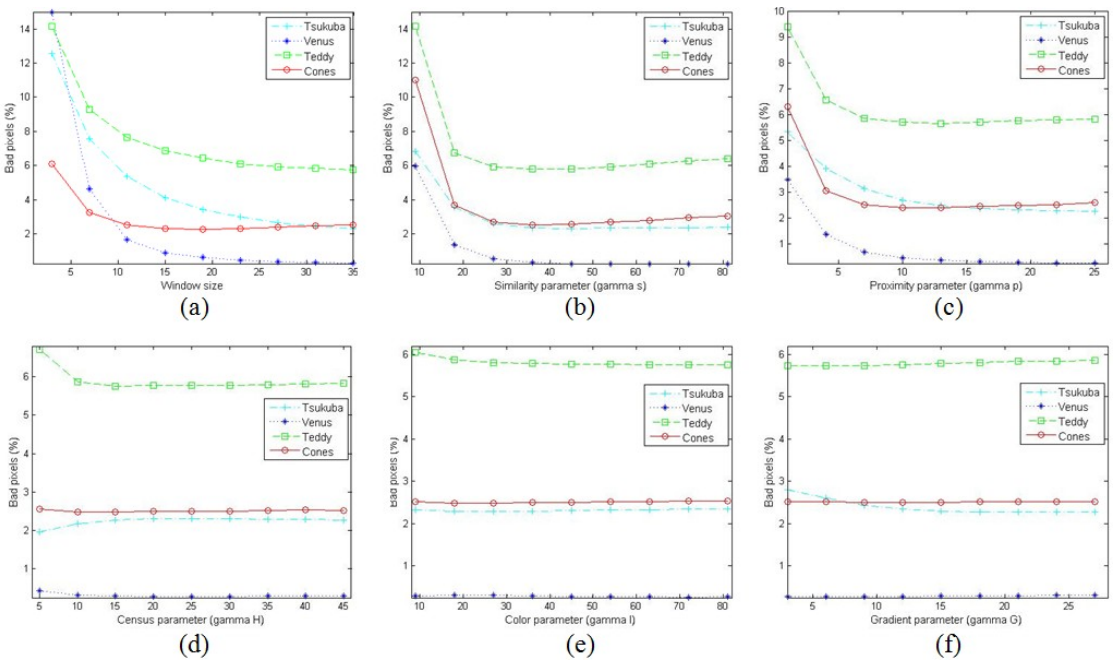

Figure 3.13: Sensitivity to the window size and 5 parameters on four stereo images. (a) Changing the window size. (b) Changing γ_s. (c) Changing γ_p. (d) Changing γ_H. (e) Changing γ_I. (f) Changing γ_G while the other parameters are kept constant.

3.3.4 Sensitivity to the parameters

The robustness of the proposed method when changing the parameters is examined. Fig. 3.13(a) shows the performance evaluation of different support window size for four Middlebury stereo images. It illustrates that the proposed method is fairly insensitive to the support window size when the size is larger than 15×15. This implies that the advanced support weight can segment the same depth region well, and thus that outliers do not increase even though the window size increases. Fig. 3.13(b) and (c) show the performance according to changing the similarity parameter (γ_s) and the proximity parameter (γ_p). They also illustrate that the proposed method is robust to the different parameter setting when they are larger than a certain value. As shown in Fig. 3.13(d), (e), and (f), the performance is also insensitive to the three cost measure parameters (γ_H, γ_I, and γ_G, respectively). Consequently, the five parameter values are not critical for the performance of the proposed method since they are used in the effectively integrated form, as in (3.6) and (3.8).

3.4 Summary

In the local stereo matching, the accuracy of disparity estimate depends on the similarity measure and the support weight computation. We propose a novel three-moded census with a noise buffer to increase robustness to image noise in homogeneous areas and investigate cross-square census for the accuracy of similarity measure. It is verified that RGB color space provides a better distance metric at object boundaries with similar color than CIELab space does. We demonstrate that the combination of three similarity measures yields more reliable cost measures in a variety of image regions. To obtain a more precise support weight and avoid the blind aggregation, a conditional support model is introduced. Simulation results verify that the proposed method outperforms the other state-of-the-art local methods. Moreover, the proposed method is not sensitive to the parameter values.

3.5 Acknowledgements

This chapter is in part a reprint of a published paper in IEEE Transactions on Multimedia, 2013.

Chapter 4

Disparity Estimation for Stereo Video

This chapter is an extension of the previous chapter for video disparity estimation. Stereo video disparity estimation is at an early stage while stereo image disparity estimation is mature. This is the consequence of two factors. First, it is due to lack of stereo video datasets with ground-truth disparity maps. Second, it is due to temporal inconsistency problems, such as flickering, resulting from the simple application of current state-of-the-art image-based algorithms to video frames. In video processing, motion is a crucial factor and, generally, moving objects tend to have a higher degree of saliency. However, most disparity methods may have difficulty (ambiguity) in dealing with fast moving edges in video scenes.

We incorporate optical flow for enhanced support weight computation within the localized window. This approach is the first use and helps to determine the spatial ambiguities by utilizing temporally consistent information. We define the correlated relation between similarity and motion by analyzing Gestalt principles. We enforce temporal consistency by refining our video disparity estimates with the spatio-temporal consistency algorithm described in [1]. Meaningful results are achieved by incorporating optical flow near moving edges.

This chapter is organized as follows. We discuss related works in Section 4.1. The proposed video disparity algorithm is presented in Section 4.2 in detail. Section 4.3 shows simulation results and discusses their significance. Finally, section 4.4 summarizes the proposed video disparity method.

4.1 Related Work

To solve the temporal inconsistency problem in video disparity estimation, [26] uses median filtering along flow vectors computed by the Horn and Schunck method [27]. However, the results are of moderate quality. The spatio-temporal method [28] adds time as an extra dimension so that matching cost is aggregated over a 3D spatio-temporal support window. It can integrate temporal coherence but requires additional memory. The spatial-temporal TV method [1] shows impressive results by treating the video disparity as a spatio-temporal volume to improve spatial and temporal consistency. Moreover, it presents the possibility of directly extending current image-based disparity algorithms to the video domain.

4.2 Proposed Method

Fig. 4.1 shows the block diagram of the proposed video disparity estimation method. Optical flow and TV refinement algorithm are added to promote disambiguation near moving edges and reduce temporal noise, respectively. Correlated support weight incorporating optical flow replaces the conditional support weight used in Chapter 3.

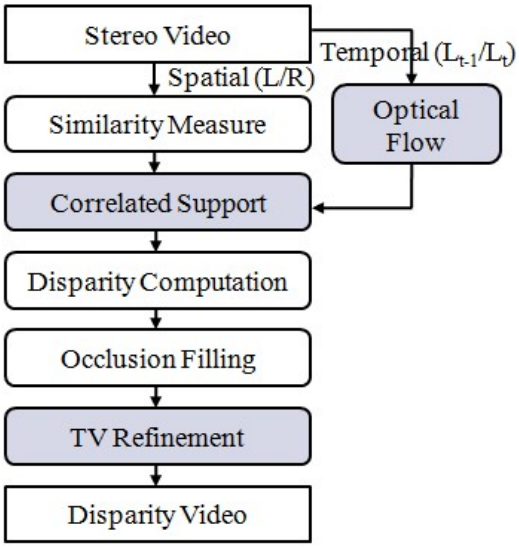

Figure 4.1: Block diagram of the proposed video disparity estimation method.

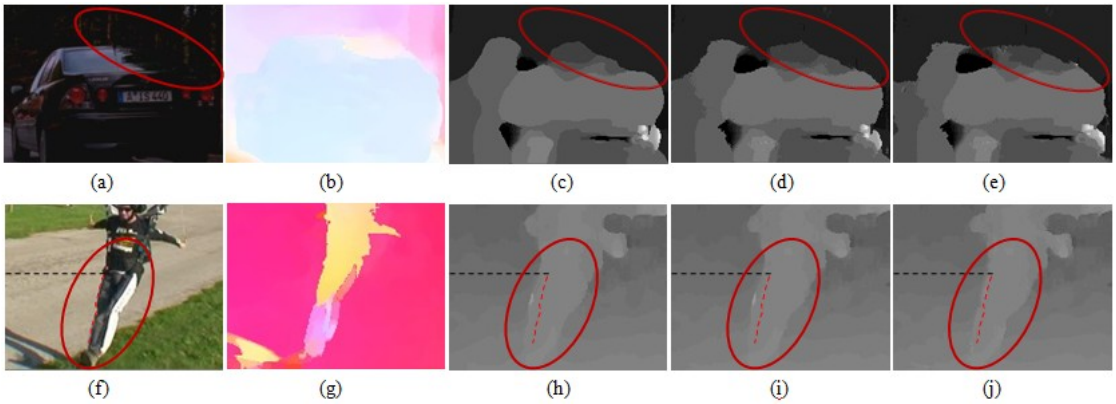

Figure 4.2: Disparity maps for "Car" in the upper row and "Skydiving" in the lower row. (a) and (f) Left view. (b) and (g) Optical flow. (c) and (h) using only proximity. (d) and (i) using proximity and similarity. (e) and (j) using proximity, similarity, and motion.

4.2.1 Benefits of a motion cue

Although motion is a key factor in video processing, it has not been investigated for support weight computation within a localized window. Fig. 4.2 illustrates the benefits of using motion cues. For the evaluation, we use the LASW method [3], in which proximity and similarity are exploited in an independent manner. We extend it to examine how the motion term affects video disparity quality. As the local methods require pixel-based computation, we use the classic optical flow method with a weighted non-local term [29], which is one of state-of-the-art optical flow methods. We use the motion information in the independently integrated support weight form. The "car" and "skydiving" video frames are processed at a resolution of 480×270 and 480×276, respectively. The parameters used are fixed throughout the experiment. In Fig. 4.2, the selected left view (Fig. 4.2(a) and (f)) and its optical flow (Fig. 4.2(b) and (g)) are shown. Fig. 4.2(c) and (h) are obtained by using only the proximity term for the support weight; Fig. 4.2(d) and (i) are computed by adding the similarity term; and Fig. 4.2(e) and (j) are obtained by adding the motion term. As shown in Fig. 4.2(a), it is challenging to discover the outline of the car since it is visually highly ambiguous. In Fig. 4.2(c), many errors are observed at the edges of the moving car (red circle). In Fig. 4.2(d), some errors are recovered by using the color cue; however, edges are not preserved. In Fig. 4.2(e), incorporating the motion term preserves the edges, even though

they are visually ambiguous. We believe that this is due to the preserved background flow as shown in Fig. 4.2(b). Although there is ambiguity in the stereo correspondence, motion between a pair of successive video frames is much more consistent, especially in a localized window in background regions. There are large forward motions in the "skydiving" video, as shown in Fig. 4.2(g). Generally, moving objects tend to have a higher degree of saliency, and viewers will fixate on the skydiver's rapid fall forward as shown in Fig. 4.2(f). Therefore, accurate disparity estimation is required at these moving edges. The dotted red line in Fig. 4.2(f) represents the real moving edge of the skydiver. In Fig. 4.2(h) and (i), a large smearing problem is observed at the edges, while in Fig. 4.2(j), the problem is much alleviated by incorporating the motion cue for the improved support weight. Note that the left edge of the foreground is compared since the occlusion appears at the right side of the foreground due to the negative disparity in Fig. 4.2(f). Disparity is estimating spatial correspondences, while motion estimates temporal correspondences, so the additional temporal information promotes spatial disambiguation. Consequently, the results in Fig. 4.2 imply that the support weight integrating the motion cue yields disparity estimates that are more accurate, especially near the edges of moving objects.

4.2.2 Correlated adaptive support weight

The effectiveness of using motion cue for support weight computation was verified in the previous section. The conditional relation between similarity and proximity has been defined in Section 3.2.2. We discuss issues on motion flow estimation and then analyze the relationship between similarity and motion in order to investigate how the motion term should be integrated. The motion difference between two pixels is calculated by using a measure of optical flow. There are two types of motion difference computation: absolute flow Endpoint Difference (ED) and Angular Difference (AD) [30]. We use ED because AD penalizes errors in larger flows less than errors in small ones [30], which is undesirable. Let $m_c = (u_c, v_c)$ and $m_q = (u_q, v_q)$ be the flow vectors of pixel c and pixel q, respectively. We suggest the truncated motion difference:

$$\Delta m_{cq} = min\left(\sqrt{(u_c - u_q)^2 + (v_c - v_q)^2}, \tau\right) \tag{4.1}$$

where τ is a truncation value. Such a model reduces the influence of flow outliers just as the truncated matching cost limits the influence of wrong matches [24]. It must be kept in mind that the optical flow is an estimated value and cannot be completely error free. The strength of grouping by motion is defined using a Laplacian kernel as

$$g_{\gamma_m}(\Delta m_{cq}) = \exp\left(-\frac{\Delta m_{cq}}{\gamma_m}\right) \tag{4.2}$$

with γ_m being an empirical motion parameter. The support weight based on the three Gestalt grouping principles (proximity, similarity, and motion) should be redefined for video disparity estimation. We suggest a correlated model, in which the conditional property should be inherited for the integrated support weight as

$$w(c,q) = \begin{cases} g_{\gamma_m}(\Delta m_{cq})^{\frac{\Delta s_{cq}}{\gamma_s}} g_{\gamma_s}(\Delta s_{cq}) & \text{if} \Delta s_{cq} \leq \eta \\ g_{\gamma_m}(\Delta m_{cq})^{\frac{\Delta s_{cq}}{\gamma_s}} g_{\gamma_s}(\Delta s_{cq}) g_{\gamma_p}(\Delta p_{cq}) & \text{otherwise.} \end{cases} \tag{4.3}$$

This model originates from the intuition that color similarity and motion tend to correlate with each other in general. For example, the center pixel and its neighboring pixel have a high likelihood of having different motion vectors if they also differ significantly in color, as expected near object edges. When this occurs, the correlated model decreases the overall support weight as compared with the independent model, since the Laplacian kernel is raised to a power based on the large color difference. Additionally, the two pixels are likely to have similar motions if they also have the same color, as in the homogeneous areas of an object surface. In this case, we can also expect to find a positive correlation between the two metrics. Therefore, the support weight will increase in reference to the independent model. However, while color is an observed quantity, motion is an estimated value. Therefore, color should take precedence over motion when there is a discrepancy between them and the correlation assumption fails. This is precisely what the model in (4.3) enforces. For example, if there is a large difference in color but a small difference in motion, then the value for the correlated support weight is decreased. As a result, the support weight depends on the color cue more than it does the motion cue. In contrast, the independent model always treats all of the Gestalt principles equally. In summary, we define conditional relation between similarity and proximity and correlated relation between similarity and motion.

Table 4.1: Performance comparison of methods on five stereo videos (bad pixel percentage with threshold of 1)

Video/ # of frames	LASW	CostFilter	Proposed method
Tunnel/ 99	1.435 %	2.157 %	**0.997** %
Book/ 40	5.933 %	4.919 %	**3.601** %
Temple/ 99	10.145 %	10.700 %	**10.362** %
Street/ 99	9.978 %	7.305 %	**7.246** %
Tanks/ 99	5.714 %	4.826 %	**4.811** %

4.3 Results

Both qualitative and quantitative experiments are performed on both synthetic and real-world videos. In addition, we examine parameter sensitivity to performance.

4.3.1 Quantitative and qualitative evaluation

To quantitatively evaluate the performance of the proposed method on stereo videos, we use five synthetic stereo videos (400×300, 64 disparity range) with ground-truth disparity map [28]. We compare three methods (LASW, CostFilter, and proposed method) without occlusion filling to compare their pure performance. The LASW method ranks 67^{th} and the CostFilter, which is one of the best performing local methods, ranks 20^{th} on the Middlebury benchmark test. All three methods are efficient one-pass local methods, having similar structure. Table 4.1 shows the average percentage of bad pixels (threshold of 1) over all frames. We ignore borders when computing statistics since they lack correspondences. Table 4.1 illustrates that the proposed method incorporating temporal correspondence information has the best performance.

To subjectively evaluate the performance of the proposed method, we perform experiments on real-world video scenes, "Jamie1" and "Ilkay," from the Microsoft i2i database (320×240, 64 disparity range). The Jamie1 video is more challenging than Ilkay because it contains large homogeneous areas and repetitive patterns, as shown in Fig. 4.3. Fig. 4.3(b) shows the disparity maps produced by LASW, Fig. 4.3(c) depicts the disparity maps yielded by CostFilter, and Fig. 4.3(d) shows the disparity maps obtained by the proposed method. Fig. 4.3 illustrates that the proposed method obtains the best

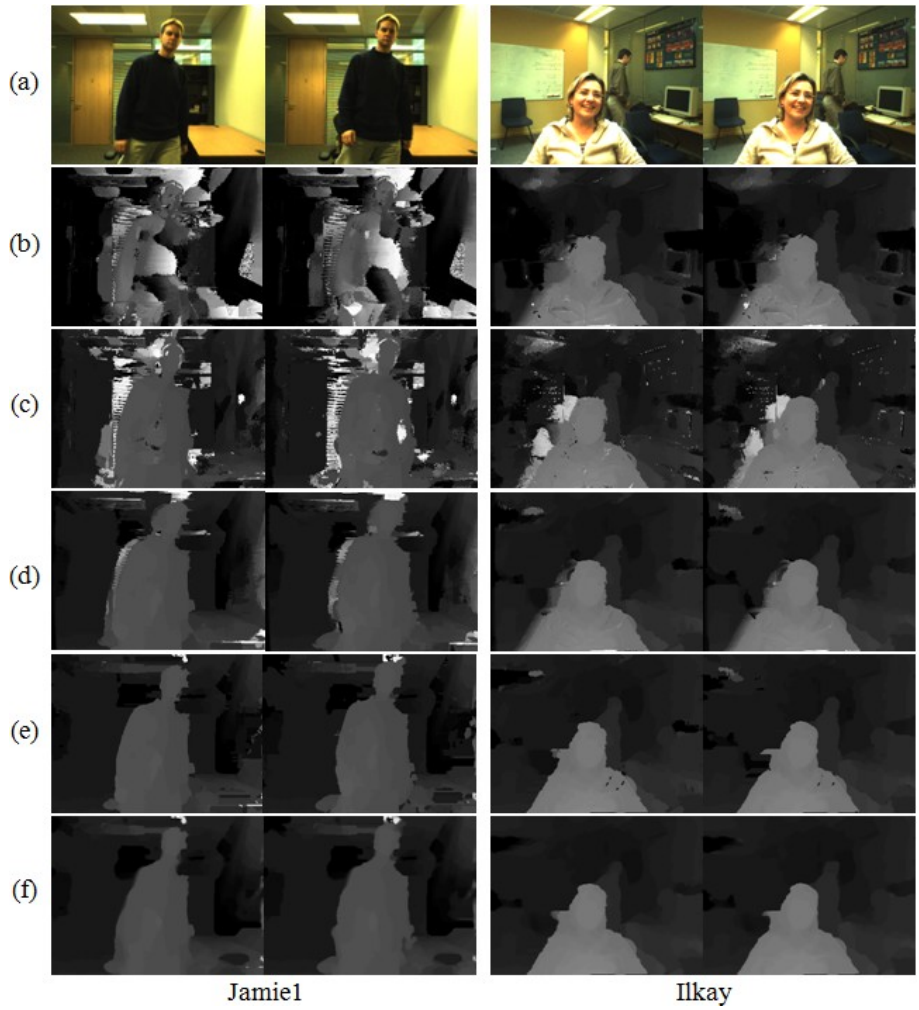

Figure 4.3: Disparity maps for "Jamie1" and "Ilkay". (a) Left frames. (b) LASW. (c) CostFilter. (d) Proposed method. (e) After occlusion filling. (f) After TV [1].

quality of disparity map among three local disparity methods. LASW yields the worst quality and CostFilter produces many errors in homogeneous and repetitive areas. In addition, Fig. 4.3(e) shows the disparity maps where the occlusion areas in Fig. 4.3(d) are filled by valid values, using the occlusion filling mentioned in Section 3.2.4. Fig. 4.3(f) depicts the disparity maps refined with spatial-temporal TV algorithm [1], which reduces errors such as spatial noise and temporal inconsistencies in the background.

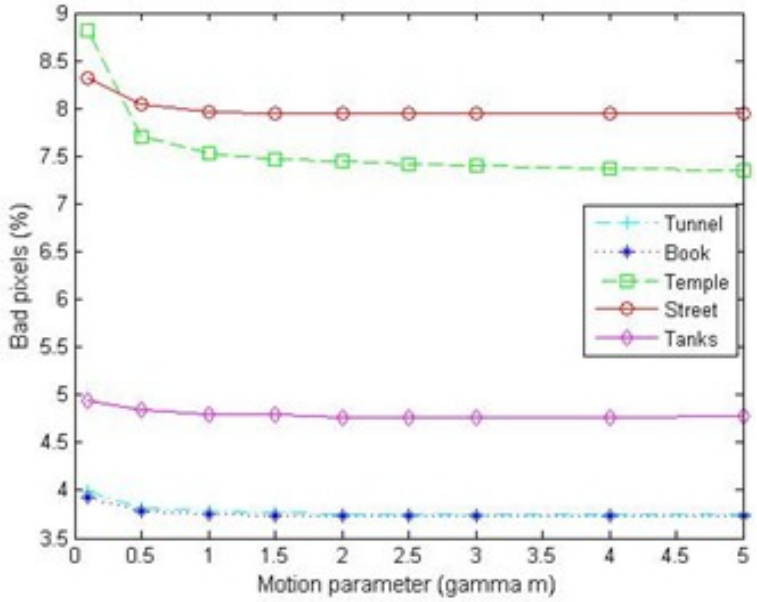

Figure 4.4: Performance evaluation according to parameter γ_m on five stereo videos while the other parameters are kept constant.

4.3.2 Sensitivity to the parameter

The robustness of the proposed method to changing the motion parameter γ_m is examined as done in Section 3.3.4. Fig. 4.4 shows the performance for different γ_m values on five stereo videos. It illustrates that the proposed method achieves almost constant performance according to the motion parameter values. The motion parameter value is not critical in the performance of the proposed method because the motion term is effectively integrated into the support weight computation, as in (4.3).

4.4 Summary

We propose to extend the proposed stereo image disparity method to stereo video domain. Optical flow is utilized to improve the support weight computation. The temporal consistent information promotes spatial disambiguation at motion boundaries with similar color. We impose spatio-temporal consistency to smooth spatial errors and reduce flickering effects. Simulation results verify that the proposed method produces high-quality video disparity maps. Moreover, the proposed method is not sensitive to

the motion parameter value.

4.5 Acknowledgements

This chapter is in part a reprint of a published paper in IEEE Transactions on Multimedia, 2013, and a reprint of a conference paper presented in European Signal Processing Conference, Aug 2013.

Chapter 5

Multi-resolution Depth Processing and Fusion for Large Stereo Panoramic View

This chapter presents a multi-resolution depth processing and fusion algorithm for large stereo panoramic images.

5.1 Introduction

Multiple images are photographed and combined in stitching software to build large-size panoramic images. This process is repeated at the left and right eye position for stereo pair [31]. Large stereo panoramic images have advantages over regular images such as wide field of view and high resolution. If it is seen on a virtual reality (VR) display, it becomes more favorable to customers. However, large-sized stereo panoramas pose a challenging problem for many computer vision tasks.

Multi-resolution or hierarchical (coarse-to-fine) depth schemes can efficiently process large stereo images by reducing matching ambiguity and computational complexity. However, it is difficult to achieve high accuracy and reduce complexity at the same time.

5.1.1 Related work

In stereo matching, local methods [32, 33] based on block matching, and global methods [34, 35, 36, 37] based on belief propagation, have utilized the hierarchical framework. In fact, the multi-resolution scheme helps to avoid local minima in correspondence matching, but it has limitations such as error propagation from coarse to fine levels and blurring at disparity discontinuity boundaries. These limitations cannot guarantee that overall matching accuracy will be improved. Therefore, most hierarchical methods focus on reducing computational complexity at the expense of accuracy. Two hierarchical stereo algorithms [38] and [39] have used a reduced disparity search range to speed up processing at the next level in the hierarchy. However, the reduced search range, which is spatially constant, may propagate error. The hierarchical stereo method with thin structure [40] emphasizes the importance of search range shifted by the disparity of the corresponding coarse point. However, there is no discussion as to how to find the optimal disparity search range, which can suppress error propagation.

The hierarchical segment-based matching scheme [33] and adaptive windowing scheme [38, 41] have been presented to reduce blurring at disparity boundaries. However, the former is not efficient because color segmentation and plane fitting are known to have high complexity, and they are required at every stage. The latter is not able to effectively segment object boundaries with rectangular windows because real 3D objects have an arbitrary shape. The boundary blurring and smearing artifacts are inherent in the hierarchical scheme. Therefore, an additional sophisticated edge-preserving refinement is required.

To better propagate the coarse results, the disparity surface, referred to as the spatio-disparity space, is introduced in [42]. It represents the quality of each possible match that corresponds to the matching correlation value. The final disparity is detected based on the disparity surface enhanced by a non-linear filter that suppresses noise and eliminates ambiguities. An adaptive scale selection mechanism [43] is presented to determine the most favorable scale level at which the surface is salient when performing hierarchical stereo matching. All of the multi-resolution methods mentioned above deal with small images, such as the Middlebury datasets [11]. In this chapter, we are dealing with panoramic stereo images about twenty times larger than those of the Middlebury

datasets. As image size increases, so does the importance of mitigating the limitations of hierarchical disparity schemes. Another challenging factor arising in the large-sized image will be discussed in Section 5.2.

In the coarse-to-fine scheme, a coarse disparity map serves as initial estimate for the next level estimation. The more reliable the initial estimate is, the more accurate the disparity result at the next level will be. In general, the coarse disparity map suffers from staircase artifacts due to disparity quantization from integer disparity estimation as well as edge-smearing artifacts. Moreover, the coarse disparity map needs to be upsampled in order to be used for the next level initialization. The upsampling process makes the artifacts more noticeable. Therefore, sub-pixel disparity estimation that can minimize the quantization artifacts is required for hierarchical methods.

Typically, real-world images and videos are susceptible to various noise factors. Disparity estimates computed in different scales and times tend to show inconsistency. Therefore, a consistency function should be integrated in the disparity estimation process. For instance, temporal consistency should be taken into account in video disparity estimation over time. Similarly, scaling consistency needs to be considered in multi-scale disparity processing, where results at different scales suffer from different error types. To the best of our knowledge, the disparity scaling consistency issue has not been investigated.

5.1.2 Contributions

In this chapter, the main contribution is the adaptively determined pixel-wise disparity search range, which is based on the local structure of image and initial disparity map. The optimal adaptive search range helps to propagate smoothness in the homogeneous areas and suppress the error appearing in the fine level estimation. It also contributes to computational speed by reducing the search complexity.

We propose a reliable multiple parabola fitting technique for sub-pixel disparity estimation, which extends the conventional fitting method. The proposed multiple fitting is not only efficient to implement but also improves the disparity accuracy while alleviating the drawback of the conventional method.

We investigate the spatial-multi-resolution TV to enforce spatial and scaling con-

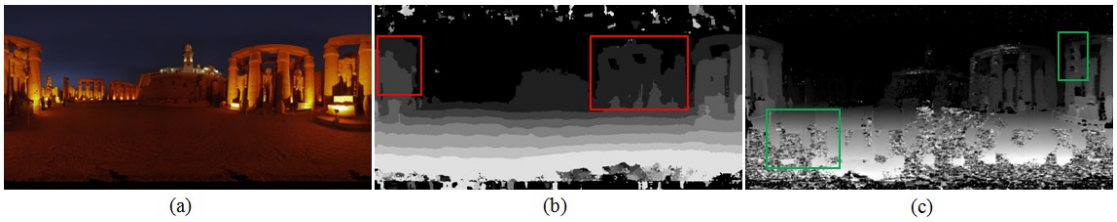

Figure 5.1: Depth maps in hierarchical framework. (a) Left panorama. (b) Coarsest level depth map (downsampled). (c) Finest level depth map.

sistency. The adaptive disparity search range and spatial-multi-resolution TV play a role in fusing multi-scale disparity results by combining their complementary information.

We evaluate the advantages and effectiveness of the proposed multi-resolution depth processing quantitatively, using the Middlebury datasets with the ground-truth depth, and then demonstrate that the proposed algorithm also achieves high quality of depth map on large panoramic views.

5.1.3 Organization

This chapter is organized as follows. Section 5.2 describes the problem that we are solving. The details of our proposed method are presented in Section 5.3. Section 5.4 shows experimental results and discusses their significance. Section 5.5 summarizes the proposed method with some remarks.

5.2 Problem Statement

Our goal is to obtain high quality and high-resolution depth maps from large-size stereo panoramas ($8,192 \times 4,096$), while using the proposed local disparity method [2]. For large-size data processing, we consider a multi-resolution approach and partitioning-stitching scheme similar to panoramic view construction. Fig. 5.1 demonstrates the problem addressed in this chapter. Fig. 5.1(a) shows the left panoramic view. Fig. 5.1(b) and (c) depict the corresponding depth maps for the low-resolution image and high-resolution image, respectively. At the coarse level, overall information and prominent features are present in the smooth form; however, details and sharp edges are lost, as shown in the red boxes in Fig. 5.1(b). In contrast, most details and edges are preserved

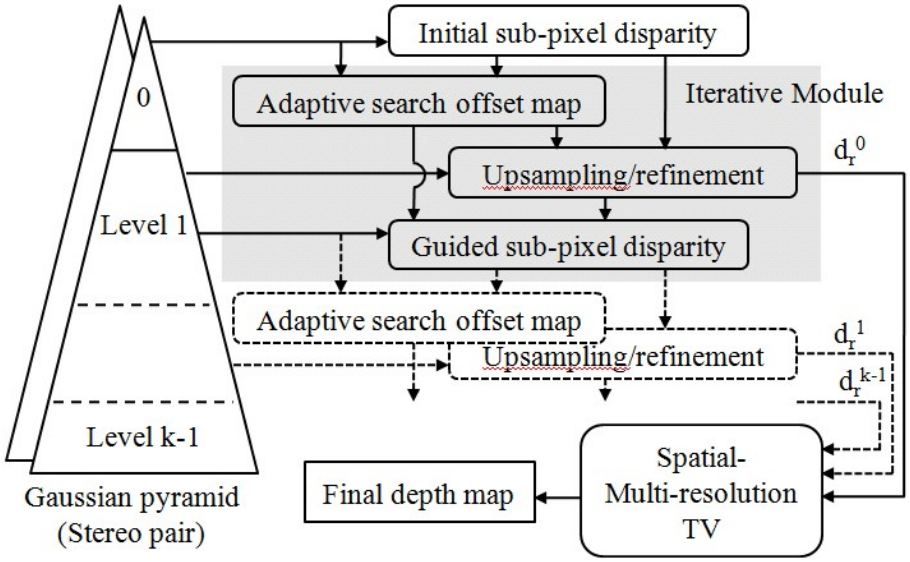

Figure 5.2: Diagram of system framework.

at the fine level; however, many errors appear in the low and high-textured areas, as shown in the green boxes in Fig. 5.1(c). It can be observed that fine detail is too small to detect at the coarse level. At the fine level with larger scale, homogeneous areas tend to increase. For instance, the local structure in the low-textured area becomes more ambiguous as image resolution increases, and the structures in high-textured areas tend to look like repetitive patterns. Such matching ambiguities cause disparity errors. The challenging problem is how to fuse only beneficial characteristics at the coarse level and fine details/edges well preserved at the finer level while suppressing undesirable errors. Finally, we seek an algorithm that increases disparity accuracy while reducing computational complexity.

5.3 Proposed Method

5.3.1 System framework

Fig. 5.2 illustrates the proposed multi-stage framework, which consists of four main blocks: sub-pixel disparity estimation, adaptive search offset map, disparity up-sampling with refinement, and spatial-multi-resolution TV. First, we build a stereo pair

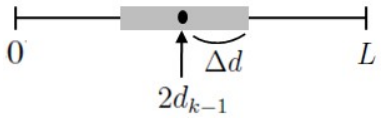

Figure 5.3: Illustration of disparity search range.

of Gaussian pyramids with k levels, and then estimate an initial sub-pixel disparity. Using the initial sub-pixel disparity and image intensity, the adaptive search offset map is constructed, which guides the next level disparity estimation and controls the weight when the initial disparity map is upsampled and refined simultaneously. The refined disparities are incorporated with the adaptive search offsets for guiding the next level disparity estimation and are combined for multi-resolution disparity fusion. The shaded box in Fig. 5.2 indicates an iterative module. The final disparity map is obtained by applying the spatial-multi-resolution TV to the refined disparity maps.

5.3.2 Adaptive search range based on eigenvalues of structure tensor

Most single-resolution disparity methods attempt to find the best matching point by searching the entire disparity range. In the hierarchical scheme, we can take advantage of given initial priors to adaptively minimize the search range without loss of accuracy. The reduced disparity search range (r_k) for the next level k can be defined as

$$2d_{k-1} - \Delta d \leq r_k < 2d_{k-1} + \Delta d \tag{5.1}$$

where d_{k-1} is the disparity estimate at level $k-1$, and Δd is the local disparity search offset. Fig. 5.3 illustrates the disparity search offset (Δd) and full search range (L). The search offset (Δd) should be properly determined at each pixel location so that it prevents the next level estimate from being trapped in a local minimum of the matching cost function. It is very important to choose the optimal Δd because a small Δd at a certain point reduces matching ambiguity and increases speed, while a larger Δd at a different point is able to resolve complex object boundaries more effectively [40]. Therefore, the estimation quality and speed directly depend on d_0 and Δd.

In general, image structure is closely related to disparity estimation. Ambiguous or complicated image structures cause disparity error. Disparity discontinuities occur at the object border, where there is mostly distinct structure. Therefore, we investigate a structure-based search offset approach. We take into account the structure tensor, known as the second-moment matrix, which is based on the summation of the outer product components of the local gradient from a neighborhood [44]. We consider a local neighborhood because it provides reliable structure and orientation information about image features, even in the presence of noise. We propose to adaptively determine Δd according to the structure tensor information. Two-dimensional features of an image I can be detected by 2D structure tensor:

$$J = \sum \nabla I \nabla I^T = \begin{pmatrix} \sum I_x^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^2 \end{pmatrix}. \tag{5.2}$$

J is a symmetric positive semi-definite matrix with two non-negative eigenvalues: λ_{max} and λ_{min}. The eigenvalues of J can be analytically derived as

$$\begin{aligned} \lambda_{max} &= \frac{\Sigma I_x^2 + \Sigma I_y^2 + \sqrt{(\Sigma I_x^2 - \Sigma I_y^2)^2 + 4(\Sigma I_x I_y)^2}}{2} \\ \lambda_{min} &= \frac{\Sigma I_x^2 + \Sigma I_y^2 - \sqrt{(\Sigma I_x^2 - \Sigma I_y^2)^2 + 4(\Sigma I_x I_y)^2}}{2}. \end{aligned} \tag{5.3}$$

There are three distinct cases for the relative values of these two eigenvalues in (5.3) [45]:

- $\lambda_{max} \approx \lambda_{min} \approx 0$: low-textured area with almost no structure where both partial derivatives (I_x and I_y) are small.
- $\lambda_{max} \gg 0, \lambda_{min} \approx 0$: one dominant orientation, like edges where both derivatives are large (diagonal edge) or only one of them is large (horizontal or vertical edge).
- $\lambda_{max} \gg 0, \lambda_{min} \gg 0$: high-textured area with ambiguous orientation elsewhere.

We exploit both image intensity and initial disparity map as a prior for the local structure acquisition because they may reveal different but complementary structure characteristics. We make the following observations, which justify our combination of image and disparity to determine local structure:

- Matching ambiguities tend to occur in both low and high-textured areas.
- Disparity jumps occur in real disparity edges, which generally match the corresponding image edges.
- Image edge structure might not be noticeable at some disparity boundaries where the image gradient is small.
- Initial disparity estimates tend to be unreliable along disparity edges due to occlusion and lack of texture details.

The first observation indicates that Δd should be small enough to reduce the ambiguity at the next level. A small Δd leads to a small search range, which in turn propagates the desirable smoothness property of the coarse disparity estimates. The other observations indicate that Δd should be large enough to detect big disparity changes and, thereby, recover the initial error. For the second and third observations, we note that the image (appearance) edges do not always match the real disparity edges. There are two cases: (1) image edge structure exists without disparity edge and (2) disparity edge structure exists without image edge. The former is not critical because the image edges from significant textures in the flat region do not cause matching errors. On the other hand, the latter implies that structure information from an image is not sufficient for obtaining accurate structures. For the second case, the initial disparity map can be utilized as complementary information. However, the initial disparity estimate is likely to be unreliable near object boundaries as in the last observation. The combination of two priors will be a good solution. We can define a function of eigenvalue of the matrix J satisfying the two main observations (first and second) as

$$f(\lambda_{max}, \lambda_{min}) = \frac{\lambda_{min} + \epsilon}{\lambda_{max} + \epsilon} \tag{5.4}$$

where ϵ is used to increase the robustness near zero eigenvalue. The extremely small value of ϵ makes the function sensitive to eigenvalues. A reasonable value of ϵ is 0.1, which is found empirically. The function f outputs approximately "1" in case of both low ($\lambda_{max} \approx \lambda_{min} \approx 0$) and high-texture ($\lambda_{max} \gg 0, \lambda_{min} \gg 0$) and approximately "0" around edges ($\lambda_{max} \gg 0, \lambda_{min} \approx 0$). In the ellipsoidal representation of the matrix

J, two eigenvalues indicate scaling in each orthogonal direction. Note that two eigenvectors of the matrix J are orthogonal since J is symmetric [46]. Let J_I and J_D be the matrices from the image intensity and initial disparity, respectively. We have four eigenvalues from two matrices: two maximum eigenvalues ($\lambda_{Imax}, \lambda_{Dmax}$) and two minimum eigenvalues ($\lambda_{Imin}, \lambda_{Dmin}$). In most areas of the disparity map, zero eigenvalues representing no local structure ($\lambda_{Dmax} \approx \lambda_{Dmin} \approx 0$) are observed due to the homogeneous characteristic of disparity maps except at disparity edges. At the disparity edges, it shows the same dominant orientation as that of the image. This implies that corresponding eigenvectors from J_I and J_D have the same direction except in homogeneous areas of the disparity map, where zero eigenvalues are obtained. Therefore, the eigenvalues can be linearly combined as $\lambda_{max} = \lambda_{Imax} + \lambda_{Dmax}$ and $\lambda_{min} = \lambda_{Imin} + \lambda_{Dmin}$. Note that the eigenvalue computation is performed on the normalized image intensity and disparity map for direct linear combination. An exponential function based on the combined eigenvalues at the pixel p can be defined as

$$g(p) = e^{-f_p(\lambda_{max}, \lambda_{min})} \tag{5.5}$$

For simplicity, we define this function as a local edge strength function, which produces a high value along the edges and a low value on the low- and high-textured area.

Fig. 5.4 depicts the Teddy image, initial disparity, and ground-truth disparity including three local edge strength maps, which motivate the reason for combining two strength maps. Fig. 5.4(d) and (e) show the local edge strength map from the image intensity and initial disparity map, respectively. Some image edge structures do not appear at the real object boundary in the green box, as shown in Fig. 5.4(d), while they are present in Fig. 5.4(e). This corresponds to the third observation above. The wrong and very weak edge structures are shown in the red box in Fig. 5.4(e), while the true and strong edge structures are present in Fig. 5.4(d). This corresponds to the fourth observation. Therefore, we need to combine the two structures by adding their eigenvalues so that the structure lost in either edge strength map is recovered, as shown in Fig. 5.4(f).

Finally, the search offset Δd is adaptively determined according to four eigen-

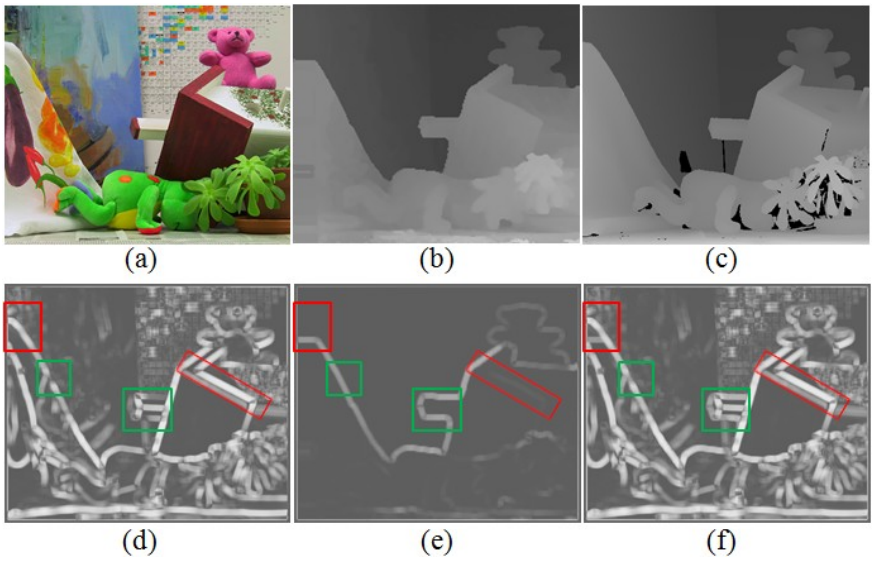

Figure 5.4: Examples of the local edge strength map. (a) Left image. (b) Initial disparity. (c) Ground truth. (d) Local edge strength of (a). (e) Local edge strength of (b). (f) Local edge strength of the combined local structure using (5.5).

values of two structure tensors as

$$\Delta d = [g(p) \times \frac{L}{2}] \tag{5.6}$$

where $[\cdot]$ represents nearest integer operator, and L is the full disparity search range, as shown in Fig. 5.3. The different Δd at each pixel point forms an adaptive disparity search offset map, and then the search range, r_k, is calculated from initial disparity value d_{k-1} and offset Δd. The adaptive disparity search offset will be used in both next-level estimation and depth refinement as a core function of the proposed algorithm.

5.3.3 Sub-pixel disparity

For the sub-pixel disparity estimation, we extend the local disparity method [2]. The local method provides an accurate cost function by effectively combining three cost measures: census, color, and gradient. However, it can handle only integer disparity. Therefore, a sub-pixel algorithm is needed to integrate with the integer disparity method in order to increase matching precision and, in turn, provide better multi-scale disparities for fusion. One of most popular algorithms is the quadratic polynomial interpolation known as parabolic fitting. It finds the fractional minimum point by fitting a parabolic

function to three discrete matching costs: an initial minimum cost (center point) and two adjacent costs. Thus, the performance of parabolic fitting directly depends on the cost function. In addition, such a fitting method is simply applicable to various local disparity methods as well as reduces the disparity discontinuity caused by the quantization [47, 48]. However, it results in the drawback known as the pixel-locking effect, which represents systematic bias toward the minimum integer point [49]. This bias corresponds to an erroneous ripple in 3D reconstruction [50].

A simple and efficient way to reduce the pixel locking effect by taking advantage of the accurate cost measure provided in [2] is to investigate more cost points for enhanced fitting. However, we observe that the brute-force extension performs worse than the original parabolic fitting does. This is because the reliability of the cost function sharply decreases as additional points grow further from the minimum integer point. To increase the sub-pixel accuracy as well as alleviate the locking effect, we propose using a multiple fitting method with reliability check. The key of the proposed algorithm is to check the reliability of the additional cost points. We assume that matching cost increases at least linearly as the matching window slides from the center point. This assumption is justified by the fact that the matching cost is aggregated over a 2D window. For the fitting operation, the cost function can be approximated as a piece-wise quadratic function. Five integer disparity points d_{--}, d_{-}, d_0, d_{+}, and d_{++} are considered, where $d_{+} = d_0 + 1$ and $d_{++} = d_0 + 2$. The quadratic cost function can be expressed as

$$A(d) = ad^2 + bd + c \tag{5.7}$$

where d is a continuous disparity value, and a, b, c are the parabolic parameters to be estimated. The minimum point d_m of the quadratic cost function passing three disparity points (d_{-}, d_0, and d_{+}) is estimated to sub-pixel precision:

$$d_m = d_0 - \frac{A(d_{+}) - A(d_{-})}{2A(d_{+}) - 4A(d_0) + 2A(d_{-})}. \tag{5.8}$$

The algorithm is as follows:

Algorithm 1: Multiple fitting for sub-pixel disparity

1: Define three cost functions: $A_1(d_{--}, d_0, d_+)$,
$A_2(d_-, d_0, d_+)$ and $A_3(d_-, d_0, d_{++})$ as in (5.7)

2: Find three minima: d_{m1}, d_{m2} and d_{m3} using (5.8)

3: **if** slope$(d_{--}, d_0) \geq$ slope(d_-, d_0) &&
slope$(d_{++}, d_0) \geq$ slope(d_+, d_0) **then**

$d_m = \frac{2d_{m1}+3d_{m2}+2d_{m3}}{7}$

4: **elseif** slope$(d_{--}, d_0) \geq$ slope(d_-, d_0) **then**

$d_m = \frac{2d_{m1}+3d_{m2}}{5}$

5: **elseif** slope$(d_{++}, d_0) \geq$ slope(d_+, d_0) **then**

$d_m = \frac{3d_{m2}+2d_{m3}}{5}$

6: **else then**

$d_m = d_{m2}$

7: **endif**

The final sub-pixel disparity d_m is computed by taking the weighted mean of those satisfying the reliability condition out of minimum points (d_{m1}, d_{m2}, d_{m3}).

To evaluate the pixel locking effect, sub-pixel disparity estimation is performed on the planar region (red box) in Fig. 5.5(a). The conventional parabolic fitting produces many peaks with a bias toward integer disparity, as shown in Fig. 5.5(c), while the proposed multiple fitting spreads out and reduces the pixel locking effect, as shown in Fig. 5.5(d). It is worth noting that the proposed multiple fitting method is also applicable to various local disparity methods. The fitting technique becomes most effective when it is integrated into the disparity method providing accurate cost function, since it is cost function-dependent.

5.3.4 Disparity refinement with upsampling

In general, smearing (blurring) error at disparity boundaries shows big deviation from the ground-truth disparity. This error is critical for multi-resolution scheme

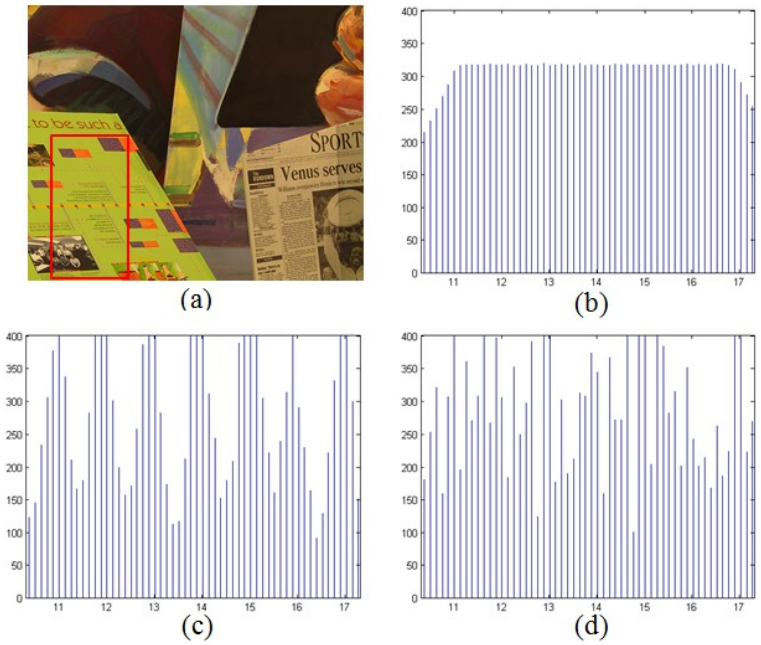

Figure 5.5: Histograms of sub-pixel disparities for a planar region on Venus. (a) Left image. (b) Ground-truth. (c) Parabolic fitting. (d) Our approach (Multiple fitting).

where upsampling process is required. Traditional upsampling methods, such as bilinear and bicubic interpolation, result in the blurring of sharp edges. The Joint Bilateral Upsampling (JBU) [51], based on the edge-preserving bilateral filter [52], upsamples and enhances the low-resolution range map with a high-resolution color image. However, JBU method causes two artifacts: texture copying on object surface with noticeable texture and blurring in the edge area with similar color. To overcome these problems, the Pixel Weighted Average Strategy (PWAS) [53] extends JBU with an additional factor, credibility map. The credibility map is obtained from the absolute gradient of the low-resolution disparity map. It controls the contribution of the weighted averaging by assigning a lower weight to disparity edge pixels. The texture copying artifacts frequently occur when a noisy Time-Of-Flight (TOF) range map is upsampled. This is because noises in the flat depth area cause textures from the color image to be transferred into geometry patterns of the upsampled range map [54]. The texture copying does not appear in our stereo matching method using two views since stereo matching approaches are not likely to produce errors on the object surface with textures.

We propose reuse of two functions (the support weight and local edge strength

already computed) in order to refine the smearing error, similar to the PWAS. The advanced support weight function [2] has been used to put a reasonable weight on each pixel cost for cost aggregation. It can be hereby reused as a sort of bilateral filter function:

$$w(p,q) = \begin{cases} e^{-\frac{\Delta s_{pq}}{\gamma_s}} & \text{if } \Delta s_{pq} \leq \tau \\ e^{-\frac{\Delta s_{pq}}{\gamma_s}} e^{-\frac{\Delta p_{pq}}{\gamma_p}} & \text{otherwise} \end{cases} \tag{5.9}$$

where Δs_{pq} and Δp_{pq} are the color and the spatial distance between pixel p and pixel q, respectively, and τ is a color threshold determining color similarity between two pixels. γ_s and γ_p are photometric and geometric scaling parameters that are determined empirically. The support weight $w(p,q)$ is functionally similar to two Gaussian kernels (spatial and range) in PWAS except that it uses Laplacian kernel and incorporating a condition to minimize blurring at edges with similar colors. The spatial term is not considered if the color difference is less than the threshold. This is because the spatial term will worsen blurring at edges with similar colors by blind aggregation. The local edge strength map $g(p)$ in (5.5) and the credibility map in PWAS share the similar goal of providing edge structure. The difference is that $g(p)$ is based on the combined structure tensor of both image intensity and initial disparity, while the credibility map is based on the gradient of the initial disparity. To reduce the influence of errors near an edge and then refine the edge blurring, local edge weakness is taken into account. The local edge weakness function is inversely proportional to the normalized edge strength function. It can be defined as

$$h(p) = 1 - g(p). \tag{5.10}$$

In the refinement process with upsampling, both support weight $w(p,q)$ and edge weakness $h(q)$ are used to put each different weight on the initial disparity value. The initial disparity value is replaced by the weighted average of a neighborhood. The disparity refinement is defined as

$$d_r(p) = \frac{\sum_{q \in W} w(p,q)h(q)d_0(q)}{\sum_{q \in W} w(p,q)h(q)} \tag{5.11}$$

where W represents the window patch and $d_0(q)$ is the low-resolution disparity value at pixel q. A small $h(q)$ along the edge implies that the corresponding disparity value is not reliable, and thus does not have as much influence on refinement.

5.3.5 Scaling consistency and multi-resolution fusion

The multi-resolution scheme is well known to be advantageous in terms of speed and memory usage. However, scale dimensional inconsistency can occur due to the frequent upsampling/downsampling process, especially for real-world images. This inconsistency is related to the fact that objects in the world appear in different ways, depending on the scale of observation [55]. To alleviate the scaling inconsistency as well as the spatial inconsistency at the same time, we investigate TV regularization algorithm over spatial and multi-resolution dimensions. The spatial-multi-resolution TV is based on the augmented Lagrangian method for image restoration presented in [56], which enforces the spatial-temporal consistency for video disparity maps. We adapt it for scale consistency where disparity estimates at several scales are used, instead of temporal consistency where several frames of video disparity estimate are used. We assume that typically, disparity value varies smoothly in both spatial and scaling dimension except at 3D boundaries. We treat a sequence of multi-resolution disparity maps as a scale-space volume: a 3D function $f(x, y, s)$ with the spatial coordinate (x, y) and the scale dimensional coordinate s. The multi-resolution disparity maps must be scaled to the same size so that it can be regarded as one volume. To reduce spatial and scale dimensional noise while preserving sharp edges, we solve the following TV regularized l_1 minimization problem:

$$\underset{\mathbf{f}}{\text{minimize}} \quad \mu||\mathbf{f} - \mathbf{g}||_1 + ||\mathbf{Df}||_{TV} \tag{5.12}$$

where the vector $\mathbf{f}$ is the unknown disparity map, the vector $\mathbf{g}$ is the multi-resolution disparity map, and the operator $\mathbf{D} = [\beta_x \mathbf{D}_x^T, \beta_y \mathbf{D}_y^T, \beta_s \mathbf{D}_s^T]^T$ denotes the forward difference operators along the horizontal, vertical, and scaling directions. The parameter μ is the regularization constant that controls the relative emphasis of the objective and regularization terms. The parameters $(\beta_x, \beta_y, \beta_s)$ also control the relative emphases of the spatial and scale dimensional terms. This minimization problem is solved by using an iterative method based on the augmented Lagrangian and alternating direction method,

as in [56].

Using the spatial-multi-resolution TV, we can fuse complementary disparities at different scale resolution while maintaining spatial-scaling consistency. It is meaningful that it can recover disparity features that are not visible in a certain scale disparity map. We will demonstrate that this approach for multi-resolution disparity estimation can be a new and effective extension of TV image restoration application. Additionally, the adaptive search range mentioned in Section 5.3.2 plays a role in fusing the initial disparity to the high-scale direction by guiding the next level estimation.

5.4 Results

We have proposed a multi-resolution depth processing and fusion approach for large panoramic views. First, we evaluate the proposed scheme quantitatively using the Middlebury datasets with ground-truth depths available. Second, we present the performance of the proposed method on large real-world panoramic views compared to conventional methods.

5.4.1 Overall performance of the multi-resolution scheme

We apply the proposed hierarchical scheme to the Middlebury datasets. All parameters are fixed throughout the experiment. The sub-pixel disparity estimation parameters are the same as those used in [2]. For the disparity upsampling/refinement, the support window size is 9×9, and the other parameters for the spatial and range term are reduced proportionally ($\gamma_p = 6$ and $\gamma_s = 9$).

Fig. 5.6 shows intermediate results from the proposed multi-resolution scheme on the Middlebury datasets. Fig. 5.6(b) shows initial disparity maps at level 0, where images are downsampled by a factor of 2. Fig. 5.6(c) depicts the local edge strength function $g(p)$ defined in (5.5), which is computed from the structure tensors of image intensity and initial disparity. This map will form the adaptive search offset (Δd) map, using (5.6). Fig. 5.6(d) shows the upsampled/refined disparity map from the initial disparity map, incorporating the edge weakness function $h(p)$ defined in (5.10). Fig. 5.6(e) depicts the final disparity map guided by the adaptive search offset map. As shown in

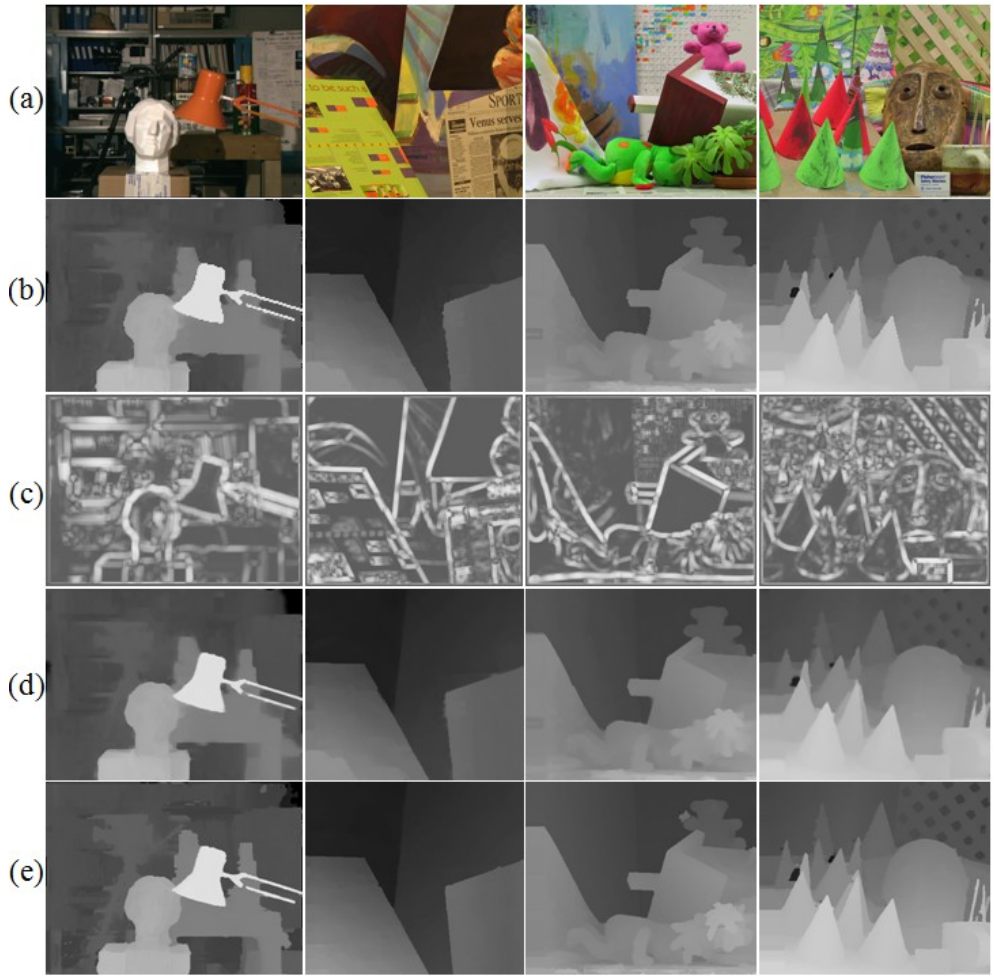

Figure 5.6: Intermediate results from our multi-resolution (2-level) disparity processing on the Middlebury datasets. (a) Left image. (b) Initial disparity at the level 0. (c) Local edge strength map $g(p)$. (d) Upsampled/refined disparity. (e) Final disparity.

Table 5.1: Performance comparison of three methods on the Middlebury benchmark test (bad pixel (error) rates (nonocc/all/disc) with threshold of 1 and search range percentage which denotes the average percentage of the shaded area in Fig. 5.3)

Dataset	Single-res. Full-search		Classical multi-res. Fixed-search		Proposed multi-res. Initial	Refined	Fixed-search		Adaptive-search	
	error (%)	range	error(%)	range	error (%)	error (%)	error (%)	range	error (%)	range
Tsukuba	2.30/2.82/**8.37**	100%	**2.21**/2.76/8.82	50%	5.10/6.10/20.1	5.04/5.84/19.9	2.39/2.91/9.59	50%	2.24/**2.76**/8.86	50%
Venus	0.13/0.43/1.67	100%	0.19/0.52/2.35	50%	0.95/1.77/9.87	0.64/1.39/6.71	0.14/0.47/1.91	50%	**0.12/0.43/1.67**	46%
Teddy	5.11/11.1/13.9	100%	5.57/11.5/14.2	50%	7.97/14.7/21.9	7.64/14.6/20.6	5.10/11.0/13.8	50%	**5.07/11.0/13.8**	47%
Cones	**1.83/7.55/5.45**	100%	1.93/7.61/5.67	50%	5.87/12.3/16.6	5.58/11.7/15.5	2.03/7.79/6.07	50%	1.91/7.67/5.70	48%

Fig. 5.6, the guided disparity map achieves the best quality, preserving fine detail as well as a smooth surface.

We conduct experiments to quantitatively evaluate the proposed multi-resolution scheme on Middlebury datasets. Table 5.1 shows performance comparisons of three schemes: single-resolution, conventional, and the proposed multi-resolution. Moreover, it depicts how disparity search range determination affects disparity performance. The single-resolution method is performed with full search range, while the multi-resolution schemes are done with reduced search range (fixed and adaptive). The initial disparity errors in the proposed multi-resolution scheme are moderately reduced by the upsampling/refinement process on three kinds of statistics (non-occlusion, all, discontinuity). The large discontinuity error reduction (from 9.87 to 6.71) on Venus implies that the refinement process works well along edges. Both the conventional and proposed multi-resolution with reduced fixed-search range perform worse than the single-resolution method. This demonstrates that a proper choice for search range is crucial for the hierarchical scheme because the fixed-search range can propagate errors. On the other hand, the proposed adaptive-search scheme achieves overall better performance than the other schemes, while reducing disparity search range. However, the full-search single resolution approach shows better results on the Cones image containing the most complicated structures. This implies that larger search range might be required for stereo images with high structure complexity. However, larger disparity search range directly leads to high computational complexity.

To reduce computational complexity, hierarchical schemes typically focus on search range reduction. The complexity of the single-resolution disparity estimation [2] is $O(NWr)$, where N and W are the size of image and support window, respectively, and r is the disparity search range. In case of 2-level multi-resolution scheme, the total complexity of the hierarchical disparity estimation is reduced to $\frac{5}{8}O(NWr)$ if the search range is reduced by 50%. As a result, the proposed multi-resolution scheme is able to have complexity gain of approximately $\frac{3}{8}O(NWr)$. It is feasible to further decrease the complexity by applying higher scale pyramid scheme. For Middlebury datasets, this is undesirable because their image size is too small to apply three levels or more. For the Tsukuba image in the three-level pyramid, the coarsest image size becomes 96×72 while the support window size for stereo matching is fixed to 35×35 along scale levels. Moreover, the disparity range ($0 \sim 4$) becomes indistinguishable. For the proposed

Table 5.2: Robustness of the proposed mutli-resolution scheme to other initial local disparity methods

Dataset	LASW error rates (%) - nonocc/all/disc				CostFIlter error rates (%) - nonocc/all/disc			
	Single-res.	Proposed multi-res.			Single-res.	Proposed multi-res.		
	Full-search	Initial	Refined	Adapt.-search	Full-search	Initial	Refined	Adapt.-search
Tsukuba	2.94/4.82/**9.50**	14.9/16.7/25.7	14.3/16.1/26.0	**2.65/4.54**/9.85	2.52/3.30/**8.74**	13.7/14.5/26.7	14.8/15.4/24.6	**2.36/3.00**/8.78
Venus	3.98/5.55/15.1	6.38/7.97/20.2	**2.89/4.51**/18.5	3.26/4.84/**14.7**	2.04/3.19/16.1	4.05/5.06/19.5	2.54/3.50/16.3	**1.85/2.95/15.5**
Teddy	14.3/22.9/**24.4**	23.2/31.0/37.6	20.7/28.8/35.0	**13.8/22.5**/24.5	8.47/17.0/19.0	14.5/21.4/29.0	13.1/20.4/26.2	**8.14/16.5/18.5**
Cones	9.43/19.5/17.2	23.0/31.5/38.2	20.3/29.1/34.4	**9.28/19.4/17.0**	**3.62**/12.6/**9.62**	14.2/20.9/28.6	15.1/21.5/27.3	3.74/**12.5**/10.0

scheme, there are two additional steps: the adaptive search offset (Δd) construction and disparity upsampling/refinement ($d_r(p)$) process compared to the conventional scheme. Their additional computation load is negligible. For the Tsukuba image, it takes about $12s$ to complete disparity estimation as in [2], while it takes about $0.1s$ and $0.4s$ to compute Δd and $d_r(p)$, respectively, using a Central Processing Unit (CPU). The local disparity method used in the proposed scheme is suitable for real-time processing using a Graphics Processing Unit (GPU) [2]. The parallel computation using GPU would decrease the final processing time significantly. Run-time efficiency in local methods depends on how many computations in an algorithm can be run in parallel processors. Note that for Middlebury datasets ideally acquired in the laboratory, spatial and scaling consistency enforcement is not necessary.

We evaluate how robust the proposed multi-resolution scheme is to other initial disparity algorithms. For experiment, we select two algorithms: LASW [3] and Cost-Filter [4], which are well-known local algorithms. Experiments are performed without filling process since the left-right filling is not obviously discussed in [3], while the local method [2] includes a region-voting algorithm for occlusion filling. Table 5.2 shows the performance comparison of the single and proposed multi-resolution method implemented with different initial disparity algorithms. Similar to Table 5.1, it shows that the proposed multi-resolution scheme achieves better overall performance while reducing complexity compared to the single-resolution scheme. This verifies that the proposed hierarchical scheme does not depend on initial results and is robust to a variety of local algorithms.

Table 5.3 shows the performance comparison of hierarchical methods including global methods. Letters G and L in the first column denote global and local method, respectively. In fact, most of high ranked methods on Middlebury benchmark site are

Table 5.3: Comparison of hierarchical methods (bad pixel rate with threshold of 1)

Methods /Type	Error rate(%) - nonocc/all/disc			
	Tsukuba	Venus	Teddy	Cones
DoubleBP [35] /G	**0.88/1.29/4.76**	0.13/0.45/1.87	**3.53/8.30/9.63**	2.90/8.78/7.79
Proposed /L	2.24/2.76/8.86	**0.12/0.43/1.67**	5.07/11.0/13.8	**1.91/7.67/5.70**
RealtimeBP [36] /G	1.49/3.40/7.87	0.77/1.90/9.00	8.72/13.2/17.2	4.61/11.6/12.4
AdaptiveScale [43] /L	2.37/4.05/9.91	1.50/2.49/10.0	10.2/17.0/21.5	4.44/12.3/10.3
HBpStereoGPU [37] /G	3.37/5.34/13.6	1.12/2.06/14.1	12.2/19.0/27.2	6.29/14.2/16.4
StereoBoundary [38] /L	10.2/11.5/20.3	4.58/5.22/14.2	8.39/13.7/20.0	5.03/10.8/13.9
HstereoSeg [33] /L	-	-	13.3/21.7/24.8	9.10/17.5/21.3

global methods while local methods are efficient to real-time process because many associated computations can be run in parallel. Due to their different characteristics, they are not directly compared. Nevertheless, there is only one global method (DoubleBP) that performs better than the proposed method on the Middlebury benchmark site. The proposed scheme outperforms other local methods and even some global methods, as shown in Table 5.3.

5.4.2 Sub-pixel results

Sub-pixel disparity method can increase the matching precision and significantly reduce the quantization error (staircase error) on the disparity map. In the proposed entire system, it contributes to providing more precise multi-scale disparities that will be combined for fusion. Fig. 5.7 illustrates that the proposed sub-pixel estimation reduces the disparity quantization error. The integer disparity yields many ripples, as shown in Fig. 5.7(c), while the proposed multiple fitting produces a smooth disparity surface, as shown in Fig. 5.7(d).

Sub-pixel algorithm should show robustness to various surface types. Therefore, we perform the experiment on specific areas, such as a slanted area and areas with different curvatures. Fig. 5.8 shows the sub-pixel disparity results and specific regions for the additional experiment. For the slanted region, the proposed method is slightly better than the parabolic fitting, as shown in Fig. 5.8(d) and (e). This is why the constant slope in such a slanted area does not provide the multiple fitting algorithm with more information.

Table 5.4 demonstrates that the proposed method shows overall better quantita-

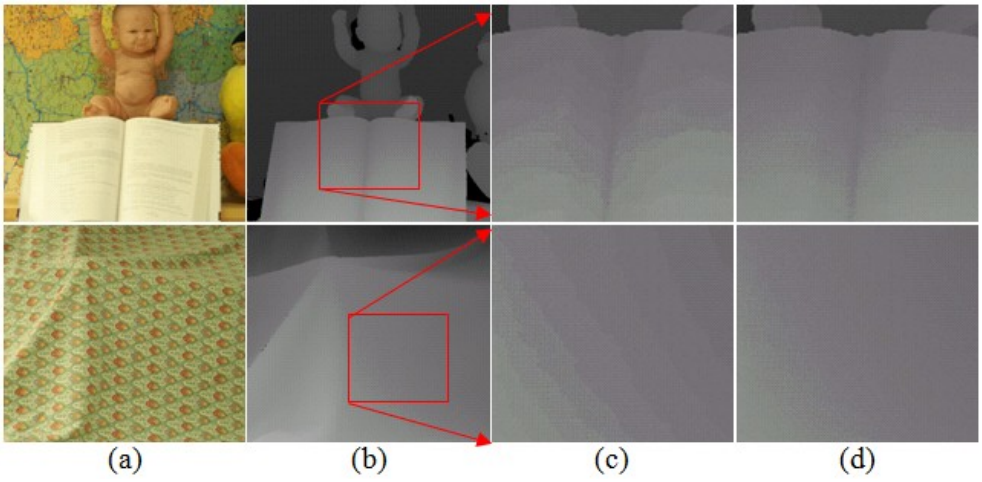

Figure 5.7: Integer vs. sub-pixel disparity. (a) Left image. (b) Ground truth. (c) Close-up of integer disparity. (d) Close-up of the Proposed.

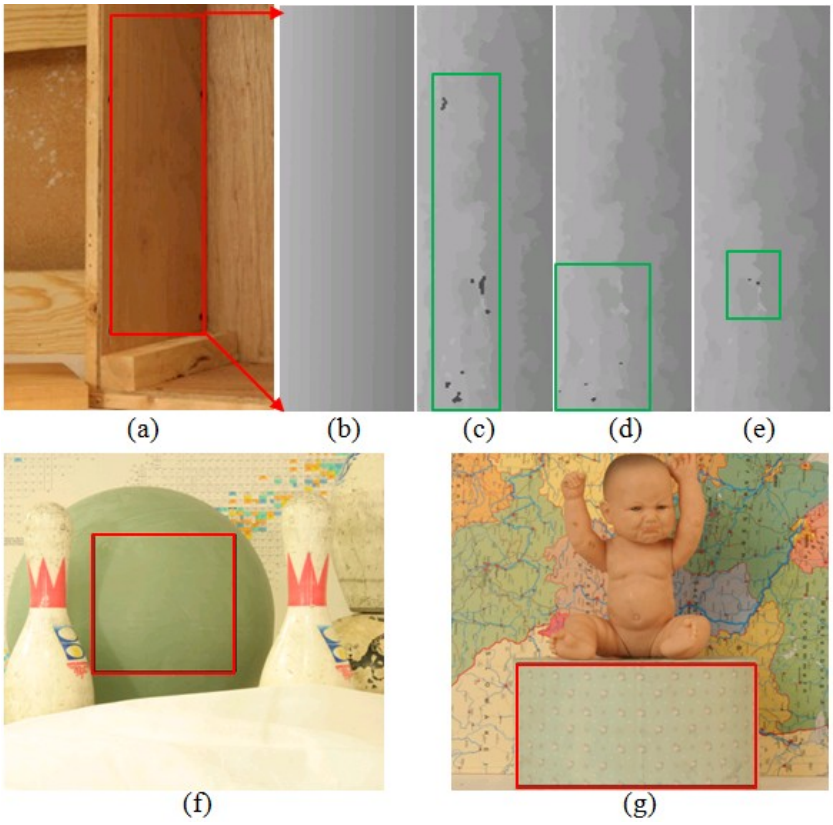

Figure 5.8: Specific areas and zoomed-in disparity maps (a) Left image including a slanted area (red box). (b) Ground truth. (c) Integer disparity. (d) Parabolic fitting. (e) Proposed multiple fitting. (f) Area with large curvature. (g) Area with small curvature.

Table 5.4: Sub-pixel performance evaluation (bad pixel (error) percentage in the non-occlusion area)

Dataset	Threshold	Integer	Parabolic	Multiple
Tsukuba	0.75	22.7%	12.8%	**12.6%**
	0.5	22.7%	22.5%	**20.2%**
Venus	0.75	0.44%	**0.17%**	0.22%
	0.5	5.42%	0.77%	**0.65%**
Teddy	0.75	6.52%	**5.92%**	5.93%
	0.5	11.4%	8.17%	**8.07%**
Cones	0.75	3.01%	2.28%	**2.25%**
	0.5	7.17%	3.88%	**3.72%**
Slanted in Fig. 5.8(a)	0.75	46.5%	42.3%	**41.9%**
	0.5	66.1%	61.7%	**61.6%**
Rounded in Fig. 5.8(f)	0.75	0.569%	0.085%	**0.076%**
	0.5	11.4%	3.35%	**2.46%**
Rounded in Fig. 5.8(g)	0.75	5.01%	2.95%	**2.67%**
	0.5	14.3%	6.94%	**6.73%**

tive performance on each image and surface type than do either the conventional sub-pixel or the integer method. Especially, the proposed method obtains noticeable performance gain (27% in threshold of 0.5) on the rounded surface area in Fig. 5.8(f) against the conventional one. It demonstrates that the proposed multiple fitting is more effective on disparity surface with various curvatures. We note that performance comparison with other sub-pixel algorithms is limited since the quality of the sub-pixel algorithms based on cost function depends mainly on how accurate the cost function is.

5.4.3 Stereo panoramic results and fusion effects

For the stereo panorama disparity estimation, we apply a multi-resolution scheme with a four-level pyramid, including the spatial-multi-resolution TV. The coarsest image size is 1024×512, which becomes basic size for partitioning. From level 1, we partition panoramic images and the partitioned images are overlapped by disparity range to avoid lack of correspondence at border areas. The partitioned disparity results are stitched for the original size disparity map. Throughout the experiment, the spatial-multi-resolution parameters are set to constant values: $\mu = 1$ and $(\beta_x, \beta y, \beta_s) = (1, 1, 2.5)$.

First, we examine the effectiveness of the disparity refinement reusing already-computed functions. Fig. 5.9 shows disparity maps before and after the refinement. As

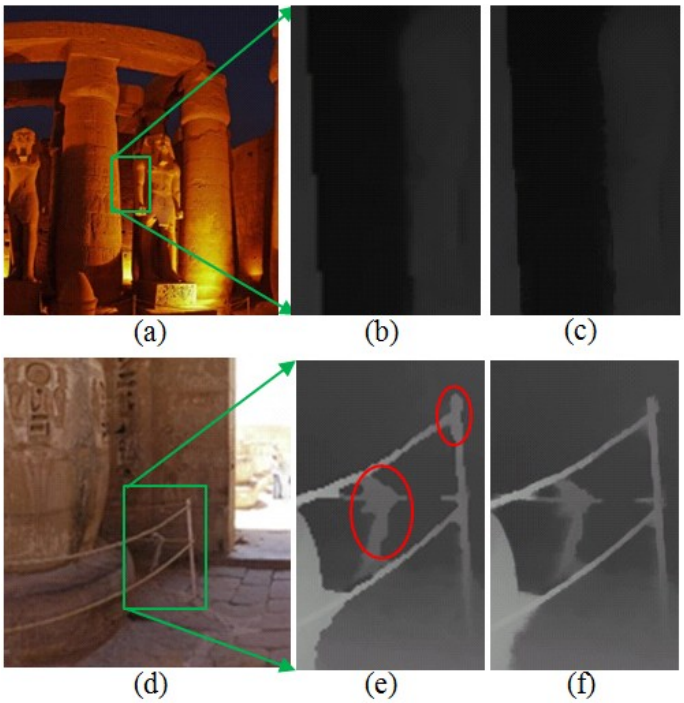

Figure 5.9: Disparity refinement results. (a) Left image 1. (b) Zoomed-in initial disparity map. (c) Zoomed-in refined disparity map. (d) Left image 2. (e) Zoomed-in initial disparity map. (f) Zoomed-in refined disparity map.

shown in Fig. 5.9(b) and (c), the blurring of sharp edges is reduced, and the wiggle edge is also recovered to be sharp after the refinement. The better-refined disparity near erroneous edges yields better guidance of the next level disparity estimation. Similarly, the erroneous thin structures, such as strings and poles, are refined with the aid of the color image and local edge weakness function $h(p)$, as shown in Fig. 5.9(f). However, large smearing errors in the red circle, which are likely to occur at thin structure with similar color background, cannot be refined completely. Note that the process of refining large and consistent disparity errors makes the refined result look blurred. In fact, such a blurring effect is equivalent to incomplete error reduction by attempts to reduce the large smearing errors. It is not an unwanted edge blurring from general upsampling process.

Fig. 5.10 illustrates how the proposed scheme improves the quality of disparity map for large panoramic views. Fig. 5.10(b) and (c) show the coarsest and finest level disparity maps from a single-resolution scheme, respectively. Fig. 5.10(d) depicts the initial sub-pixel disparity map where the surface is much smoother than Fig. 5.10(b).

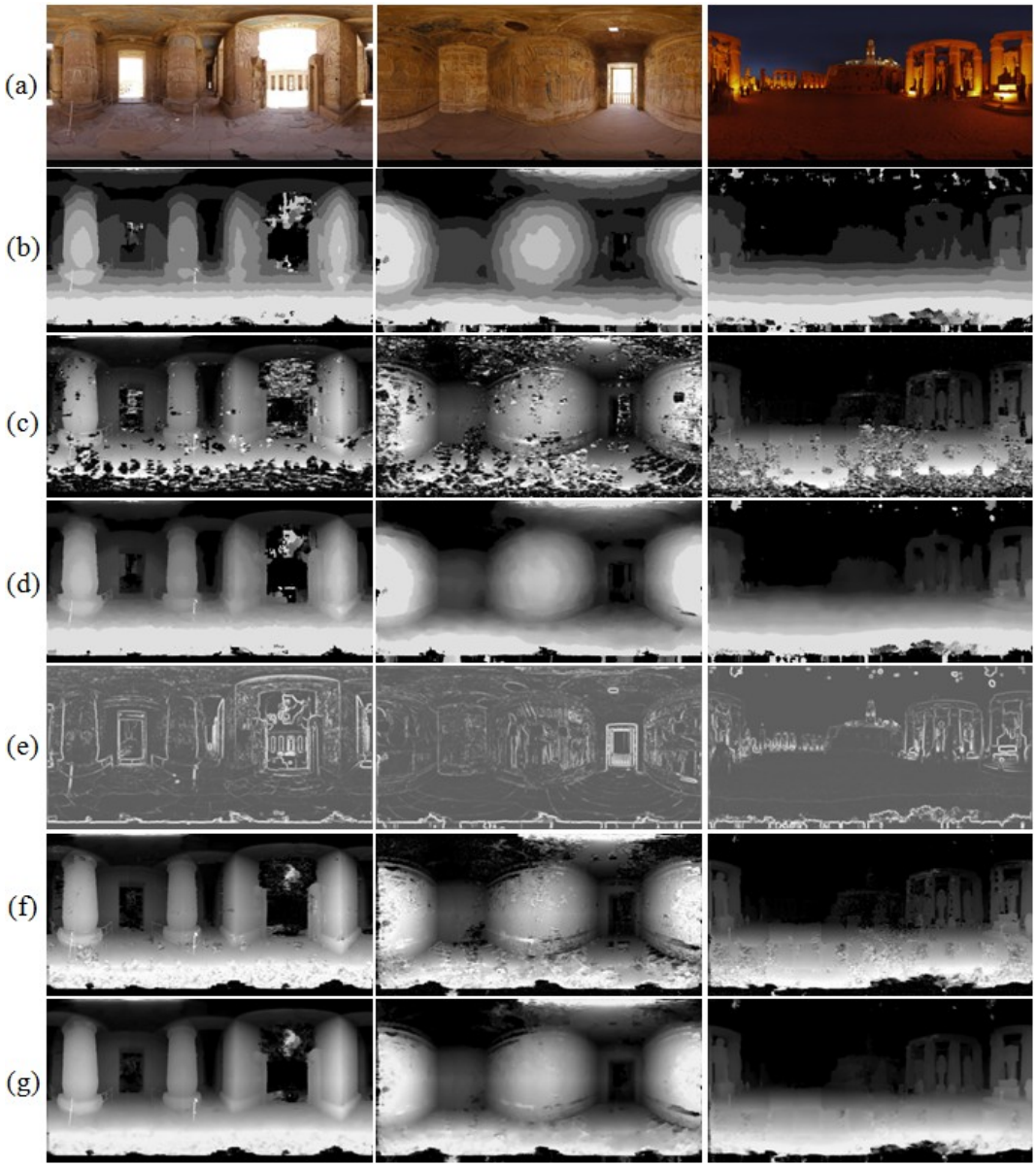

Figure 5.10: Intermediate disparity results. (a) Left panoramic image. (b) Coarsest integer disparity. (c) Finest integer disparity. (d) Initial sub-pixel disparity. (e) Local edge strength. (f) Guided finest sub-pixel disparity. (g) Final disparity fused by spatial-multi-resolution TV.

However, there are still some staircase disparity errors and poor details. Fig. 5.10(e) shows the local edge strength map, and Fig. 5.10(f) depicts the finest-scale disparity map guided by the refined disparity and search offset. As shown in Fig. 5.10(g), the final results show a significant improvement in disparity quality. Spatial and scaling

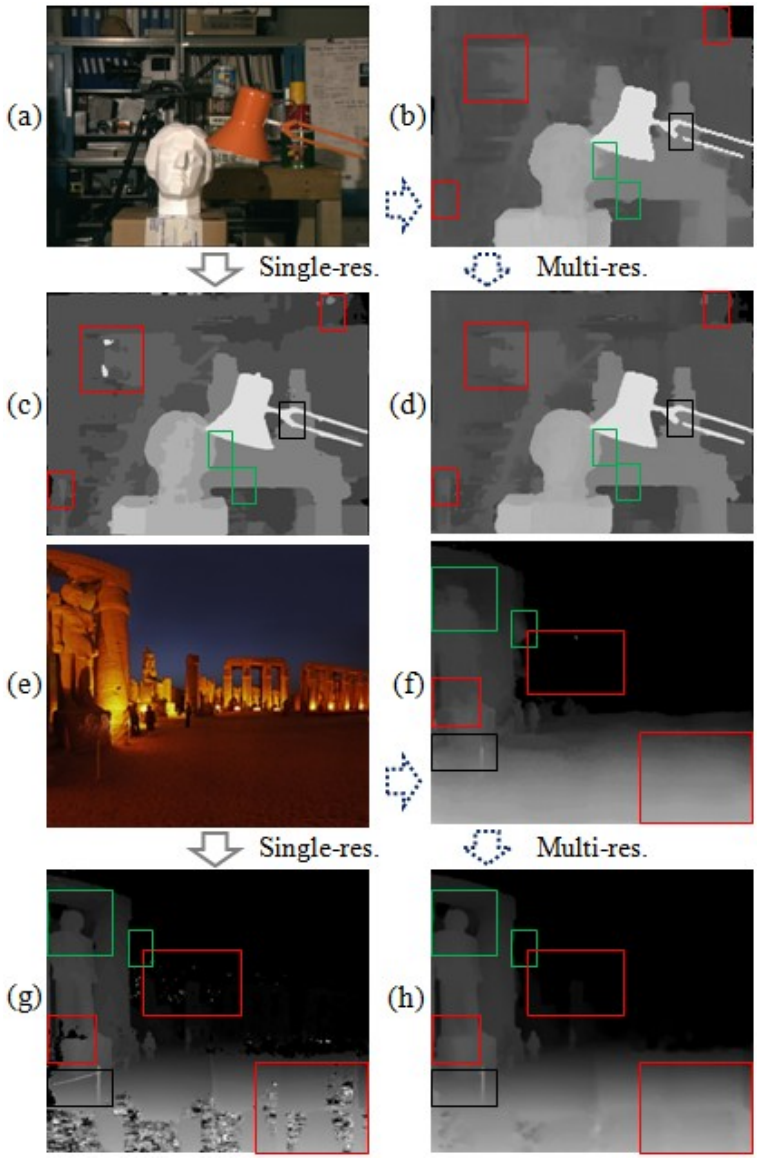

Figure 5.11: Fusion process. (a) Left image 1. (b) Initial disparity from multi-resolution. (c) Single-resolution disparity. (d) Final disparity. (e) Left image 2. (f) Initial disparity from multi-resolution. (g) Single-resolution disparity. (h) Final disparity.

errors are recovered by applying the spatial-multi-resolution TV, while fine details are preserved. This is a result of fusion from individual 4-level disparity maps including Fig. 5.10(d) and (f).

Fig. 5.11 illustrates the process of the fusion driven by the adaptive search offset and spatial-multi-resolution TV. Fig. 5.11(b) and (d) show the initial disparity and final disparity guided by the adaptive search offset Δd, respectively. Fig. 5.11(c) depicts

the disparity result from single-resolution scheme. In the red box, the smoothness of the initial disparity is propagated into the next level estimate by the adaptive search offset which should be small enough here, while the single-resolution scheme yields errors due to the matching ambiguities in these repetitive and homogeneous regions, as shown in Fig. 5.11(c). In the green box, the errors and lost details present in the initial disparity are recovered without the error propagation, with the aid of the adaptive search offset, which should be large enough. These are important achievements of the adaptive search offset. However, there is a limitation as shown in the black box where the initial disparity error propagates. It occurs only when the edge structure cannot be detected from both initial disparity and image intensity and at the thin structure, which is typically difficult to detect. From our observations, such cases are rare and highly dependent on the scene. The precise thin structure reconstruction in disparity estimation is still a difficult problem.

For real-world images, the spatial-multi-resolution TV, which is one of two fusion functions in the proposed scheme, is applied to enforce scaling consistency. Similarly, Fig. 5.11(h) shows the final disparity guided by the offset and then combined by the spatial-multi-resolution TV. The error suppression in the red box is more significant, and the recovery of lost details in the green box is also achieved. As a result, it is verified that the adaptive search offset and spatial-multi-resolution TV contribute to fusing the advantages taken from both coarse and fine level estimate.

Fig. 5.12 depicts final disparity maps of three hierarchical schemes on panoramic images: (1) RealtimeBP [36], which is a well-known global hierarchical method, (2) conventional hierarchical scheme using the local method [2], and (3) proposed hierarchical scheme. The RealtimeBP has a memory warning on such a large image test, as with most methods. We partition the large images up to half of the basic size we use. The RealtimeBP produces very noisy disparity maps that may result from the propagation failure during the global optimization process. The conventional scheme still shows a lot of errors along object boundaries and staircase errors on the surface. On the other hand, the proposed scheme shows the best quality of disparity map compared to the others.

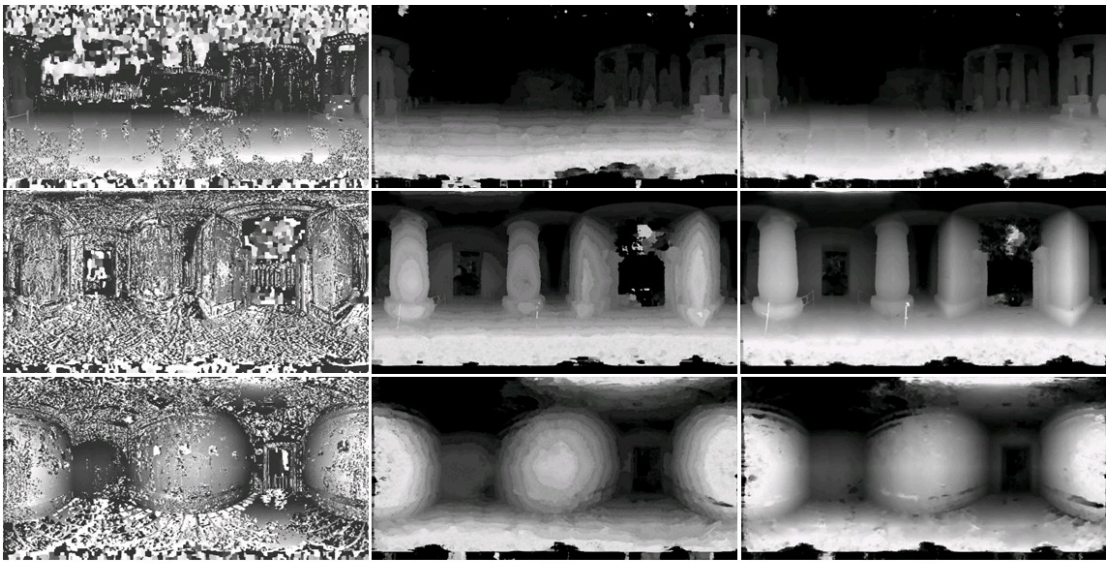

Figure 5.12: Disparity results on large images. (a) RealtimeBP. (b) Conventional hierarchical scheme. (c) Proposed.

5.5 Summary

It is challenging to estimate disparity for large stereo panoramas and moreover, achieve better accuracy and reduce complexity at the same time. To obtain high quality and high-resolution disparity maps from large-sized images, we propose an adaptively determined pixel-wise disparity search range, which is based on the combined eigenvalues of structure tensor matrices of image intensity and initial disparity. For the sub-pixel disparity, the multiple fitting algorithm extending the parabolic fitting is proposed to better represent rounded surfaces while alleviating the pixel locking effect. To enforce the spatial and scaling consistency, we use the spatial-multi-resolution TV method. Simulation results verify that the proposed method fuses the multi-resolution disparity maps effectively and, thereby, produces high quality disparity maps for large stereo panoramas. It will be an interesting research to extend the proposed algorithm to cylindrical or spherical panoramic views.

5.6 Acknowledgements

This chapter is in part a reprint of a conference paper presented in IEEE International Conference on Acoustics, Speech and Signal Processing, May 2014, and a reprint of a submitted paper to IEEE Transactions on Multimedia, 2014.

Chapter 6

Multi-array Camera Disparity Enhancement

In the previous chapters, we deal with disparity estimation for stereo image, video, and panoramic view. This chapter presents a disparity enhancement algorithm for multi-array camera system, which is applicable to any standard binocular stereo matching methods.

6.1 Introduction

Multiple array cameras have been recently developed with many features such as refocusing after taking the photo, focusing on multiple objects, and combining the multiple images over stereo camera. In particular, it might be able to provide accurate depth information of the captured scene, which is fundamental information for a wide range of 3D applications.

Multi-array camera systems have greater potential for 3D depth-based application development compared to stereo camera systems. However, there are very few research results on multi-array-based disparity estimation, due to lack of data. We provide synthetic multi-array images and videos created by 3DS MAX software as well as associated ground-truth disparity maps. In this chapter, we propose alternate use of local and global fusion of multi-array disparities to maximize the disparity enhancement in array camera system.

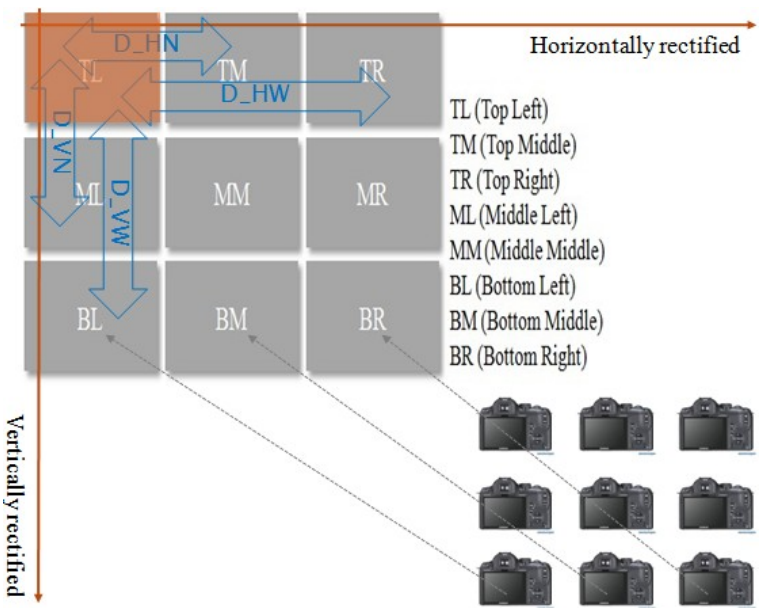

Figure 6.1: A 3×3 array camera model.

Fig. 6.1 shows a planar multi-array camera model with 3×3 array configuration as a basic setup. The array camera produces nine captured images (Top Left (TL), Top Middle (TM), etc.) which are rectified horizontally and vertically as shown in Fig. 6.1. In this camera setup, we can obtain four different disparity estimates for TL image: two horizontal disparity maps with narrow baseline (D_HN) and wide baseline (D_HW) and two vertical disparity maps with narrow baseline (D_VN) and wide baseline (D_VW). To enhance disparity accuracy in multi-array camera systems where there are multiple disparity estimates, fusion processes will be required, such as locally replacing unreliable pixels by valid pixels and globally optimizing multiple estimates over multi-dimension.

6.1.1 Related work

A stereo matching method using multiple stereo pairs is presented in [57, 58]. However, these methods are limited to a 1D array camera model in which cameras are displaced only in the horizontal direction. A multi-view stereo reconstruction algorithm is proposed for planar camera arrays, which can be 2×2 and larger [59]. This focuses on 3D surface reconstruction using a unique layered depth image. Various multi-array camera applications are introduced in [60], where various capabilities of a large number

of inexpensive cameras, such as high-dynamic range, high-resolution, high-speed video, and wide synthetic aperture, are presented.

An LRC check is introduced to detect invalid pixels on disparity maps based on a stereo image pair [61, 62]. The method has been applied to numerous stereo matching algorithms to ensure that both left and right disparities are consistent. This can be categorized as post-processing of disparity estimation, which fills occlusion error and mismatch with a valid value within a horizontal neighborhood. However, this simple filling strategy causes horizontal streaks [24]. To avoid the horizontal streaks and achieve better filling strategy, a local region voting approach, in which all reliable pixels within a neighborhood vote for the most reliable disparity candidate, is used in [63, 21, 22]. The unreliable pixel is replaced by the majority of the reliable pixels in voting region. However, this advanced filling strategy based on high-confidence voting is designed for refining only stereo disparity pair. Therefore, it should be extended to enhance multiple disparity pairs in array camera systems.

Image restoration such as denoising and deblurring is an inverse problem. Recovering an original image from a noisy image via an inverting process is ill posed since it does not have a unique solution and the solution is very sensitive to noise. To obtain a meaningful and stable solution, regularization is introduced. Two well-known regularization techniques are the Tikhonov regularization [64, 65, 66, 67, 68] and TV regularization [69, 70, 71, 72]. Tikhonov regularization tends to make images overly smooth [71], and TV regularization is known to be an advanced variational method. An augmented Lagrangian method with TV regularization [56] is proposed to enhance initial video disparity maps by combining them temporally and spatially. It results in enforcing spatio-temporal consistency on video disparity maps, which demonstrates that the TV regularization method is a good solution for refining multiple disparity maps. However, the method is memory intensive due to 3D regularization treating the video disparity as a space-time volume. In a multi-array camera system, there exists one more dimension, the "multi-array dimension," which comes from camera geometric difference. It is obvious that the memory problem will be worsened as the number of dimensions under consideration increases.

6.1.2 Contributions

The main contribution of this chapter is a new cascade regularization approach, which can better restore diagonal structures by further regularizing diagonal variations. Detailed analyses and experimental results are presented to verify the advantages of the cascade regularization against conventional approaches on images.

For multi-array camera disparity enhancement, a cascade TV regularization is proposed to globally combine complementary multiple disparities along array dimension and, thereby, achieve high performance gain.

We propose a multiple cross-filling algorithm to further obtain cross consistency between multiple estimates by locally replacing mismatched values by valid ones.

Synthetic multi-array images and video are created for simulation. In addition, the associated ground-truth disparity maps can be used for quantitative evaluation. Various experiment results show that the proposed twofold algorithm can enhance the initial disparity map up to about 50%, and possibly more, on poor initial disparity estimates.

6.1.3 Organization

This chapter is organized as follows. Section 6.2 describes the problem observations that we make. The details of our proposed method are presented in Section 6.3. Section 6.4 shows simulation results and discusses their significance. Section 6.5 concludes with some remarks.

6.2 Problem Observation

In a planar multi-array camera system, multiple cameras are displaced horizontally and vertically with narrow and wide baseline as shown in Fig. 6.1. For simulation, multi-array datasets (image, video, and disparity ground-truth map) are created by a 3D modeling and rendering software (e.g. 3DS MAX) since no ground-truth depth map from array camera is available. The array setup allows us to compute multiple disparity maps for one view with different scan-line directions and baseline lengths, using standard binocular stereo algorithms. For an $n \times n$ array camera system, $2 \times (n-1)$ disparity

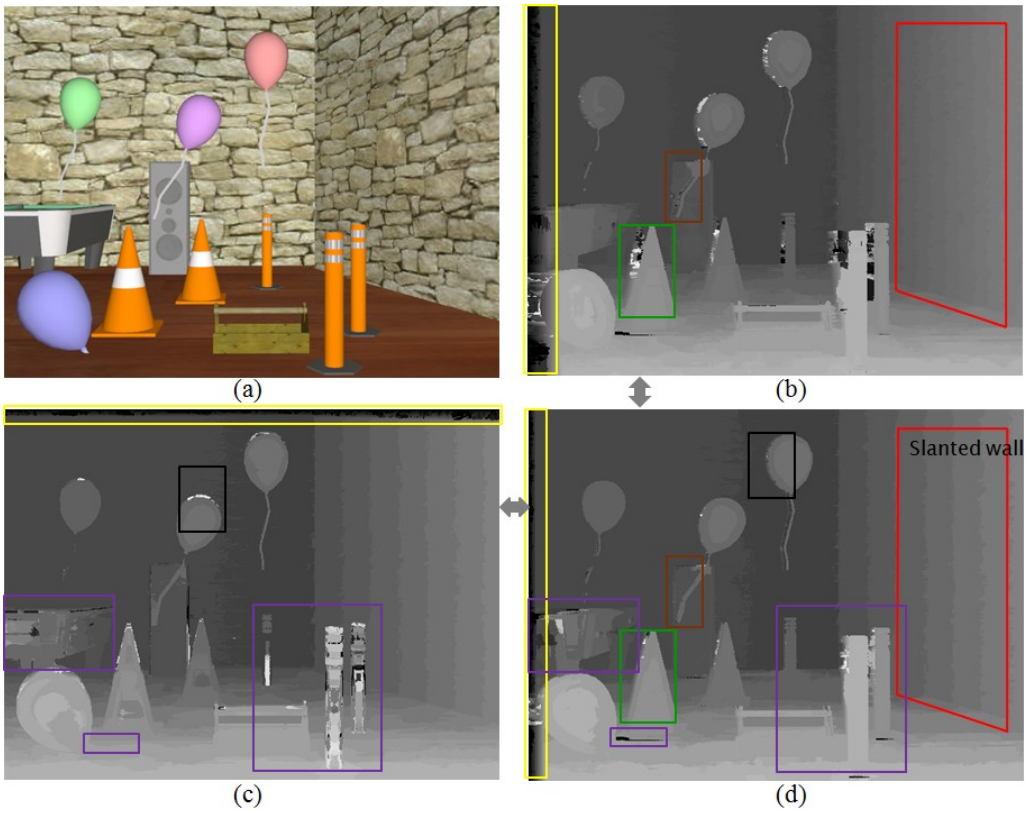

Figure 6.2: Three different disparity maps. (a) Top Left image. (b) Horizontal-wide baseline (D_HW). (c) Vertical-narrow baseline (D_VN). (d) Horizontal-narrow baseline (D_HN).

maps can be estimated individually as shown in Fig. 6.1.

Fig. 6.2 depicts a left image (Top Left) and three different estimates (D_HW, D_VN, and D_HN) calculated by the local stereo method "LM3C" [2]. The initially computed disparity maps reveal different characteristics according to the baseline and scan-line type. First, the narrow-baseline disparity in Fig. 6.2(d) shows fewer border errors (yellow box) and fewer occlusion errors (green box), comparing to wide-baseline disparity in Fig. 6.2(b). However, it shows poor thin structure representation (brown box) and noticeable staircase errors on the slanted surface (red box). Consequently, there are advantages and disadvantages. The advantages come from fewer corresponding pixel shifts and smaller occluded regions due to small displacement of two cameras. The disadvantages are due to low disparity precision. Disparity value (d_i) at pixel i is

defined as

$$d_i = \frac{BF}{Z_i} \tag{6.1}$$

where F is the camera focal length and Z_i is the real distance between camera and object. B is the baseline, which is the distance between two cameras. For the same object, the disparity range (precision) is proportional to only B since F is assumed to be fixed in the multi-array camera system. The narrow baseline makes the tracking of feature easier but less precise, whereas the wide baseline makes the matching more precise but there is greater possibility of false match due to the larger search area [57]. Second, we investigate how the horizontal and vertical scan-line matching affect the disparity result. The former causes vertical border and occlusion errors (yellow and black box in Fig. 6.2(d), respectively), whereas the latter results in horizontal border and occlusion errors, as shown in Fig. 6.2(c). They show very different error types (purple boxes) as in real world, structures on the same object look different at different view angles.

To summarize, multiple disparity estimates calculated individually contain complementary information with different type of errors that can be eliminated by utilizing the baseline and scan-line properties. It is necessary to fuse the complementary multiple estimates to enhance the initial disparities. A strategy of stereo matching on each pair followed by refinement process is inspired by these observations. In addition, such a separate process scheme gives us more flexibility for camera array extension than a combined process. Our goal is to develop an algorithm that effectively fuses multiple estimates while reducing the initial disparity errors as much as possible.

6.3 Proposed Method

6.3.1 Overall algorithm

Fig. 6.3 illustrates the proposed multi-array disparity enhancement algorithm based on a basic array model (3×3), which consists of two main functions: local fusion (multiple cross-filling) and global fusion (cascade TV regularization). After the binocular stereo matching algorithm is performed on each camera pair, the multiple

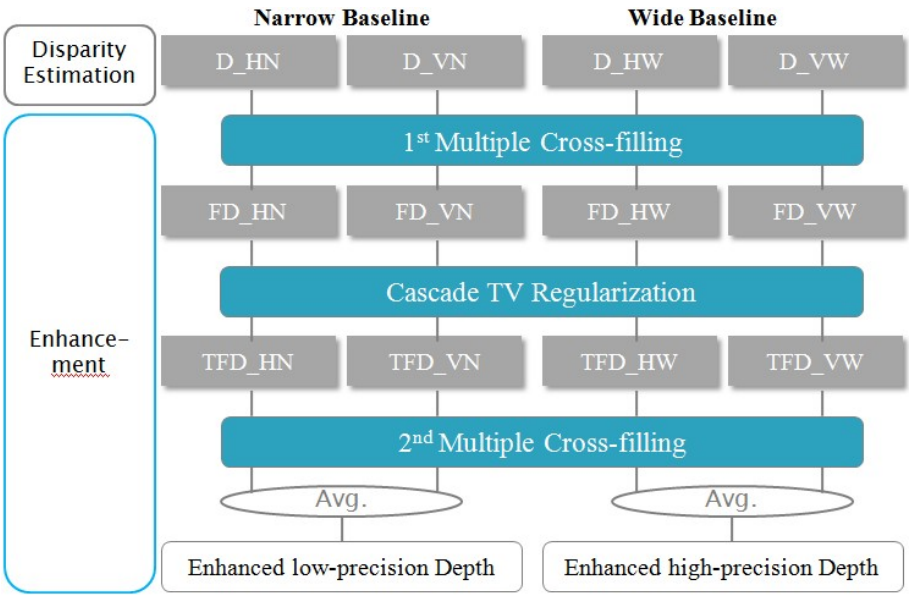

Figure 6.3: Functional diagram of proposed algorithm.

cross-filling is applied to the initial disparity estimates. To enforce array camera geometrical and spatial consistency, the cascade TV regularization is applied to the cross-filled disparity maps (FD_HN, FD_VN, etc.). To obtain further cross consistency, the multiple cross-filling is repeated on the TV-refined estimates (TFD_HN, TFD_VN, etc.). Alternately applying the local and global fusion is able to extend disparity enhancement limit that can be maximally achieved by the individual fusion. Two final disparity maps with low and high-precision, respectively, are constructed by a simple average function. The proposed algorithm can be easily extended to $n \times n$ array camera system and is applicable to any binocular stereo method.

6.3.2 Cascade regularization

TV regularization also known as TV denoising, is one of the most advanced variational methods for noise removal [69]. Based on the observation that noises with spurious detail have high total variation, the algorithm minimizes noise by reducing total variation while preserving important structures such as edges. In other words, it maximizes consistency over spatial dimension by reducing total variation along x and y directions, except for edges. The isotropic TV regularization problem is defined as in

[69]

$$\begin{aligned} \underset{f}{\text{minimize}} \ & \frac{1}{2}\sum_{i,j}(f_{i,j}-g_{i,j})^2 \\ & +\lambda\sum_{i,j}\sqrt{|f_{i+1,j}-f_{i,j}|^2+|f_{i,j+1}-f_{i,j}|^2} \end{aligned} \tag{6.2}$$

where $f_{i,j}$ and $g_{i,j}$ are the unknown and observed 2D signal at pixel location (i,j), respectively. λ is the regularization parameter that controls the relative emphasis. The solution (f) is obtained by minimizing the first discrete term (objective) and second discrete term (regularization) at the same time. It can be rewritten in the vector form for simple notation:

$$\underset{\mathbf{f}}{\text{minimize}} \ \ \frac{1}{2}||(\mathbf{f}-\mathbf{g})||_2^2+\lambda||\mathbf{D}_{x,y}\mathbf{f}||_{TV} \tag{6.3}$$

where operator $\mathbf{D}_{x,y}$ is defined as a collection of suboperators $\mathbf{D}_{x,y}=[\mathbf{D}_x^T \ \ \mathbf{D}_y^T]^T$. $\mathbf{D}_x$ and $\mathbf{D}_y$ are the first-order forward finite-difference operators along x-axis and y-axis, respectively. Since the traditional regularization always considers 2D (x,y) variations simultaneously, as in (6.2), the algorithm can be referred to as simultaneous regularization.

We utilize a cascade approach in which x-directional regularization is followed by one in the y direction. This setup also prepares the overall algorithm for the memory efficiency issue expected in the multi-array camera system. Note that such a minimization problem is not separable, unlike 2D Gaussian convolution, which is separable. Convolving along the x-axis and then y-axis yields exactly same solution as that of the 2D simultaneous convolution. The new approach might produce a different solution. Therefore, we need to verify the difference and advantage against the conventional (simultaneous) approach.

We analyze how the cascade regularization approach affects the result of image restoration. The Tikhonov regularization is used since the L-2 norm of the Tikhonov regularization has a closed-form solution [69, 73] whereas the L1 and TV norm do not. It is straightforward to generalize from 2D to 3D and for general case as well. A 2D Tikhonov regularized least square problem is set up as

$$\underset{\mathbf{f}}{\text{minimize}} \ \ ||(\mathbf{f}-\mathbf{g})||_2^2+\lambda||\mathbf{D}_{x,y}\mathbf{f}||_2^2. \tag{6.4}$$

To solve this problem, we take the derivative and set it to 0:

$$2(\mathbf{f}-\mathbf{g})+2\lambda(\mathbf{D}_{x,y}^{T}\mathbf{D}_{x,y}\mathbf{f})=0. \tag{6.5}$$

The Pseudo-inverse solution is:

$$\hat{\mathbf{f}}=[\mathbf{I}+\lambda(\mathbf{D}_{x}^{T}\mathbf{D}_{x}+\mathbf{D}_{y}^{T}\mathbf{D}_{y})]^{-1}\mathbf{g}. \tag{6.6}$$

On the other hand, a new approach with cascade regularization consists of two minimization problems:

$$\begin{aligned}\underset{\bar{\mathbf{f}}}{\text{minimize}} \quad & ||(\bar{\mathbf{f}}-\mathbf{g})||_2^2+\lambda||\mathbf{D}_x\bar{\mathbf{f}}||_2^2\\ \underset{\mathbf{f}}{\text{minimize}} \quad & ||(\mathbf{f}-\bar{\mathbf{f}})||_2^2+\lambda||\mathbf{D}_y\mathbf{f}||_2^2\end{aligned} \tag{6.7}$$

where the solution of the first problem is used as initial value for the second problem in order to find a final solution. The intermediate solution ($\bar{\mathbf{f}}$) is found similarly to (6.5) and (6.6)

$$\bar{\mathbf{f}}=[\mathbf{I}+\lambda\mathbf{D}_{x}^{T}\mathbf{D}_{x}]^{-1}\mathbf{g}. \tag{6.8}$$

The final solution is

$$\hat{\mathbf{f}}=[\mathbf{I}+\lambda(\mathbf{D}_{x}^{T}\mathbf{D}_{x}+\mathbf{D}_{y}^{T}\mathbf{D}_{y})+\lambda^{2}\mathbf{D}_{y}^{T}\mathbf{D}_{y}\mathbf{D}_{x}^{T}\mathbf{D}_{x}]^{-1}\mathbf{g}. \tag{6.9}$$

The difference between the proposed and conventional approaches can be examined by comparing (6.6) and (6.9). If λ^2 is taken out from the brackets in two solutions, the only difference becomes $\mathbf{D}_{y}^{T}\mathbf{D}_{y}\mathbf{D}_{x}^{T}\mathbf{D}_{x}$ added in the cascade solution. We analyze the difference from two different perspectives as follows.

Fourier transform perspective

The forward difference operators can be expressed as convolutions:

$$\begin{aligned}\mathbf{D}_{x}\mathbf{f}&=[1\ -1]*f\\ \mathbf{D}_{y}\mathbf{f}&=[1\ -1]^{'}*f\end{aligned} \tag{6.10}$$

where $'$ and $*$ denote vector transpose and convolution, respectively. Convolution is a linear operation, so it can be expressed by circulant matrix, which can be diagonalized

using Discrete Fourier Transform (DFT) matrices [74, 75]:

$$\begin{aligned}\mathbf{D}_x\mathbf{f} &= \mathbf{W}^T\mathbf{\Lambda_x}\mathbf{W}\mathbf{f} \\ \mathbf{D}_x^T\mathbf{D}_x\mathbf{f} &= \mathbf{W}^T\mathbf{\Lambda_x^2}\mathbf{W}\mathbf{f} \\ \mathbf{D}_y^T\mathbf{D}_y\mathbf{D}_x^T\mathbf{D}_x\mathbf{f} &= \mathbf{W}^T\mathbf{\Lambda_y^2}\mathbf{\Lambda_x^2}\mathbf{W}\mathbf{f}\end{aligned} \tag{6.11}$$

where $\mathbf{W}$ and $\mathbf{\Lambda_x}$ are DFT matrix and a diagonal matrix having the Fourier coefficients of $\mathbf{D}_x$, respectively. The first equation in (6.11) can be interpreted as follows. The last term ($\mathbf{Wf}$) represents the Fourier transform of $\mathbf{f}$. The second term ($\mathbf{\Lambda_x}$) operates as a scaling matrix that scales signals showing characteristics similar to the horizontal difference operator $\mathbf{D}_x$. The first term ($\mathbf{W}^T$) represents the inverse Fourier transform. Consequently, $\mathbf{W}^T\mathbf{\Lambda_x}\mathbf{Wf}$ amplifies the signals with horizontal variation in $\mathbf{f}$ by the amount of corresponding component of $\mathbf{\Lambda_x}$. The solution of the conventional regularization in (6.6) can be rewritten in the Fourier transform using (6.11) as

$$\hat{\mathbf{f}} = \mathbf{W}^T[\mathbf{I} + \lambda(\mathbf{\Lambda_x^2} + \mathbf{\Lambda_y^2})]^{-1}\mathbf{W}\mathbf{g}. \tag{6.12}$$

The solution of the cascade regularization in (6.9) can be rewritten as

$$\hat{\mathbf{f}} = \mathbf{W}^T[\mathbf{I} + \lambda(\mathbf{\Lambda_x^2} + \mathbf{\Lambda_y^2}) + \lambda^2(\mathbf{\Lambda_y}\mathbf{\Lambda_x})^2]^{-1}\mathbf{W}\mathbf{g}. \tag{6.13}$$

The new term $(\mathbf{\Lambda_y}\mathbf{\Lambda_x})^2$ is added in the inverse function, compared to (6.12). It reduces diagonal variation similar to the operations of $\mathbf{\Lambda_x^2}$ and $\mathbf{\Lambda_y^2}$ on horizontal and vertical variations, respectively. As a result, the cascade approach is able to regularize the diagonal variations better than the conventional approach due to the new term ($(\mathbf{\Lambda_y}\mathbf{\Lambda_x})^2$) added in the inverse function.

Filtering perspective

The first-order difference operators in (6.10) can be extended to high-order difference operators as follows:

$$\begin{aligned}\mathbf{D}_x^T\mathbf{D}_x\mathbf{f} &= [-1 \quad 2 \quad -1] * f \\ \mathbf{D}_y^T\mathbf{D}_y\mathbf{D}_x^T\mathbf{D}_x\mathbf{f} &= \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix} * f\end{aligned} \tag{6.14}$$

where $\mathbf{D}_x^T\mathbf{D}_x$ operates as the second-order difference filter along horizontal direction. $\mathbf{D}_y^T\mathbf{D}_y\mathbf{D}_x^T\mathbf{D}_x$ operates as a high-pass filter along diagonal direction. The inverse term in (6.6) represents combination of three filters: identity ($\mathbf{I}$), second-order horizontal ($\mathbf{D}_x^T\mathbf{D}_x$), and second-order vertical filter ($\mathbf{D}_y^T\mathbf{D}_y$). On the other hand, the inverse term in the cascade solution in (6.9) also includes a high-pass filter ($\mathbf{D}_y^T\mathbf{D}_y\mathbf{D}_x^T\mathbf{D}_x$) along diagonal direction. Since the term is inverted, signals with diagonal variation can be further filtered out in the cascade approach.

To solve large-scale Tikhonov minimization problem, we use an iterative method based on the augmented Lagrangian method [71, 72] and Alternating Direction Method (ADM) [76, 77]. We introduce intermediate variables $\mathbf{u}$ and transform problem (6.4) into an equivalent constrained problem:

$$\begin{aligned} &\underset{\mathbf{f},\mathbf{u}}{\text{minimize}} \;\; ||(\mathbf{f}-\mathbf{g})||_2^2 + \lambda||\mathbf{u}||_2^2 \\ &\text{subject to} \;\; \mathbf{u} = \mathbf{D}_{x,y}\mathbf{f}. \end{aligned} \tag{6.15}$$

The augmented Lagrangian of problem (6.15) is:

$$\begin{aligned} L(\mathbf{f},\mathbf{u},\mathbf{y}) =& ||\mathbf{f}-\mathbf{g}||_2^2 + \lambda||\mathbf{u}||_2^2 \\ &- \mathbf{y}^T(\mathbf{u}-\mathbf{D}_{x,y}\mathbf{f}) + \frac{\rho_r}{2}||\mathbf{u}-\mathbf{D}_{x,y}\mathbf{f}||_2^2 \end{aligned} \tag{6.16}$$

where $\mathbf{y}$ is the Lagrangian multiplier associated with the constraint $\mathbf{u} = \mathbf{D}_{x,y}\mathbf{f}$ and ρ_r is a regularization parameter associated with the quadratic penalty term $||\mathbf{u}-\mathbf{D}_{x,y}\mathbf{f}||_2^2$. ADM is used to find the minimum of $L(\mathbf{f},\mathbf{u},\mathbf{y})$. The following subproblems are solved iteratively:

$$\begin{aligned} \mathbf{f}_{k+1} =& \underset{\mathbf{f}}{\text{argmin}} \;\; ||(\mathbf{f}-\mathbf{g})||_2^2 \\ &- \mathbf{y}_k^T(\mathbf{u}_k-\mathbf{D}_{x,y}\mathbf{f}) + \frac{\rho_r}{2}||\mathbf{u}_k-\mathbf{D}_{x,y}\mathbf{f}||_2^2 \\ \mathbf{u}_{k+1} =& \underset{\mathbf{u}}{\text{argmin}} \;\; \lambda||\mathbf{u}||_2^2 \\ &- \mathbf{y}_k^T(\mathbf{u}-\mathbf{D}_{x,y}\mathbf{f}_{k+1}) + \frac{\rho_r}{2}||\mathbf{u}-\mathbf{D}_{x,y}\mathbf{f}_{k+1}||_2^2 \\ \mathbf{y}_{k+1} =& \mathbf{y}_k - \rho_r(\mathbf{u}_{k+1}-\mathbf{D}_{x,y}\mathbf{f}_{k+1}). \end{aligned} \tag{6.17}$$

The first and second problems are solved similarly to (6.5) and (6.6). The multiplier $\mathbf{y}$ is updated as in (6.17).

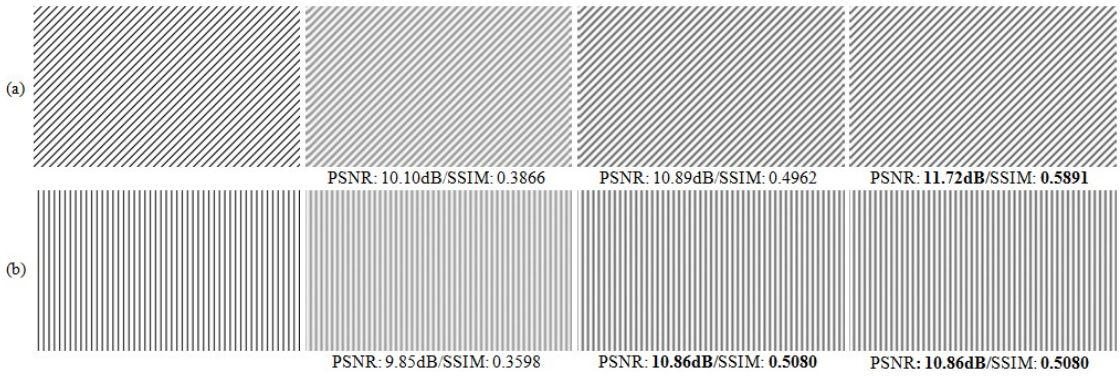

Figure 6.4: Conventional vs. cascade regularization on synthetic images. (First column) Original. (Second column) Blurred. (Third column) Conventional. (Fourth column) Cascade. (a) Diagonal streak. (b) Vertical streak.

To verify that the cascade approach regularizes the diagonal variation better, we perform experiments on synthetic and real-world images. In the simulation, the parameters ($\lambda = 0.001, \rho_r = 2$) are fixed and images are blurred by a Gaussian blur kernel of size 9×9 and $\sigma = 2$. Gaussian noise, distributed as $\mathcal{N}(0, 10^{-5})$ is added to the images. The blurred and noisy images are restored by the conventional and cascade Tikhonov regularization method individually, defined in (6.4) and (6.7), respectively.

Fig. 6.4 depicts the restored results of two synthetic images with Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) index. For the Diagonal streak image, the cascade regularization approach exhibits much better PSNR and SSIM than the conventional approach. On the other hand, for the Vertical streak image, the two approaches achieve the same PSNR and SSIM results. These experimental results demonstrate that the cascade approach can handle diagonal structures better than the conventional one. Consequently, the proposed approach achieves better restoration performance on diagonal structures while showing the same performance on vertical and horizontal structures. These simulation results agree with the analysis discussed previously. To confirm the advantage of the cascade approach, we perform the same experiment on real-world images (Barbara, Salesman). As shown in Fig. 6.5, the cascade regularization performs better than the conventional regularization, especially on the grid texture area (red box). However, note that the performance gain might depend on the scene structure. We will show more experimental results on this issue in Section 6.4.4.

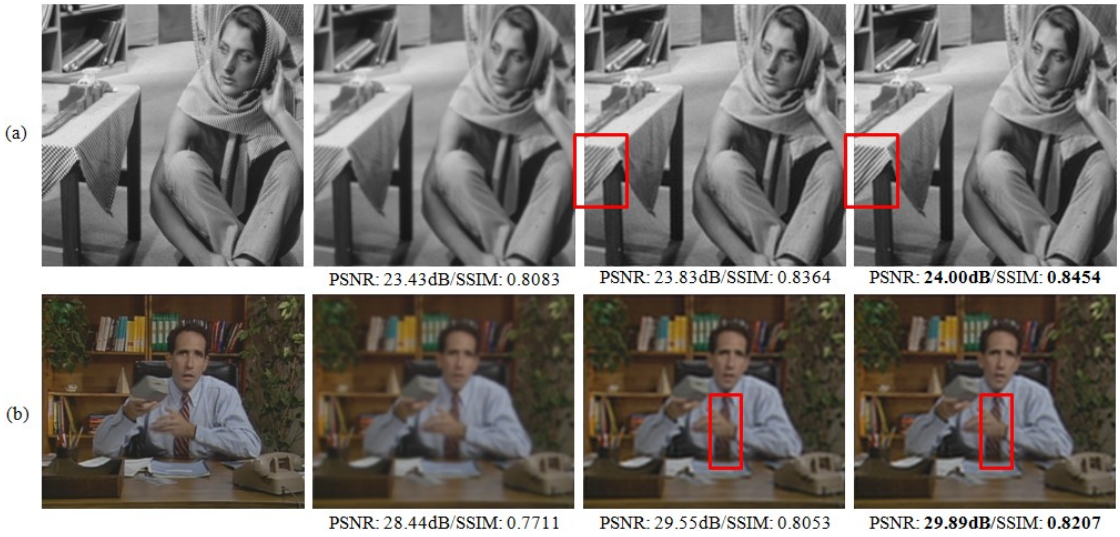

Figure 6.5: Conventional vs. cascade regularization on real-world images. (First column) Original. (Second column) Blurred. (Third column) Conventional. (Fourth column) Cascade. (a) Barbara. (b) Salesman.

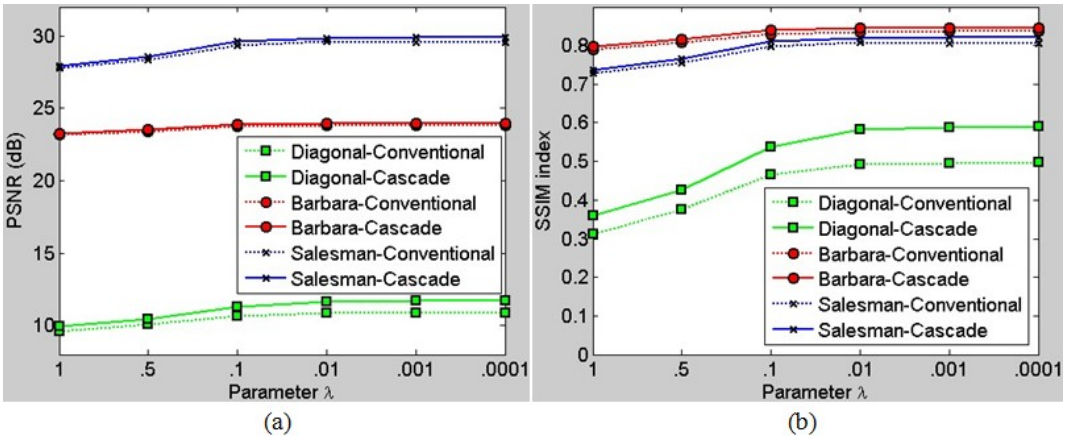

Figure 6.6: Regularization parameter sensitivity. (a) PSNR. (b) SSIM.

The influence of the regularization parameter λ on two approaches is examined. Fig. 6.6 illustrates that the cascade regularization approach achieves consistently better performance than the conventional one in terms of both quality metrics (PSNR, SSIM) regardless of parameter (λ). However, the performance gain is different on each input image. On the Barbara image, it shows almost the same performance, while large difference is shown on the Diagonal streak image.

Table 6.1 shows complexity evaluation of the two regularizations. Overall, the

Table 6.1: Complexity evaluation of two regularizations. Complexity is the ratio of computation time of the cascade approach to that of the conventional one.

Dataset	Conventional	Cascade	Complexity
Diagonal	**1.576s**	2.172s	1.38
Vertical	0.648s	**0.621s**	0.96
Barbara	**0.805s**	0.987s	1.23
Salesman	**1.015s**	1.236s	1.22

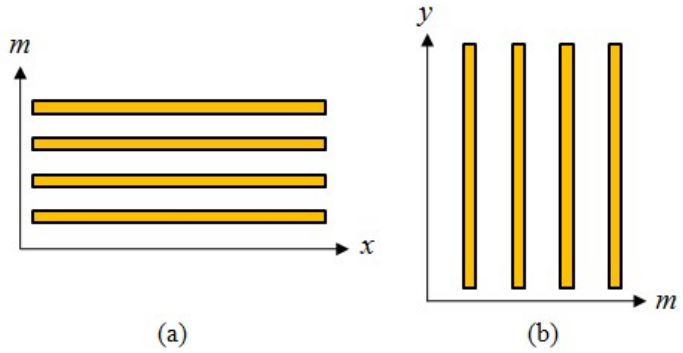

Figure 6.7: Dimension coupling. (a) (x, m). (b) (y, m).

cascade approach is slower than the conventional one because it consists of two minimization problems. However, the computation time is not double - it is at most $\times 1.38$ since the cascade approach uses simple 1D regularization instead of 2D regularization. It is meaningful to note that it is even faster on the Vertical streak image. In addition, it is evident that the cascade approach has better memory efficiency by considering just 1D variation.

We adapt a cascade TV regularization to multi-array camera system for disparity enhancement along multi-array dimension as well as spatial dimension. We have demonstrated that the spatial cascade regularization reconstructs complicated structures better. In the multi-array model, the additional array dimension is too small to be treated as an independent dimension. For instance, there are only six points (estimates) along multi-array dimension in a 4×4 array camera system. Because of these factors, we couple the horizontal (x) and vertical dimension (y) with the array dimension (m) as (x, m) and (y, m) shown in Fig. 6.7. In the cascade regularization, (x, m) and (y, m) are each treated as a single volume.

The cascade TV minimization problem for multi-array camera system is defined as

$$\begin{aligned} \underset{\bar{\mathbf{f}}}{\text{minimize}} \quad & \mu||(\bar{\mathbf{f}}-\mathbf{g})||_1 + ||\mathbf{D}_{x,m}\bar{\mathbf{f}}||_{TV} \\ \underset{\mathbf{f}}{\text{minimize}} \quad & \mu||(\mathbf{f}-\bar{\mathbf{f}})||_1 + ||\mathbf{D}_{y,m}\mathbf{f}||_{TV} \end{aligned} \tag{6.18}$$

where μ is a regularization parameter. A new parameter (μ) is introduced to avoid confusion since λ is associated with the Tikhonov minimization problem. Analogously, the intermediate solution ($\bar{\mathbf{f}}$) is fed back to the next minimization problem. For reasons why TV regularized L1 minimization is used for disparity refinement, the reader can refer to [56]. The first-order horizontal forward difference operator is defined as $\mathbf{D}_x\mathbf{f} = \mathbf{vec}(f(x+1,m) - f(x,m))$ where $\mathbf{vec}(\cdot)$ denotes the vectorization operator. Generally, it can be justified as a regularization factor, based on the assumption that disparity varies smoothly along horizontal and vertical direction since disparity is a piecewise constant function, except for sharp edges [56]. However, multiple disparity estimates along array dimension (m) are supposed to be the same even though they come from different geometric matchings. Thus, the multi-array difference operator ($\mathbf{D}_m$) should be redefined as

$$\mathbf{D}_m\mathbf{f} = \mathbf{vec}(\sum_{m^* \neq m} f(x,m^*) - f(x,m)). \tag{6.19}$$

Total difference over array dimension is used instead of the forward difference ($\mathbf{D}_x$ or $\mathbf{D}_y$).

6.3.3 Multiple cross-filling

In the previous section, we propose fusion of multiple disparity estimates by solving the cascade regularization minimization problem. To maximize disparity enhancement, we perform a multiple cross-filling before and after the cascade regularization. The first cross-filling provides the cascade regularization processor with multiple estimates that are more cross-consistent. The second cross-filling further enforces the cross consistency on the TV-refined disparity estimates. The multiple cross-filling achieves cross consistency in the manner of locally replacing pixels, whereas the cascade

TV regularization achieves consistency by globally optimizing variation along spatially coupled array dimensions.

First, it is important to detect invalid pixels having different disparities on the multiple disparity maps. We perform multiple cross-checks to detect invalid pixels, comparing the other corresponding array estimates. The invalid pixel is determined by a majority decision, and then an invalid pixel index map is constructed. Once an invalid pixel is detected, the algorithm searches in four directions (up, down, left, and right) in order to find valid pixels ($p_u, p_d, p_l,$ and p_r) and generates a valid pixel set. The most valid pixel in the valid pixel set is selected under two conditions: (1) high color similarity between the invalid pixel under consideration and valid pixel, and (2) the smallest disparity based on the fact that background pixels are occluded. There are three differences from the conventional region voting method: (1) multiple cross-checking, (2) four-directional search for finding valid pixels, and (3) color similarity condition. The algorithm procedure is as follows:

Algorithm:

1: Multiple cross-check for invalid pixel set P_I.

2: **if** pixel $p \in P_I$ **do**

3: Generate a valid pixel set $P_V = \{p_u, p_d, p_l, p_r\}$.

4: Select a pixel (p_s) in P_V which has similar color to p and has minimum disparity value.

5: Construct a neibhborhood for p_s based on the cross-based aggregation method [21].

6: Vote and the majority value is filled

7: **endif**

Table 6.2: Overall performance evaluation of the proposed algorithm on multi-array images with bad pixel % (threshold of 1 on all regions).

Dataset	Initial	1^{st} MCF	CTV	2^{nd} MCF	Avg.	Enhance.
Room	22.0833	18.2604	15.1582	10.2650	**10.2585**	53%
Cones	22.5700	19.3480	14.9958	13.0755	**13.0729**	42%
Bike	29.7653	26.3929	21.6810	20.1058	**20.1016**	32%

6.4 Results

For simulation, we create three multi-array images and video using 3DS MAX. For a quantitative evaluation, we provide a disparity ground-truth map converted from depth information encoded by 3DS MAX for each synthetic image. To simulate a real camera environment, we add camera noise which is known to be dependent on the pixel intensity level and whose variance is proportional to intensity [18, 19]. As a basic binocular stereo matching algorithm, we use LM3C [2]. We will show the robustness of the proposed algorithm to other disparity methods. There is no multi-array disparity algorithm available for performance comparison.

6.4.1 Overall performance on multi-array camera system

We apply the proposed multi-array disparity enhancement algorithm to the 3×3 camera system. Fig. 6.8 shows three multi-array images and one video created at the top left camera position. The multi-array dataset with associated disparity maps are available on our project website (http://videoprocessing.ucsd.edu/∼zucheul/multi-array.html) so that other researchers can use them for comparison. The proposed algorithm consists of three main steps: 1^{st} Multiple Cross-Filling (MCF), Cascade TV (CTV), and 2^{nd} MCF, as shown in Fig. 6.3. By evaluating each process individually, we show the step-by-step performance enhancement result. The regularization parameter μ in (6.18) is set to 1. Table 6.2 shows that the bad pixel rate decreases consistently as each step is performed. For the Room array image, we obtain the performance gain of 53% compared to the initial disparity estimates. The initial disparity estimates are computed by LM3C [2], which is among top local methods. It is meaningful that the high gain is

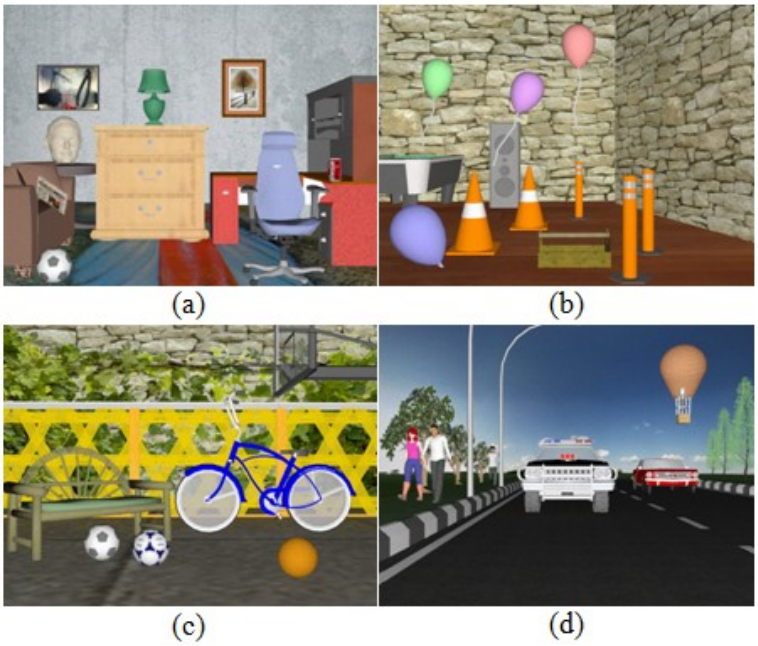

Figure 6.8: Multi-array images and video (at Top Left (TL) position). (a) Room. (b) Cones. (c) Bike. (d) Cars video.

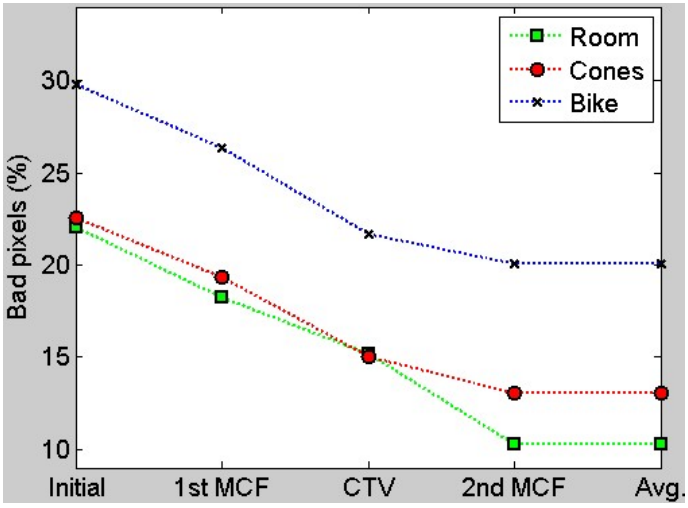

Figure 6.9: Performance graph of the proposed algorithm.

further achieved from the reasonable initial disparity estimates.

Fig. 6.9 illustrates performance improvement of each step in the proposed algorithm. This graph demonstrates that main functions for disparity enhancement are MCF and CTV. The average function is a simple step to make a representative dispar-

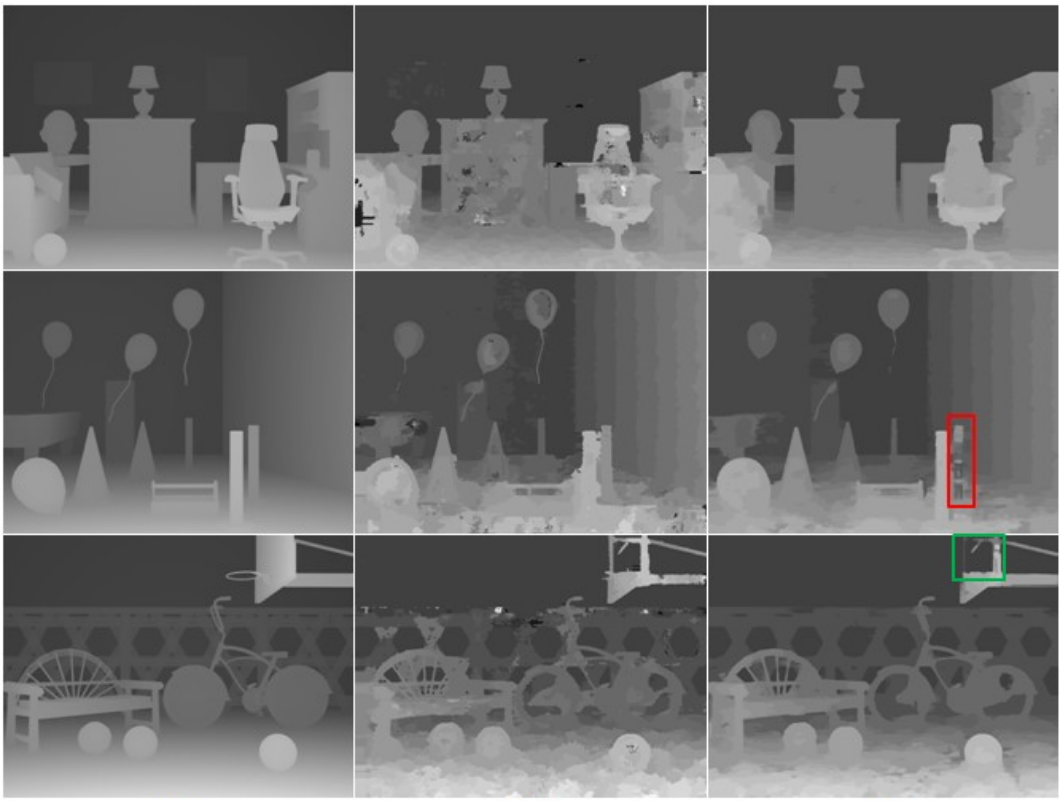

Figure 6.10: Multi-array image disparity maps (First row) Room. (Second row) Cones. (Third row) Bike. (a) Ground truth. (b) Initial. (c) Enhanced disparity maps.

ity map for low and high-precision results, as shown in Fig. 6.3. It does not contribute significantly to performance improvement.

Fig. 6.10 shows ground-truth, initial, and enhanced disparity maps. Disparity error reduction is significant, as shown in Fig. 6.10(b) and (c). However, there are two limitations. The first limitation (as indicated in the red box) is the fused error, where multiple estimates are fused, with a majority of the estimates containing large error. This is likely to occur on the thin structures with homogeneous texture. The second limitation (indicated by the green box) is caused by a transparent object, such as glass, which poses a challenging problem in stereo matching research.

Fig. 6.11 shows ground-truth, initial, and enhanced disparity maps on 5 consecutive array video frames. The initial video disparity quality is enhanced noticeably by the proposed algorithm. The initial disparity estimate at a certain point varies along

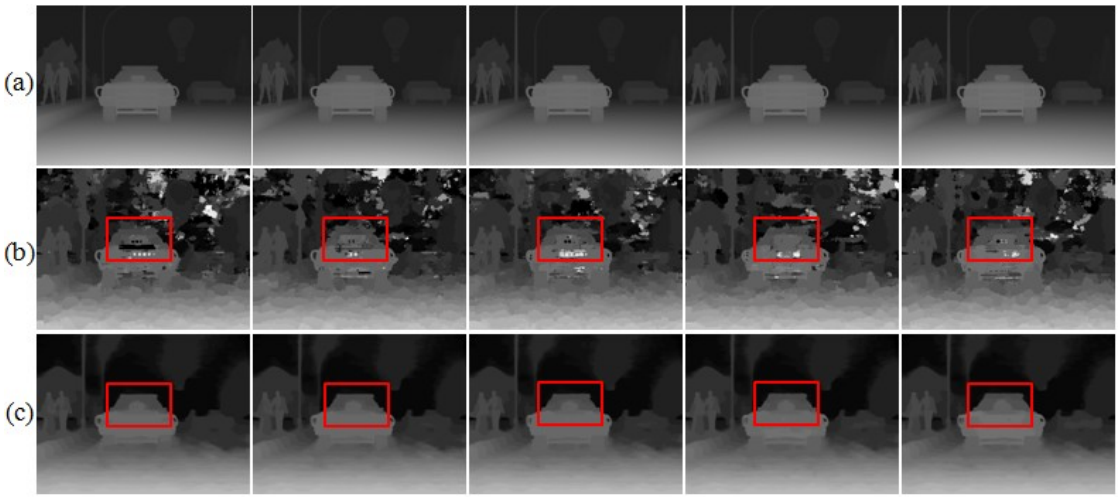

Figure 6.11: Multi-array video disparity results on five consecutive frames. (a) Ground truth. (b) Initial. (c) Enhanced.

the time axis, as shown in the red boxes in Fig. 6.11 (flickering), whereas the enhanced disparity estimate does not. This demonstrates that the proposed algorithm is able to achieve temporal consistency for the array video disparities. For temporal consistency in multi-array video system, the cascade TV regularization is defined similarly to (6.18):

$$\begin{aligned} \underset{\mathbf{f}}{\text{minimize}} \quad & \mu||(\mathbf{f}-\mathbf{g})||_1 + ||\mathbf{D}_{x,t}\mathbf{f}||_{TV} \\ \underset{\mathbf{f}}{\text{minimize}} \quad & \mu||(\mathbf{f}-\bar{\mathbf{f}})||_1 + ||\mathbf{D}_{y,t}\mathbf{f}||_{TV} \end{aligned} \tag{6.20}$$

where subscript t represents time domain and (x,t) and (y,t) are coupled.

6.4.2 Algorithm robustness

To evaluate algorithm robustness, we use two well-known local disparity methods: LASW [3] and CostFilter [4] for acquiring multi-array initial disparity estimates. Tables 6.3 and 6.4 show performance evaluation results using LASW and CostFilter, respectively. We obtain performance enhancement gain of about 50% regardless of the choice of disparity methods. It is important to note that the proposed algorithm performs better on poor initial disparity estimates (performance gain of 65% on the Room image in Table 6.3).

Table 6.3: Performance evaluation using LASW [3] with bad pixel % (threshold of 1 on all region).

Dataset	Initial	1^{st} MCF	CTV	2^{nd} MCF	Avg.	Enhance.
Room	32.4056	18.7337	15.1156	**11.2150**	10.2210	65%
Cones	27.6618	18.5322	14.9561	13.4255	**13.4160**	51%
Bike	40.9941	27.0120	23.5850	21.8451	**21.8239**	47%

Table 6.4: Performance evaluation using CostFilter [4] with bad pixel % (threshold of 1 on all regions).

Dataset	Initial	1^{st} MCF	CTV	2^{nd} MCF	Avg.	Enhance.
Room	35.9323	22.4186	18.7507	**15.0537**	15.0573	58%
Cones	32.5374	23.8333	17.9440	**16.5778**	16.5827	49%
Bike	42.6782	31.1995	24.5876	24.1462	**24.1429**	43%

6.4.3 Multiple cross-filling performance

The comparison of the conventional region voting [21] and proposed multiple-cross filling are performed. Table 6.5 demonstrates that the proposed filling algorithm outperforms the conventional region voting on all multi-array dataset. The local multiple cross-filling algorithm is simple and effective. In Table 6.4, the first filling process achieves high performance gain of 27% on the Cones image.

6.4.4 Simultaneous vs. cascade TV regularization

We presented a comparison of 2D simultaneous and cascade Tikhonov regularization in Section 6.3.2. Here, we extend the Tikhonov to TV regularization for 3D as well as 2D. 3D regularization is taken into account in multi-array systems, while 2D regularization is done in stereo systems. First, to evaluate the effectiveness of the cascade TV regularization on a single disparity map for stereo systems, we use the stereo dataset provided by Middlebury benchmark site [11]. Gaussian noise distributed as $\mathcal{N}(0, 10^{-3})$ is added to the ground-truth disparity maps. We solve the minimization problem by using two regularization approaches: (1) simultaneous TV and (2) proposed cascade TV.

Table 6.5: Conventional regions voting vs. proposed multiple filling with bad pixel % (threshold of 1 on all regions).

Dataset	Initial	Conventional	Proposed
Room	22.0833	19.7441	**18.2604**
Cones	22.5700	20.4541	**19.3480**
Bike	29.7653	26.5420	**26.3929**

Table 6.6: Simultaneous vs. cascade TV regularization on Middlebury dataset (bad pixel rates (on all regions) with threshold of 1)

Dataset	Measurement	Input	Simul. TV	Cascade TV	
				Intermediate	Final
Tsukuba	Bad Pixel(%)	1.16	0.43	0.15	**0.08**
	PSNR (dB)	30.44	39.30	37.48	**41.25**
Venus	Bad Pixel(%)	29.2	0.28	1.09	**0.11**
	PSNR (dB)	30.12	**45.67**	38.25	45.45
Teddy	Bad Pixel(%)	57.7	4.75	13.9	**4.22**
	PSNR (dB)	30.03	37.34	36.90	**37.98**
Cones	Bad Pixel(%)	57.5	5.46	16.4	**4.91**
	PSNR (dB)	30.08	37.34	36.39	**37.85**

The cascade TV regularization problem for single disparity map is expressed as

$$\begin{aligned} &\underset{\bar{\mathbf{f}}}{\text{minimize}} \;\; \mu||(\bar{\mathbf{f}}-\mathbf{g})||_1 + ||\mathbf{D}_x\bar{\mathbf{f}}||_{TV} \\ &\underset{\mathbf{f}}{\text{minimize}} \;\; \mu||(\mathbf{f}-\bar{\mathbf{f}})||_1 + ||\mathbf{D}_y\mathbf{f}||_{TV} \end{aligned} \tag{6.21}$$

where μ is set to a medium value of the parameter range recommended in [56] and $||\mathbf{D}_y\mathbf{f}||_{TV}$ is equivalent to $||\mathbf{D}_y\mathbf{f}||_1$. Table 6.6 demonstrates that the cascade TV regularization shows overall better performance in terms of bad pixel rate and PSNR.

We extend the comparison to 3D regularization so that multi-array disparities can be handled. The simultaneous one can be expressed as

$$\underset{\mathbf{f}}{\text{minimize}} \;\; \mu||(\mathbf{f}-\mathbf{g})||_1 + ||\mathbf{D}_{x,y,m}\mathbf{f}||_{TV} \tag{6.22}$$

whereas the cascade method is defined in (6.18). For multiple initial disparity estimates, we use two binocular stereo matching methods: (1) LM3C [2], producing relatively smooth disparity surface and (2) LASW [3], producing relatively noisy disparity surface.

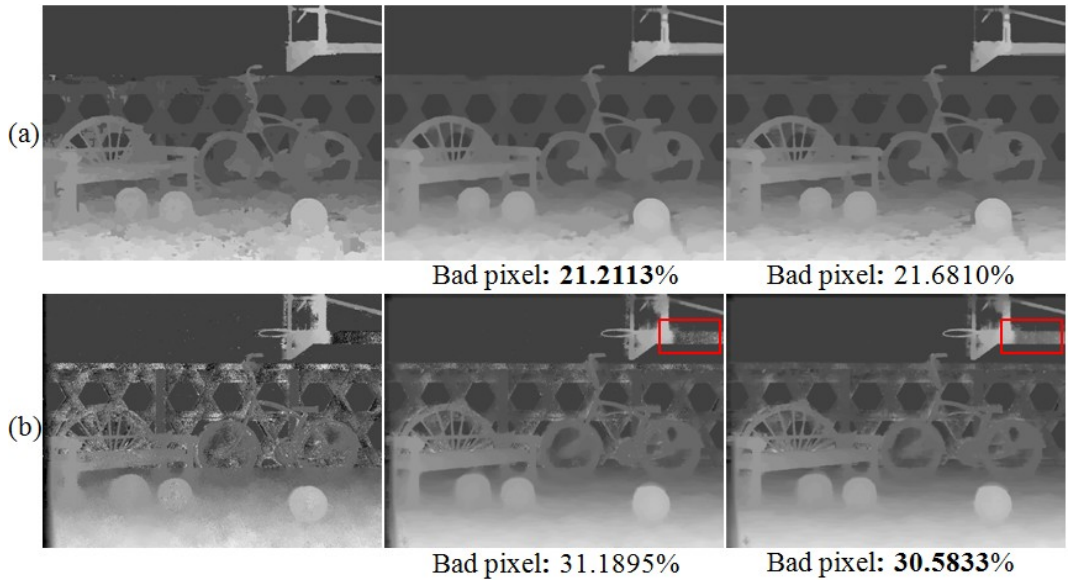

Figure 6.12: Simultaneous vs. cascade TV regularization for 3 dimensions. (First column) initial. (Second column) Simultaneous. (Third column) Cascade. (a) Using LM3C [2]. (b) Using LASW [3].

Fig. 6.12 shows initial and refined disparity maps. The cascade TV regularization shows better refinement on the noisy initial disparity surface while the simultaneous TV shows a slightly better bad pixel rate on the smooth initial disparity map. The performance of two approaches might be scene-dependent, as mentioned above. However, the proposed approach is more effective on poor initial and complicated disparity maps.

6.5 Summary

In conventional regularization problems, simultaneous regularization has always been used. In this chapter, a cascade regularization approach is proposed, and its advantages are investigated by detailed analyses and simulation results. To effectively fuse the multiple disparity estimates in multi-array camera system, we propose the cascade TV regularization, which can better reconstruct complicated structures by globally optimizing diagonal variations while reducing memory limitation. A multiple cross-filling algorithm is proposed to locally refine the initial disparities by achieving cross consistency between array disparity estimates. Simulation results show that the proposed

algorithm can enhance the initial disparities up to 65%.

6.6 Acknowledgements

This chapter is in part a reprint of a submitted paper to IEEE Transactions on Multimedia, 2014.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this dissertation, we present an accurate and efficient local disparity estimation algorithm, which can be used for stereo image and be extended for stereo video and large panoramic view. We focus on local stereo matching rather than global stereo matching because the local method has simple structure and more efficiency in real-time processing. However, it shows some drawbacks compared to the global method. We propose a novel algorithm to tackle each drawback. In addition, we present an algorithm to enhance the initial disparities computed by various binocular stereo matching methods. The enhancement algorithm, performed on the multi-array camera systems, achieves remarkable performance gain.

In Chapter 3, a new three-moded census with a noise buffer is proposed to increase robustness to image noise in homogeneous area. Moreover, we find that the cross-square census and combination of three similarity measures achieve more reliable similarity cost in a variety of image regions. To obtain more precise support weight window, we first define the relation among Gestalt principles and then model the advanced support weight computation function. Simulation results demonstrate that the proposed method is the best-performing local method on the Middlebury benchmark test.

In Chapter 4, we present a stereo video disparity algorithm by incorporating motion information as well as imposing temporal consistency. To reduce spatial ambiguities near moving edges, we propose to utilize optical flow, which can provide temporally

consistent information. Incorporating motion cue leads to further improving the reliability of the support aggregation. The experimental results show that the proposed method produces better video disparity estimates compared to other methods.

In Chapter 5, we present an effective multi-resolution depth processing and fusion for large panoramic images. We propose the adaptive disparity search range based on the combined local structure from image intensity and initial disparity. The adaptive range value can propagate the smoothness property at the low-resolution to the high-resolution while suppressing undesirable errors and preserving fine details. To reduce the disparity quantization error in the hierarchical scheme, we propose a reliable multiple fitting algorithm based on the conventional sub-pixel estimation. The spatial-multi-resolution TV method is employed to enforce consistency in both spatial and scaling dimensions. The experimental results on real-world panoramic images demonstrate that the proposed multi-resolution scheme produces high quality and high-resolution panoramic depth map by fusing individual multi-scale depth maps effectively. Moreover, it is performed with reduced complexity.

In Chapter 6, we present a new cascade regularization-based approach, which can restore diagonal structures better than conventional techniques. The detailed analyses and experimental results verify that the cascade approach better regularizes the diagonal variations and, in turn, achieves better image enhancement. We adapt the cascade TV regularization to the multi-array camera system in order to globally combine multiple disparity estimates. A local multiple cross-filling algorithm is proposed to achieve cross consistency between array disparity estimates by effectively filling the mismatches. Experimental results show that the proposed multi-array disparity enhancement algorithm can improve the accuracy of the initial array disparity estimates up to 65% while alleviating memory limitation. Moreover, the proposed algorithm is applicable to any binocular disparity methods.

7.2 Future Work

For future research directions, we have the following suggestions.

- We demonstrated that the advanced local support weight based on three Gestalt

principles is able to segment the same depth region well at various image resolutions. However, as the image resolution becomes extremely larger or smaller, the optimal window size should vary for the best performance. In these cases, the adaptive window size in addition to the adaptive support weight would help further increase the accuracy of disparity estimation.

- In the coarse-to-fine scheme, the proposed adaptive search range based on the combined local structure from image intensity and initial disparity is able to reduce error propagation as well as computational complexity. However, there are a few exceptions, such as specific areas where the local edge structure is not detected from both image intensity and initial disparity and where there are many thin structures.

- The multi-resolution depth fusion algorithm is performed on planar panoramic views. We think that it will be an interesting research topic to extend the proposed algorithm to cylindrical or spherical panoramic views.

- We have developed the disparity enhancement algorithm for multi-array camera system. In this dissertation, the basic array camera configuration is 3×3. The proposed algorithm can be theoretically extended to $n \times n$ array camera but it is not proved in real environment. In addition, simulations have been performed on only synthetic multi-array images. Therefore, it will be meaningful to apply the proposed algorithm to various real multi-array camera systems and then observe how much performance gain is achieved.

Bibliography

[1] R. Khoshabeh, S. Chan, and T. Nguyen, "Spatio-temporal consistency in video disparity estimation," in *Proc. IEEE ICASSP*, pp. 885–888, 2011.

[2] Z. Lee, J. Juang, and N. T.Q., "Local disparity estimation with three-moded cross census and advanced support weight," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 1855–1864, 2013.

[3] K. Yoon and I. Kweon, "Adaptive support weight approach for correspondence search," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, pp. 650–656, Apr. 2006.

[4] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *Proc. IEEE CVPR*, pp. 3017–3024, 2011.

[5] M. YI, S. Soatto, J. Kosecka, and S. S. Sastry, *An invitation to 3-D Vision From Images to Geometric Models*. Verlag: Springer, 2003.

[6] M. Pollefeys, R. Koch, and L. Van Gool, "A simple and efficient rectification method for general motion," in *Proc. IEEE ICCV*, vol. 1, pp. 496–501, 1999.

[7] R. Zabih and J. Ll, "Non-parametric local transforms for computing visual correspondence," in *ECCV*, pp. 151–158, 1994.

[8] D. Angens, *From Gestalt theory to image analysis : a Probabilistic approach*. NY: Springer, 2008.

[9] G. Papari and N. Petkov, "Adaptive pseudo dilation for gestalt edge grouping and contour detection," *IEEE Trans. Image Processing*, vol. 17, no. 10, pp. 1950–1962, 2008.

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

[11] D. Scharstein and R. Szelisk, "Middlebury stereo evaluation version 2," http://vision.middlebury.edu/stereo/eval, 2010.

[12] Autodesk, "3dx max," http://www.autodesk.com/products/autodesk-3ds-max/overview, 2013.

[13] H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1582–1599, 2009.

[14] A. Fusiello, V. Roberto, and E. Trucco, "Efficient stereo with multiple windowing," in *Proc. IEEE CVPR*, pp. 858–863, 1997.

[15] S. Mattoccia, F. Tombari, and L. Stefano, "Segmentation-based adaptive support for accurate stereo correspondence," in *Proc. PSIVT*, pp. 427–438, 2007.

[16] Y. Liu, Z. Gu, X. Xu, and Q. Zhang, "Local stereo matching with adaptive support-weight, rank transform and disparity calibration," in *Pattern Recognition Letters 29*, pp. 1230–1235, 2008.

[17] C. Rhemann, M. Bleyer, and C. Rother, "Patchmatch stereo - stereo matching with slanted support windows," in *Proc. BMVC*, 2011.

[18] C. Liu, W. Freeman, R. Szeliski, and S. B. Kang, "Noise estimation from a single image," in *Proc. IEEE CVPR*, pp. 901 – 908, 2006.

[19] L. Zhang, S. Vaddadi, H. Jin, and S. Nayar, "Multiple view image denoising," in *Proc. IEEE CVPR*, pp. 1542 – 1549, 2009.

[20] C. Connolly and T. Fleiss, "A study of efficiency and accuracy in the transformation from rgb to cielab color space," *IEEE Trans. Image Processing*, vol. 6, no. 7, pp. 1046–1048, 1997.

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

[22] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang, "On building an accurate stereo matching system on graphics hardware," in *Proc. IEEE ICCV*, pp. 467–474, 2011.

[23] H. Tao and H. Sawhney, "Global matching criterion and color segmentation based stereo," in *Proc. IEEE WACV*, pp. 246–253, 2000.

[24] A. Hosni, M. Bleyer, M. Gelautz, and C. Rhemann, "Local stereo matching using geodesic support weights," in *Proc. IEEE ICIP*, pp. 2093 – 2096, 2009.

[25] L. Wang, M. Gong, R. Yang, and M. Gong, "A performance study on different cost aggregation approaches used in real-time stereo matching," in *Proc. IJCV*, vol. 75, pp. 283–296, 2007.

[26] M. Bleyer and M. Gelautz, "Temporally consistent disparity maps from uncalibrated stereo videos," in *Proc. ISPA*, pp. 383–387, 2009.

[27] B. Horn and B. Schunck, "Determining optical flow," *Artificial Intelligence*, pp. 185–203, 1981.

[28] C. Richardt, D. Orr, A. Davies, I. Criminisi, and N. A. Dodgson, "Realtime spatiotemporal stereo matching using the dual-cross-bilateral grid," in *Proc. ECCV*, 2010.

[29] D. Sun, S. Roth, and M. Black, "Secrets of optical flow estimation and their principles," in *Proc. IEEE CVPR*, pp. 2432–2439, 2010.

[30] S. Baker, S. Roth, D. Scharstein, M. Black, J. P. Lewis, and R. Szeliski, "A database and evaluation methodology for optical flow," in *Proc. IEEE ICCV*, pp. 1–8, 2007.

[31] R. Ainsworth, D. Sandin, A. Prudhomme, J. Schulze, and T. DeFanti, "Acquisition of stereo panoramas for display in vr environments," *SPIE Electronic Imaging, The Engineering Reality of Virtual Reality*, 2011.

[32] M. Accame, F. De Natale, and D. Giusto, "Hierarchical block matching for disparity estimation in stereo sequences," in *Proc. IEEE ICIP*, vol. 2, pp. 374–377, 1995.

[33] T. Kudo, K. Shirai, and M. Ikehara, "Hierarchical stereo matching via color segmentation," in *Proc. IEEE DSP/SPE*, pp. 522–525, 2006.

[34] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," in *Proc. IEEE CVPR*, pp. 261–268, 2004.

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

[36] Q. Yang, L. Wang, R. Yang, S. Wang, M. Liao, and D. Nister, "Real-time global stereo matching using hierarchical belief propagation," in *Proc. BMVC*, 2006.

[37] S. Grauer-Gray and C. Kambhamettu, "Hierarchical belief propagation to reduce search space using cuda for stereo and motion estimation," in *Proc. IEEE WACV*, pp. 1–8, 2009.

[38] M. Sizintsev and R. Wildes, "Efficient stereo with accurate 3-d boundaries," in *Proc. BMVC*, 2006.

[39] G. Van Meerbergen, M. Vergauwen, M. Pollefeys, and L. Van Gool, "A hierarchical stereo algorithm using dynamic programming," in *Proc. IEEE SMBV*, pp. 166–174, 2001.

[40] M. Sizintsev, "Hierarchical stereo with thin structures and transparency," in *Proc. CRV*, pp. 97 –104, may 2008.

[41] M. Sizintsev and R. Wildes, "Coarse-to-fine stereo vision with accurate 3d boundaries," *Image and Vision Computing (IVC)*, vol. 28, pp. 352–366, 2010.

[42] Y. Yang and A. Yuille, "Multi-level enhancement and detection of stereo disparity surfaces," *Artificial Intelligence*, vol. 78, pp. 121–145, 1995.

[43] Y. Jen, E. Dunn, P. Gite-Georgel, and J. Frahm, "Adaptive scale selection for hierarchical stereo," in *Proc. BMVC*, pp. 95.1–95.10, 2011.

[44] B. Jahne, *Digital Image Processing*. Springer, Jan. 2006.

[45] T. Brox, F. Boomgaard, F. Lauze, J. Weijer, F. Weickert, and P. Kornprobst, *Adaptive Structure Tensors and their Applications*. Visualization and Processing of Tensor Fields, Springer Berlin Heidelberg, 2006.

[46] G. Strang, *Linear Algebra and Its Applications, 3rd ed.* Harcourt Brace Jovanovich, 1988.

[47] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. IEEE CVPR*, pp. 1–8, 2007.

[48] Z. Zhang, X. Ai, N. Canagarajah, and N. Dahnoun, "Local stereo disparity estimation with novel cost aggregation for sub-pixel accuracy improvement in automotive applications," in *Proc. IEEE IVS*, pp. 99–104, 2012.

[49] M. Shimizu and M. Okutomi, "Precise sub-pixel estimation on area-based matching," in *Proc. IEEE ICCV*, vol. 1, pp. 90–97, 2001.

[50] A. Stein, A. Huertas, and L. Matthies, "Attenuating stereo pixel-locking via affine window adaptation," in *Proc. IEEE ICRA*, pp. 914–921, 2006.

[51] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM SIGGRAPH*, p. 96, 2007.

[52] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE ICCV*, pp. 839–846, 1998.

[53] F. Garcia, B. Mirbach, B. Ottersten, F. Grandidier, and A. Cuesta, "Pixel weighted average strategy for depth sensor data fusion," in *Proc. IEEE ICIP*, pp. 2805 –2808, sept. 2010.

[54] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A noiseaware filter for real-time depth upsampling," in *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008.

[55] T. Lindeberg, *Scale-Space Theory in Computer Vision*, vol. 256. Springer, 1994.

[56] S. Chan, R. Khoshabeh, K. Gibson, P. Gill, and T. Nguyen, "An augmented lagrangian method for total variation video restoration," *IEEE Trans. Image Processing*, vol. 20, pp. 3097 –3111, nov. 2011.

[57] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, pp. 353 – 363, 1993.

[58] K. Yamada, T. Ichikawa, T. Naemura, and K. Aizawa, "Generation of a disparity panorama using a 3-camera capturing system," in *Proc. IEEE ICIP*, pp. 772 – 775, 2000.

[59] M. Maitre, Y. Shinagawa, and M. Do, "Symmetric multi-view stereo reconstruction from planar camera arrays," in *Proc. IEEE CVPR*, pp. 1 – 8, 2008.

[60] B. Wilburn, N. Joshi, V. Vaish, E. Talvala, and E. Antunez, "High performance imaging using large camera arrays," *ACM Trans. Grpahics*, vol. 24, pp. 765 – 776, 2005.

[61] S. D. Cochran and G. Medioni, "3-d surface description from binocular stereo," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, pp. 981 – 994, 1992.

[62] P. Fua, "A parallel stereo algorithm that produces dense depth maps and preserves image features," *Machine Vision and Applications*, vol. 6, pp. 35 – 49, Winter 1993.

[63] J. Lu, G. Lafruit, and F. Catthoor, "Anisotropic local high-confidence voting for accurate stereo correspondence," in *Proc. SPIE-IS&T Electron. Imaging*, vol. 6812, p. 68120, Jan. 2008.

[64] D. Geman, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, pp. 367 – 383, 1992.

[65] P. Hansen, "Analysis of discrete ill-posed problems by means of the l-curve," *SIAM Review*, vol. 14, pp. 561 – 580, 1992.

[66] M. Bertero and P. Boccacci, *Introduction to inverse problems in imaging*. IOP Publishing Ltd, 1998.

[67] C. R. Vogel, *Computational Methods for Inverse Problems*. SIAM Frontiers in Applied Mathematics, 2002.

[68] A. Beck and A. Ben-Tal, "On the solution of the tikhonov regularization of the total least squares problem," *SIAM Journal on Optimization*, vol. 17, pp. 98 – 118, 2006.

[69] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, pp. 259 – 268, Nov. 1992.

[70] T. Chan, G. Golub, and P. Mulet, "A nonlinear primal-dual method for total variation-based image restoration," *SIAM Journal on Scientific Computing*, vol. 20, pp. 1964 – 1977, Nov. 1999.

[71] Y. Wang, J. Yang, W. Yin, and Y. Zhang, *An efficient TVL1 algorithm for deblurring multichannel images corrupted by impulsive noise*. CAAM, Rice Univ., Sep. 2008.

[72] M. Afonso, J. Bioucas-Dias, and M. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Trans. Image Process*, vol. 19, pp. 2345 – 2356, Sep. 2010.

[73] E. Shechtman, Y. Caspi, and M. Irani, "Space-time super-resolution," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, pp. 531 – 545, apr. 2005.

[74] G. Golub and C. Van Loan, *Matrix Computation, 2nd ed.* Baltimore, MD: Johns Hopkins Univ. Press,, 1989.

[75] P. J. Davis, *Circulant Matrices, 2nd ed.* New York: Chelsea, 1994.

[76] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for l1-minimization with applications to compressed sensing," *SIAM J. Imag. Sci.*, vol. 1, no. 1, pp. 143 –168, 2008.

[77] A. Szlam, G. Zhaohui, and S. Osher, "A split bregman method for non-negative sparsity penalized least squares with applications to hyperspectral demixing," in *Proc. IEEE ICIP*, pp. 1917 – 1920, 2010.