Depth-constrained Feature-based Stitching System for Stereoscopic Panoramic Video Generation

BY

HAOYU WANG
B.Sc. Electrical and Information Eng., Huazhong University of Science and Technology, 2010

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Chicago, 2019

Chicago, Illinois

Defense Committee:

Dan Schonfeld, Chair and Advisor
Daniel J. Sandin, Co-advisor, Computer Science
Andrew Johnson, Computer Science
Hulya Seferoglu
Mojtaba Soltanalian

ACKNOWLEDGMENT

Primarily, I want to pay my authentic thankfulness to my advisor, Professor Dan Schonfeld, for his vision, help, and guidance during my Ph.D. period. I would also like to thank Professor Dan Sandin, who is the supervisor of my primary research achievements during the SENSor Environment Imaging project. To my best appreciates, I am grateful to his considerable patience in communication and effective suggestions towards the entire project. Without the help of my supervisor, I will miss the chance to explore this exciting field and couldn't finish this dissertation of appropriate standards. I wound like to extend my gratitude to my dissertation committee members (Professor Andrew Johnson, Professor Hulya Seferoglu, and Professor Mojtaba Soltanalian) for their unwavering support and assistance. They provided guidance and helps in all areas that helped me accomplish my research goals and enjoy myself in the process.

Here I also want to thank those who have been encouraging me and supporting me all the time. Without their trust and help, I would not have the confidence to overcome all the challenges and obtain my Ph.D. degree. Thank you all.

HW

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

FIGURE **PAGE**

LIST OF FIGURES (Continued)

FIGURE **PAGE**

LIST OF FIGURES (Continued)

FIGURE **PAGE**

LIST OF ABBREVIATIONS

AANAP	Adaptive As-Natural-As-Possible
CIF	Commonly Identified Feature
CP	Control Point
CSPS	Casual Stereoscopic Panorama Stitching
DoG	Difference of Gaussian
FOV	Filed of View
GBVS	Graph-Based Visual Saliency
KLT	Kanade Lucas Tomasi
NCC	Normalized Cross-Correlation
NFS	No Feature Selection
POF	Partially Occluded Feature
PPD	Pixel Per Degree
PSI	Panoramic Stereo Imaging
RANSAC	Random sample consensus
RMSE	Root Mean Square Error
ROI	Region of Interest
SFS	Saliency-based Feature Selection

LIST OF ABBREVIATIONS (Continued)

SIFT	Scale-invariant feature transform
SPHP	Shape-Preserving Half-Projective
SSIM	Structural Similarity
SURF	Speeded Up Robust Feature
TPS	Thin Plate Spline
VR	Virtual Reality

SUMMARY

The thesis presents one feature-based stitching system for high-quality stereoscopic panoramic video generation. Although panorama stitching is a well-studied topic, and various algorithms and software are available to create high-quality monocular panoramas, generalizing those methods for both stereo and video modes is not an easy task. One satisfying output of the stitching system needs to meet several requirements in different senses. For every single frame, the output monocular-view panorama should have minimal spatial stitching errors or distortion. For every pair of output stereoscopic panorama, no vertical disparity can be perceived, and the horizontal disparity should be appropriately distributed across the scenario. For the output video, discontinuities, such as shakiness or abrupt changes of depth between consecutive frames, are not desired.

The main contribution of our work consists of the definition of depth-constrained feature in stitching framework, the introduction of human-visual interest to control point refinement, and the post-stitching correction of the artifact and depth anomaly. First, the stitching, based on the proposed depth-constrained feature, can ensure fewer visible artifacts in the generated monocular panorama and better stereo consistency between the left and right views. Furthermore, we utilize human visual sensitivity to refine and qualify the input control points and tracking results of the Kanade-Lucas-Tomasi tracker in the video sequences. Finally, the pixel-based monocular geometric correction and feature-based depth control enable us to minimize all the visible stitching errors and adjust the perceived depth from the stereoscopic panoramic video into one reasonable range.

CHAPTER 1

INTRODUCTION

The content of this chapter is based on our works that are published in [1–3].

1.1 Goal Description and Motivation

The increasing interest in virtual reality (VR) has heightened the need for high-quality VR video content using real-world data. Although panorama stitching is a well-studied topic, and various algorithms and software are available to create high-quality monocular panoramas, generalizing those methods for both stereo and video modes is not an easy task. Some early studies about the monocular panorama stitching, such as the original spinning PSI system, fail to extend to stereo or video case since they don't take stereo and video into account. Some newer algorithms suffer from expensive computation for densely sampled data or high-accuracy depth information [4]. Some omnistereo projection-based algorithms [5] utilize view interpolation techniques to synthesize each column from real captured image views. But all these methods heavily depend on the high-quality, dense depth maps, which is still not easy to generate [6], especially for high-resolution images. Base on the literature review we make, there are no efficient sparse feature-based stereoscopic panoramic video generation systems. Thus, in this thesis, we intend to construct a unified stitching framework that operates feature-based 360 by 180 stereoscopic panoramic video panorama stitching tasks with low computational costs.

To construct a stitching system for stereoscopic panoramic video, we assume that a good output video has the following properties. First, the single-frame panorama should have minimal spatial stitching errors or distortion. Second, the stereoscopic panorama should be consistently stitched so that no vertical disparities can be detected, and the perceived depths of all objects should be properly distributed. In an excellent panoramic video, discontinuities, such as shakiness or abrupt changes of depth between consecutive frames, are not desired. In summary, we face four challenges with the stitching framework:

1. good monocular view panorama stitching quality
2. stereo consistency between binocular views
3. reasonable depth carried with the stitched video
4. temporal consistency between consecutive frames

To solve these problems, we formulate a framework for the stereoscopic panoramic video generation, based on commonly identified features, such that the monocular stitching quality, stereo consistency, and temporal consistency can be achieved simultaneously. The commonly identified feature (CIF) is one designed feature structure to describe the points, edges, and areas that can be observed and precisely described by multiple camera views in the stereoscopic camera arrangement. Since this feature structure characterizes the same content from different view perspectives, the corresponding depth information can be easily obtained based on the triangulation and known baselines between stereo camera pairs. Thus, compared to independently detected features, the operation of control points matching between CIF can benefit from the incorporated depth term and get more reliable corresponding results.

Given the initially detected CIF set from input images, we conduct one feature refinement before sending them into the feature tracker in the video sequence. To construct one reasonable and efficient stitching system, we employ human visual sensitivity to eliminate the redundant control points and adjust the distribution of sampled control points. The proposed energy map combines depth maps, gradient maps, and saliency maps to indicate the visual importance of each pixel. According to the generated energy map, we redistribute all selected commonly identified grid-wise features to shrink the size of control point lists and to fit the sampled control points in tune to human visual perception.

Because video stitching largely relies on temporal coherence of selected control points, we utilize commonly identified features in successive frames to achieve well-stitched content in the video generation task. The temporal consistency between consecutive frames can be interpreted as consistencies in geometry, vertical, and horizontal disparities. For more detail, the geometric consistency indicates the shape, size, and relative location of objects that should remain identical between the prior and consecutive frames. The temporal consistency in vertical and horizontal directions guarantees there will be no abrupt changes in the perceived depth from the same object. We employ human visual sensitivities to generate a grid- and saliency-based energy map to indicate the visual importance of pixels. Then, the global temporal feature-tracking can be decomposed into several grid-based local tracking tasks according to changes in pixel energy. To further improve the accuracy of commonly identified features, we extend the underlying assumption of small-displacement into the depth domain, removing the falsely tracked control points. Moreover, to compensate removed CIF, we detect and construct new features that have the minimal L2 norm distance to the position of CIF in the previous frame.

During the image alignment step, the standard operation in the conventional framework is to utilize random-sample consensus (RANSAC) for homography estimation for the left-view and right-view independently. However, the random draw from the control point pool at left-view and right-view respectively won't guarantee the consistency of the fitted homography. Thus, we modified the original random-sample consensus (RANSAC)-based homography estimation algorithm so that it can work for the stereo camera pose estimation.

After the final composition of the stereoscopic panorama, we can always find some artifacts in the overlapping region, especially near the blending seams. Those stitching errors, such as object cropping, straight-line discontinues, and region distortion, are not expected according to the first challenge we stated above. Thus, to mitigate all these undesirable viewing experiences, we propose one flow-map-guided correction framework to solve those problems. Since there is no ground truth in the monocular stitching task, the correct reference we can utilize to make the artifact correction should come from the input images. In our proposed correction technique, we intend to fix all the visible stitching errors in the output panorama via ROI reconstruction from the corresponding artifact-free area from the input images.

After the artifact correction in the monocular sense output panorama, we still need to take care of possible depth issues carried with the independently corrected panoramas. Considering the depth from the input image pairs as the ground truth, we intend to adjust the left-view and right-view panorama globally and locally to provide a more accurate depth distribution. The global depth adjustment aims to align the output left-view and right-view panorama via pure translation. Subsequently, the majority of the pixels in the output panorama are expected to have no vertical disparity and minimal horizontal error

distance. Then, for those regions that still suffer from depth issues, we select one of the two monocular-view panoramas as the reference and then warp another monocular view via Thin-plate-spline (TPS) to achieve depth correctness in the local region.

The main contribution of our work consists of a unified framework that facilitates the generation of high-quality stitched stereoscopic panoramic videos and the introduction of human-visual interest to control point selection and refinement in the 360 by 180 stereoscopic panorama composition. First, the stitching, based on commonly identified features, can ensure fewer visible artifacts in the generated monocular panorama and better stereo consistency between the left and right views. Furthermore, we utilize human visual sensitivity to refine and qualify the tracking results of the Kanade-Lucas-Tomasi (KLT) tracker for more reliably matched control points. Then, we extend the original monocular RANSAC-based algorithm to the stereo sense for the more consistent camera poses. Finally, the pixel-based monocular geometrical correction and feature-based depth control enable us to minimize all the visible stitching errors and adjust the perceived depth from the stereoscopic panoramic video into one reasonable range. To validate the stitching feasibility under various camera arrangements and robustness to different scenes, we conduct extensive panorama stitching experiments with various camera-based images and synthetic data from simulation software.

1.2 Proposed Framework

Figure 1 shows the pipeline of the standard framework used by PanoTools [7], and Figure 2 shows our proposed generation system. In the following of this chapter, we walk through our proposed framework step by step.

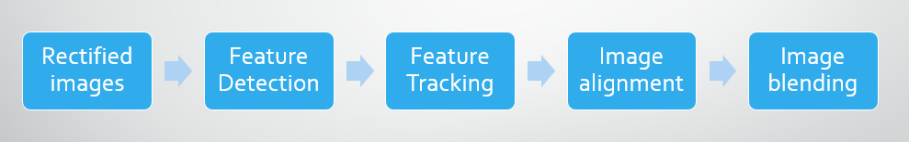


Figure 1. Pipeline of standard framework

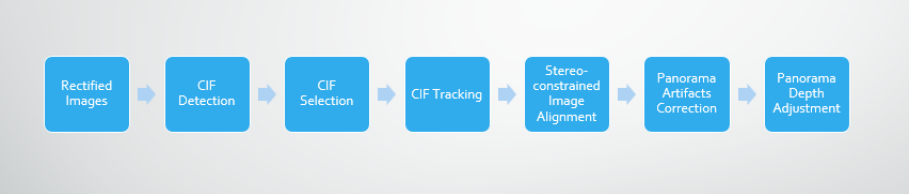


Figure 2. Pipeline of proposed framework

1.2.1 Depth-constrained Feature Detection and Matching

In the first step of panorama stitching, we detected standard SIFT features from all the input images independently. Instead of directly considering them as control points for matching, we divide them into two groups: commonly identified feature (CIF) and partially occluded feature (POF). The commonly identified feature refers to those patches that can be observed by all the input images, while a partially occluded feature indicates the absence of at least one input image. Based on the grouping result, we carefully select commonly identified features as control points and operate corresponding matching in the first frame. Since the CIF contains multiple standard features, the depth information carried with these standard features is taken into consideration in the proposed matching function to improve

the matching reliability. The nonvertical disparity term and horizontal disparity consistency term in our proposed matching function make it more robust than the standard two-dimensional (2D) features matching based on the gradient.

1.2.2 Saliency-based Feature Selection

After we obtain the generated CIF set for stereoscopic panorama stitching at the first frame, we operate one feature selection process to prune the original control point list for later feature tracking steps. Instead of sending all detected features to the video feature tracker directly, we only choose parts of original commonly identified features that can guarantee efficient feature tracking, distributed across the camera views. Thus, we incorporate saliency and gradient maps to construct one energy map, which indicates the pixel sampling weight distribution. Then, we divide the camera image into multiple grids and select the best-matched commonly identified features of each grid. The proposed energy map will determine the maximal sample number in each grid. In this way, we could improve the feature tracking efficiency and avoid the potential artifacts caused by the imbalance distribution of control points in different depth planes.

1.2.3 Depth-constrained Feature Tracking

Given the refined control points in the first frame, the next step is about tracking their positions in the video sequence. During this step, we introduce one energy-based and grid-based feature update strategy to replace the original global update. This proposed feature update strategy focuses on the change of energy in each grid between consecutive frames. Hence, only those regions with significant energy changes require feature updates; other areas can retain the same control points inherited from the previous frame. Moreover, to improve the reliability of control point tracking and to avoid drift

problems, we extend a small displacement assumption of the KLT tracking algorithm into the 3D case and propose two filtering conditions to qualify valid CIF. After the strict qualification of the tracking results, we need to fill those rejected positions in the control point list for the current frame. To attain the spatial coherence in the video sequence, we intend to select new feature descriptors that have the least distance with retrieved position parameters from the old features at previous frames.

1.2.4 Stereo-constrained Image Alignment

The next step in our proposed framework is image alignment and blending. During this step, we extend the original RANSAC-based homography estimation algorithm into the stereo sense. Instead of randomly selecting subsets of ther standard feature list and operating homography estimation for the left and right views independently, we only choose CIF from the control point list and introduce similarity penalty terms to fit consistency between left-view and right-view perspective transformation matrices. With known camera poses, we use EnBlend [8] as the blender to compose the final output panoramas from multiple camera views.

1.2.5 Post-stitching Flow-Map-Guided Panorama Correction

Usually, the standard panorama stitching framework ends up the image alignment and blending step. However, the previous feature-based technique cannot eliminate all the visual artifacts in the finally composited panorama. Those stitching errors, such as object cropping, straight-line discontinuities, and region distortion, will cause viewing discomfort. To correct those visible errors in the output panorama, we establish one dense and pixel-wise correspondence between the erroneous region of interest (ROI) and artifact-free ROI from the input image. By adjusting the pixel-wise position movement between target coordinate and reference coordinate, we expect to warp that artifact-free area into the correspond-

ing region in the output panorama. After the removal of original pixel position displacement values around ROI, we assign these neighboring pixels with smoother position displacement values. Under the guidance of pixel-wise updated position displacement maps, we can reconstruct the erroneous target ROI from the reference ROI and avoid the generation of visible stitching errors.

1.2.6 Post-stitching Feature-based Depth Adjustment

The post-stitching flow-map-guided panorama correction is designed to deal with the monocular-view stitching problems, but viewing experience of the stereoscopic panorama also heavily relies on the stereo consistency between left-view and right-view panorama. After we apply our proposed artifact correction technique to fix all the visible stitching in the original monocular view panorama, one sparse feature-based method is used to ensure the depth accuracy of output stereoscopic panorama globally and locally. The depth information obtained from input-rectified image pairs is considered ground truth in this depth adjustment. For global depth correction, we fix the reference panorama and globally shift the target panorama until the minimal sum of absolute differences in both vertical and horizontal directions is achieved. In the local depth correction, we manually label those regions that still suffering from depth issues after global correction. Then we adopt the thin-plate-spline morphing (TPS) method to warp all the pixels in the ROI to their expected positions with more reasonable disparity values.

1.3 Thesis organization

The rest of the thesis is organized as follows.

In Chapter 2, we present the definition of depth-constrained and adapted matching operation for the CIF construction for the first frame.

In Chapter 3, we present the feature selection and redistribution of control points according to the human visual interest.

In Chapter 4, we present the depth-constrained CIF tracking strategy and blending strategy in the video sequence.

In Chapter 5, we present the original RANSAC-based homography estimation and our proposed stereo-constrained image alignment.

In Chapter 6, we present all the details of the experiment and simulation setup.

In Chapter 7, we present the flow map guided artifact correction to the monocular panorama.

In Chapter 8, we present the feature-based depth adjustment to the stereoscopic panorama.

In Chapter 9, we summarize our works and contributions stated in the thesis.

CHAPTER 2

COMMONLY IDENTIFIED FEATURE CONSTRUCTION AND MATCHING

The content of this chapter is based on our work that is published in [1] .©2017 IEEE. Reprinted with permission, from [1].

2.1 Background and Related Works

Panorama stitching is a well-studied topic, and various algorithms and software are available for users to create high-quality monocular panorama [9]. The increasing interest in the field of virtual reality has heightened the need for high-quality VR content based on real-world data. However, the generalization from monocular panorama approaches to the stereo case by independently creating binocular view panorama is problematic. The inconsistency in the same objects in the left-view and right-view scenes may cause viewing discomfort and 3D fatigue to viewers and incorrectly delivered depth.

Several reliable stereoscopic image stitching methods have been proposed to deal with the stitching task in stereo mode [4, 5, 10–12]. Unfortunately, these approaches require densely sampled depth information, which requires expensive computation and may result in inaccurate estimation of the depth. Peleg *et al.* proposed the setup of a single camera that rotates in a circular trajectory for stereoscopic panorama stitching [11]. Couture *et al.* introduced a system based on a pair of rotating cameras and stitching complete frames instead of small strips [13, 14]. Because these old rotating systems can only work for static scenes, synchronized multi-camera arrays were proposed. Couture proposed a system containing six cameras with fish-eye lenses [15]. The Google Jump system used 16 static cameras

arrayed along a circle and interpolated hundreds of virtual views for stitching [5]. However, most of these extensions of the idea by Peleg lack the necessary procedure to handle the increasing distortions at the top and bottom of the camera array, which means that it is quite difficult to generalize them to the $360°$x$180°$ case.

Richardt *et al.* [12] proposed an optical flow-based blending approach to reduce visual artifacts for images captured using hand-held cameras. However, they only compensate for vertical parallax by projecting undistorted input images onto a cylindrical imaging surface. Zhang and Liu also extended a spatially varying warping method [16] in their proposed approach to perform panorama stitching from a sparse set of casually taken input images [4]. However, its stereo consistency of the stitching results is mainly manipulated by the pre-stitched reference panorama and the pre-stitched dense disparity map.

In this chapter, we mainly present a depth-constrained feature structure for the generation of high-resolution stereoscopic panoramas. Compared to generating a high-quality monocular scene, generating a high-quality stereo panorama using existing technologies is challenging owing to the inconsistency between the left and right views and difficulties in disparity control. In our proposed stitching framework, the CIF is designed to describe the identical feature in two pairs of adjacent input images instead of processing the left and right images independently. The proposed feature structure can not only fit different hardware setups well with higher reliability but also attain good extensibility to further modification to the camera arrangement designs. Moreover, the proposed CIF structure and matching operation are also flexible enough to be generalized to the 4π steradian stereoscopic panoramic video case.

2.2 Proposed Feature Structure for Stereoscopic Panorama Stitching

2.2.1 Traditional Feature Structure in Panorama Generation

In one standard panorama stitching task, various feature descriptors are employed to characterize the pose between neighboring cameras, such as scale-invariant feature transform (SIFT) [17] and speeded up robust features (SURF) [18]. These features are popular in the task of panorama stitching because they are invariant to the rotation and scaling, which is the desired property during the image alignment. Thus, in our proposed stitching system, we adopt SIFT as the basic operation unit for the feature detection from each single input image. To better understand of the proposed SIFT structure for the stereoscopic panorama task, we first discuss more details about the generation of the standard SIFT feature descriptor.

There are mainly four steps involved in SIFT algorithm as follows [19]:

1. Scale-space Extrema Detection
2. Keypoint Localization
3. Orientation Assignment
4. Keypoint Descriptor

Scale-space Extrema Detection

The generation of a standard SIFT starts with the points of interest detection, which are always considered one pre-processing operation in the standard SIFT framework. Before the operation of keypoint detection, we usually generate Gaussian-blurred images with different scales. For more de-

tails, each image will be convolved with Gaussian filters at successive scaling scales. Specifically, a $DoGimageD(x, y, \sigma)$ can be defined as:

$$D(x, y, \sigma) = L(x, y, k_i\sigma) - L(x, y, k_j\sigma). \tag{2.1}$$

where $L(x, y, k\sigma)$ is the convolution result of the original image $I(x, y)$ with the Gaussian blur $G(x, y, k\sigma)$ at the scale parameter $k\sigma$:

$$L(x, y, \sigma) = G(x, y, k\sigma) * I(x, y). \tag{2.2}$$

Thus, a DoG image between scale $k_i\sigma$ and $k_j\sigma$ is actually defined as the difference of Gaussian-blurred images at scales$k_i\sigma$ and $k_j\sigma$. After the convolution with Gaussian-blurs at various scales, the output images are classified into different groups by octave. The octave here refers to a set of size-reduced image after progressively Gaussian-blurring. In each octave, the DoG is represented as the difference between images with adjacent Gaussian blurring size. After the comparison to eight neighbors at the same scale and nine corresponding neighboring pixels in each of the adjacent scales, the pixel with the minimal or maximal value will be considered as a candidate keypoint.

Keypoint Localization

In the result of scale-space extreme detection, there are always too many keypoint candidates. To keep those stable key point candidates, one precise fit to nearby data is needed for the more accurate location, scale, and ratio of principal curvatures. Thus, if the intensity at one selected keypoint is less than the threshold value, it will be rejected. If one selected keypoint is identified as an edge, it will also be removed.

Orientation Assignment

Then, an major orientation has to been assigned to each keypoints to achieve invariance to image rotation. Those neighbors that are around the keypoint location are taken into account for gradient and magnitude and direction calculation. Gradient magnitude and direction from the neighbors around the keypoint location are used to create one histogram with 36 bins covering 360 degrees. In the histogram, only the highest peak and other peaks above 80% of it will be considered to calculate the orientation of the features. They will create key points with the same location and scale but different directions.

Keypoint Descriptor

For each keypoint, the intensity values of its 16x16 neighborhood are usually used to compute the gradient vector. The 16x16 neighborhood can be divided into 16 sub-blocks of 4x4 size. Each sub-figure can produce an 8-bin orientation histogram so that there are 128 bin values to present one standard SIFT feature. To enhance the invariance of feature to affine changes in illumination, this 1x128 gradient vector needs to be normalized to unit length.

2.2.2 Definition of Depth-constrained Feature

Though the standard SIFT works well in the monocular-view panorama stitching task, it lacks the necessary mechanism to guarantee the stereo consistency between the left-view and right-view panorama. Compared to the monocular view stitching, the stereoscopic panorama generation needs more careful selection of control points due to the extra constraints in the depth dimension. The straightforward application of standard SIFT to the stereo case can be decomposed into the original SIFT feature detection and gradient-based matching independently for left-view and right-view input images. How-

ever, those separately detected and matched feature descriptors may produce serious stereo rivalry in the output stereoscopic panorama.

To solve these problems, we proposed one depth-constrained feature structure for the feature-based stereoscopic panorama generation framework. Given the initial detection result of standard SIFT for each input image, we don't accept all of them as the control points for the later image alignment. Based on whether each feature can be observed and precisely described by multiple camera views, we can divide these initially detected features set into two groups: the commonly identified feature (CIF) and the partially occluded feature (POF).

Figure 3 is an example of the CIF. We can find those key points in four neighboring images that are linked by four red lines, which indicates those feature descriptors can be viewed and recognized in all these adjacent views. Thus, the areas represented by these key points will be under consideration for left-view and right-view panorama generation simultaneously. The consistency in control points selection here is expected to maintain the stereo consistency in the output panoramas.

Figure 4 is an example of the partially occluded feature. We can find those key-points that are linked by three or fewer red lines, which indicates those feature descriptors can only be viewed and recognized in three or fewer neighboring views. In this case, the input left-view and right-view input images will be aligned by very different control points under the independently stitching strategy.

Thus, in our proposed stitching system, CIF will be considered as the basic matching unit instead of the standard SIFT feature. To explain the whole construction more clearly, we only consider the stitching task for two pairs of input rectified stereoscopic images, I_{L1}, I_{L2}, I_{R1}, and I_{R2}. In a more complex situation, we can process more images with a similar pattern. In the traditional definition of SIFT fea-

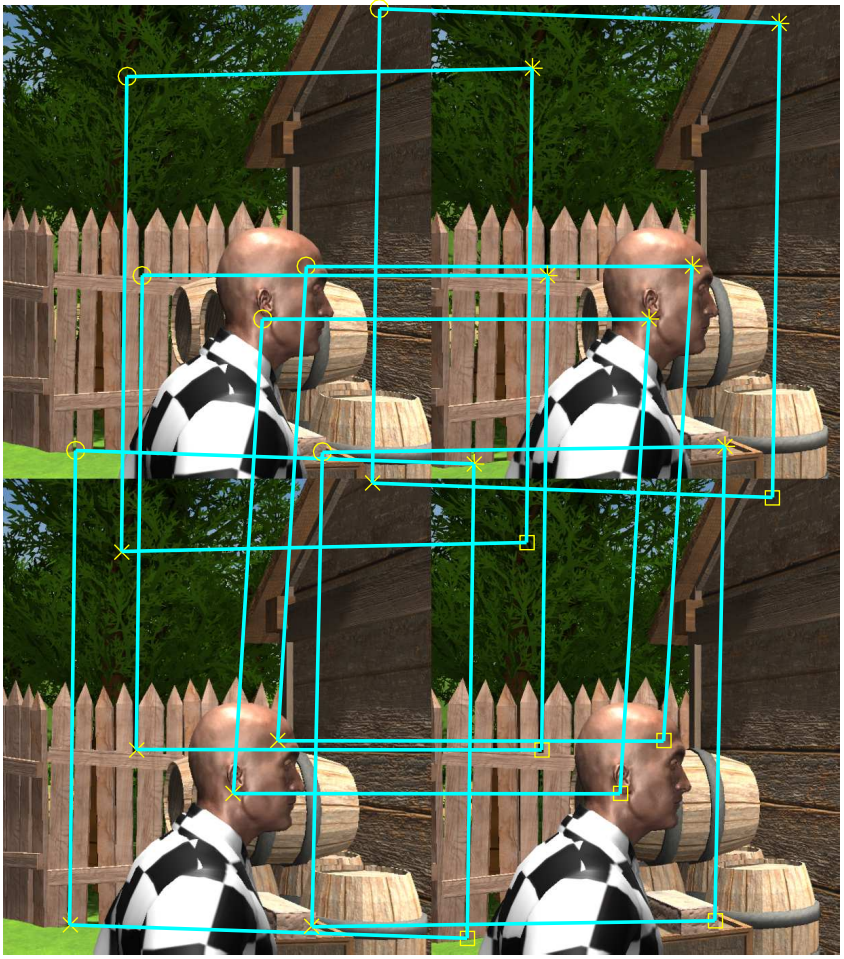


Figure 3. Commonly identified features between four neighboring image views. Images from left to right, top to bottom, are L1, L2, R1, and R2.

Figure 4. Partially occluded features between four neighboring image views. Images from left to right, top to bottom, are L1, L2, R1, and R2.

tures, each descriptor contains one 1×128 vector $d_i.v$ to record the gradient information in multiple directions and four scalars, $d_i.x, d_i.y, d_i.s, d_i.r$, for the row position, column position, scale, and orientation. Note that the commonly identified feature denotes the same corner, edge, or object that can be observed and precisely described with multiple camera views. Thus, each commonly identified feature in two pairs of input-rectified stereoscopic images consists of four standard SIFT feature descriptors:

$$\begin{aligned} CIF = \{&(d_{L1}.x, d_{L1}.y, d_{L1}.s, d_{L1}.r), (d_{L2}.x, d_{L2}.y, d_{L2}.s, d_{L2}.r), \\ &(d_{R1}.x, d_{R1}.y, d_{R1}.s, d_{R1}.r), (d_{R2}.x, d_{R2}.y, d_{R2}.s, d_{R2}.r)\}. \end{aligned} \tag{2.3}$$

2.2.3 Depth-constrained Feature Matching

In what follows, we will present the depth-constrained feature matching operation from four neighboring input images. The standard feature matching operation between two sets is always formulated as the problem to find the closest descriptor pairs in the given distance metric. The matching score between two SIFT feature in the L2 norm is defined as:

$$\epsilon(d_1, d_2) = \|d_1.v - d_2.v\|^2. \tag{2.4}$$

Then, for each descriptor d_1 in set S_{L1}, the $\hat{d}_2$ with minimal L2 distance will be selected as matched feature in set S_{L2}:

$$\hat{d}_2 = \underset{d_2 \in S_{L2}}{\operatorname{argmin}}\, \epsilon(d_1, d_2). \tag{2.5}$$

To construct the well-matched CIF set, we extend the above traditional matching operation into the four-set case. The score we used to evaluate the correspondence between four potential features is defined as follows:

$$\begin{aligned}\epsilon(d_1, d_2, d_3, d_4) = \lambda_1 \sum_{i=1}^{3} \sum_{j=i+1}^{4} (\|d_i.v - d_j.v\|^2) + \\ \lambda_2 \|d_1.y - d_3.y\|^2 + \lambda_2 \|d_2.y - d_4.y\|^2 + \\ \lambda_3 \|Depth.1 - Depth.2\|^2\end{aligned} \tag{2.6}$$

Four SIFT feature descriptors, d_1, d_2, d_3, and d_4, are randomly selected from four input images, S_{L1}, S_{L2}, S_{R1} and S_{R2}. The first score term refers to the Euclidean distance between any two feature descriptors in the L2 norm. The two following terms are the vertical disparity penalty terms, which address the matching accuracy between left and right view features. The last term is the horizontal disparity penalty term, which utilizes the depth-aware information to improve the matching reliability. $Depth.1$ can be computed from the triangulation between two matched features in image I_{L1} and I_{R1} while $Depth.2$ is from I_{L2} and I_{R2}:

$$Depth.1 = \frac{f \times b}{d_1.x - d_3.x}; \tag{2.7}$$

$$Depth.2 = \frac{f \times b}{d_2.x - d_4.x}, \tag{2.8}$$

where f is the focal length and b is the baseline distance between stereo camera pair. Thus, for any chosen feature descriptor, d_1, from the image I_{L1}, we can find its best-matched features in the other three images:

$$(\hat{d}_{1,2}, \hat{d}_{1,3}, \hat{d}_{1,4}) = \arg \min_{\substack{d_2 \in S_{L2} \\ d_3 \in S_{R1} \\ d_4 \in S_{R2}}} \epsilon(d_1, d_2, d_3, d_4). \tag{2.9}$$

Similarly, we can perform the above process for every feature descriptor from the other three detected SIFT feature sets:

$$(\hat{d}_{2,1}, \hat{d}_{2,3}, \hat{d}_{2,4}) = \arg \min_{\substack{d_1 \in S_{L1} \\ d_3 \in S_{R1} \\ d_4 \in S_{R2}}} \epsilon(d_1, d_2, d_3, d_4), \tag{2.10}$$

$$(\hat{d}_{3,1}, \hat{d}_{3,2}, \hat{d}_{3,4}) = \arg \min_{\substack{d_1 \in S_{L1} \\ d_2 \in S_{L2} \\ d_4 \in S_{R2}}} \epsilon(d_1, d_2, d_3, d_4), \tag{2.11}$$

$$(\hat{d}_{4,1}, \hat{d}_{4,2}, \hat{d}_{4,3}) = \arg \min_{\substack{d_1 \in S_{L1} \\ d_2 \in S_{L2} \\ d_3 \in S_{R1}}} \epsilon(d_1, d_2, d_3, d_4). \tag{2.12}$$

According to the different sources of chosen feature descriptors, four candidates for the CIF set (i.e., C_{L1}, C_{L2}, C_{R1}, and C_{R2}) are generated.

$$C_{L1} = \bigcup_{d_1 \in S_{L1}} \{(d_1, \hat{d}_{1,2}, \hat{d}_{1,3}, \hat{d}_{1,4})\}. \tag{2.13}$$

$$C_{L2} = \bigcup_{d_2 \in S_{L2}} \{(d_2, \hat{d}_{2,1}, \hat{d}_{2,3}, \hat{d}_{2,4})\}. \tag{2.14}$$

$$C_{R1} = \bigcup_{d_3 \in S_{R1}} \{(d_3, \hat{d}_{3,1}, \hat{d}_{3,2}, \hat{d}_{3,4})\}. \tag{2.15}$$

$$C_{R2} = \bigcup_{d_4 \in S_{R2}} \{(d_4, \hat{d}_{4,1}, \hat{d}_{4,2}, \hat{d}_{4,3})\}. \tag{2.16}$$

To ensure uniqueness and improve reliability, the verified CIF set is then defined as the intersection of the above four candidates:

$$C_v = C_{L1} \cap C_{L2} \cap C_{R1} \cap C_{R2}. \tag{2.17}$$

However, the exhaustive search and matching process suffers from low efficiency with time complexity of $O(n^4)$. We decide to accelerate the CIF construction process via some approximations. Instead of the direct matching operation among four input images, we divide the entire construction process into two steps: the matching between stereo image pairs and the matching between neighboring camera views. The first step is to find the corresponding 2D features between input stereo pairs (L_1 and L_2). The second step is to match these depth-aware features between the neighboring image capture positions (between camera position 1 and camera position 2). For better matching efficiency, the vertical and horizontal disparity obtained from the first step can also be employed as the qualifying condition to filter out these false-matched features before the computation of gradient vector distance. For more details, we first reduce the Equation 2.6 into the matching score between S_{L1} and S_{R1} as follows:

$$\epsilon(d_1, d_3) = \lambda_1 \|d_1.v - d_3.v\|^2 + \lambda_2 \|d_1.y - d_3.y\|^2. \tag{2.18}$$

With the reduced matching score, the standard matching operation can produce the corresponding SIFT pair set D_1:

$$\widetilde{d_3} = \arg \min_{d_3 \in S_{R1}} \epsilon(d_1, d_3). \tag{2.19}$$

$$\widetilde{d_1} = \arg \min_{d_1 \in S_{L1}} \epsilon(d_1, d_3). \tag{2.20}$$

$$D_1 = \bigcup_{d_1 \in S_{L1}} \{(d_1, \widetilde{d_3})\} \cap \bigcup_{d_3 \in S_{R1}} \{(d_3, \widetilde{d_1})\} \tag{2.21}$$

Similarly, we can obtain the SIFT pair set D_2 between S_{L2} and S_{R2}:

$$\epsilon(d_2, d_4) = \lambda_1 \|d_2.v - d_4.v\|^2 + \lambda_2 \|d_2.y - d_4.y\|^2 \tag{2.22}$$

$$\widetilde{d_4} = \arg \min_{d_4 \in S_{R2}} \epsilon(d_2, d_4). \tag{2.23}$$

$$\widetilde{d_2} = \arg \min_{d_1 \in S_{L2}} \epsilon(d_2, d_4). \tag{2.24}$$

$$D_2 = \bigcup_{d_2 \in S_{L2}} \{(d_2, \widetilde{d_4})\} \cap \bigcup_{d_4 \in S_{R2}} \{(d_4, \widetilde{d_2})\} \tag{2.25}$$

During these matching processes between stereo image pairs, the vertical disparity term is used as the first qualifying condition to increase efficiency. That means it will reject the potential SIFT descriptor with unaccepted vertical disparity before we compute its gradient vector distance to the target SIFT descriptor. Besides, we also utilize Kd-tree to partitions the potential SIFT set into a binary tree and then use the k-nearest neighbor to accelerate the standard matching process.

Given the two matched SIFT pair sets, D_1 and D_2, another round of the standard matching process can be conducted. At this moment, the known depth information in each element of the D_1 and D_2 are employed as other qualifying conditions. Those SIFT pairs with very different depth values will be filtered out before we start to compute their gradient vector distances.

According to the decompositions of exhaustive matching, we can improve the computation complexity from $O(n^4)$ to $O((logn)^2)$.

2.3 Feature-based Quality Metric

Feature-based Monocular Alignment Error

To quantify the performance of our proposed feature structure on various data sets, we use the projection position distance as the metric to evaluate the image alignment accuracy. After the warping of two neighboring camera views (e.g., L_1 and L_2)) into the output canvas, each key-point within overlapping region always corresponds to two source key-points from warped L_1 and L_2 respectively. In the ideal situation, these two source key-points on the output canvas are expected to have zero position displacement. Therefore, given the estimated projection function from the control points list P, we define the monocular view stitching error as the average position displacement of all the CIF in the

overlapping region. For each single $CIF = \{d_{L1}, d_{L2}, d_{R1}, d_{R2}\}$, its Root of Mean Square Error (RMSE) can be defined as:

$$E_L = \sqrt{(P(d_{L1}.x) - d_{L2}.x)(2) + (P(d_{L1}.y) - d_{L2}.y)^2}, \tag{2.26}$$

and

$$E_R = \sqrt{(P(d_{R1}.x) - d_{R2}.x)(2) + (P(d_{R1}.y) - d_{R2}.y)^2}. \tag{2.27}$$

$P(d.x)$ and $P(d.y)$ indicate the horizontal and vertical position of each CIF on the final output panorama after the projection. E_L and E_R refer to the specific CIF's projection error at left-view and right-view panorama, respectively. For each constructed CIF set C_v, the overall monocular alignment error can be defined as the average of all CIF's projection error:

$$E_{Mono} = \frac{1}{N}\sum_{i=1}^{N}(0.5 * E_L^i + 0.5 * E_R^i). \tag{2.28}$$

Usually, the smaller error implies fewer misalignment and discontinues, in other words, a better stitching quality in the monocular sense.

Feature-based Vertical Depth Error

To quantify the depth information error, we use the feature-wise depth accuracy to evaluate whether the output stereoscopic panorama carries the correct disparity value. In the ideal stitching result, the vertical disparity value between each two corresponded features in the binocular view is expected to be zero. Thus, the vertical disparity between the corresponding CIF at left-view and right-view panorama

implies the vertical depth accuracy of the system output. For each single $CIF = \{d_{L1}, d_{L2}, d_{R1}, d_{R2}\}$, the vertical depth error can be defined as:

$$E_{V1} = \sqrt{(P(d_{L1}.y) - d_{R1}.y)(2)}, \tag{2.29}$$

and

$$E_{V2} = \sqrt{(P(d_{L2}.y) - d_{R2}.y)(2)}. \tag{2.30}$$

$P(d.y)$ indicates the vertical position of each CIF on the final output panorama after the projection. E_{V1} and E_{V2} refers to the specific CIF's vertical depth error at 1st and 2nd camera view respectively. Give the constructed CIF set C_v, the overall monocular alignment error can be defined as the average of all CIF's vertical depth error:

$$E_V = \frac{1}{N}\sum_{i=1}^{N}(0.5 * E_{V1}^i + 0.5 * E_{V2}^i). \tag{2.31}$$

A small vertical disparity error between two final stitched panorama canvases indicates fewer vertical object jumping problems.

Feature-based Horizontal Depth Error

To define the depth error along the horizontal direction, we first assume that all the input image pairs can provide correct depth information for each CIF in the original spatial coordinate. Then, we construct the dense disparity map of the output stereoscopic panorama and obtain the measured horizontal disparity value for each CIF. Thus, the feature-based depth difference between the expected disparity map and the measured disparity map can be used to characterize the depth accuracy of output panorama.

Given the ground truth disparity map $Disp^{GT}$ and measured disparity Map $Disp^{M}$, for each single $CIF = \{d_{L1}, d_{L2}, d_{R1}, d_{R2}\}$, the horizontal depth error can be defined as:

$$E_{H1} = \sqrt{Disp^{M}(P(d_{L1}.x), P(d_{L1}.y)) - Disp^{GT}(d_{L1}.x, d_{L1}.y)}, \tag{2.32}$$

and

$$E_{H2} = \sqrt{Disp^{M}(P(d_{L2}.x), P(d_{L2}.y)) - Disp^{GT}(d_{L2}.x, d_{L2}.y)}. \tag{2.33}$$

$Disp^{GT}(d_L.x, d_L.y)$ indicates the ground truth of the horizontal depth in the system output, while $Disp^{M}(P(d_L.x), P(d_L.y))$ shows the measured horizontal disparity of each CIF on the final output panorama after the projection. E_{H1} and E_{H2} refer to the specific CIF's horizontal depth error at the 1st and 2nd camera view, respectively. For each constructed CIF set C_v, the overall horizontal depth error can be defined as the average of all CIF's horizontal depth error:

$$E_H = \frac{1}{N}\sum_{i=1}^{N}(0.5 * E_{H1}^{i} + 0.5 * E_{H2}^{i}). \tag{2.34}$$

A small horizontal depth error indicates the delivered depth in the stitched panoramas is close to the ground truth depth perceived from the input rectified camera views.

2.4 Quantitative Analysis

Based on the feature-based evaluation metric we defined above, we compute the monocular and stereoscopic stitching quality of four different panorama stitching solution: AutoPano, PTGui, Hugin, and our proposed method. The comparison includes one real captured indoor scene and 20 frames of synthetic outdoor scenes. Given the field of view and resolution of each input image, we can compute

the input pixel per degree (PPD). To construct one fair comparison, we usually scale the size of output panorama to $12,000$ by $6,000$ pixels for 360° x 180°. In this experiment, the input image usually hold 32.00 PPD (1,920 pixel for 60° FOV) and output panorama will hold 33.33 PPD ($12,000$ pixel for 360° FOV). Thus, the factor for panorama scaling is around 1.04. Those regions on the top and bottom of the panorama without valid pixels will be padded with the totally black pixels. The quantitative evaluation result is stated in Table I and Table II. It is noted that our proposed method attains the least monocular alignment projection error and the smallest depth errors in the vertical and horizontal direction, which indicates our proposed method outperforms other panorama stitching solutions in both the monocular and stereo sense

TABLE I

COMPARISON RESULT OF INDOOR SCENE

	Autopano	PTGui	Hugin	Proposed
RMSE	$19.61px$	$12.39px$	$9.04px$	$8.55px$
Vertical Disp	0.39°	0.20°	0.31°	0.20°
Horizontal Dist	0.42°	0.47°	0.35°	0.34°

TABLE II

COMPARISON RESULT OF SYNTHETIC OUTDOOR SCENES

	Autopano	PTGui	Hugin	Proposed
RMSE	$42.50px$	$47.97px$	$47.64px$	$11.59px$
Vertical Disp	0.56°	0.81°	0.42°	0.17°
Horizontal Dist	0.11°	1.21°	0.31°	0.04°

2.5 Visual Comparison

2.5.1 Monocular Stitching Comparison

The term monocular panorama refers to the left or right view of a stereoscopic panorama. Figure 5 shows the left-view panorama stitched by the PTGui, Autopano, Hugin, and our proposed method. In the areas marked by the yellow rectangles in the top three stitching output, the welding seam of the ceiling pipe is always misaligned or disconnected. However, our proposed method can handle this problem and produce a well-stitched overlapping region. The better stitching visual experience comes from the CIF instead of independently matched standard SIFT features. The matching information between features from one view can benefit the matching process in another monocular perspective. Thus, we can conclude that our proposed method outperforms the other three monocular-view stitching software alternatives in this situation.

(a) Autopano

(b) PTGui

(c) Hugin

(d) Proposed

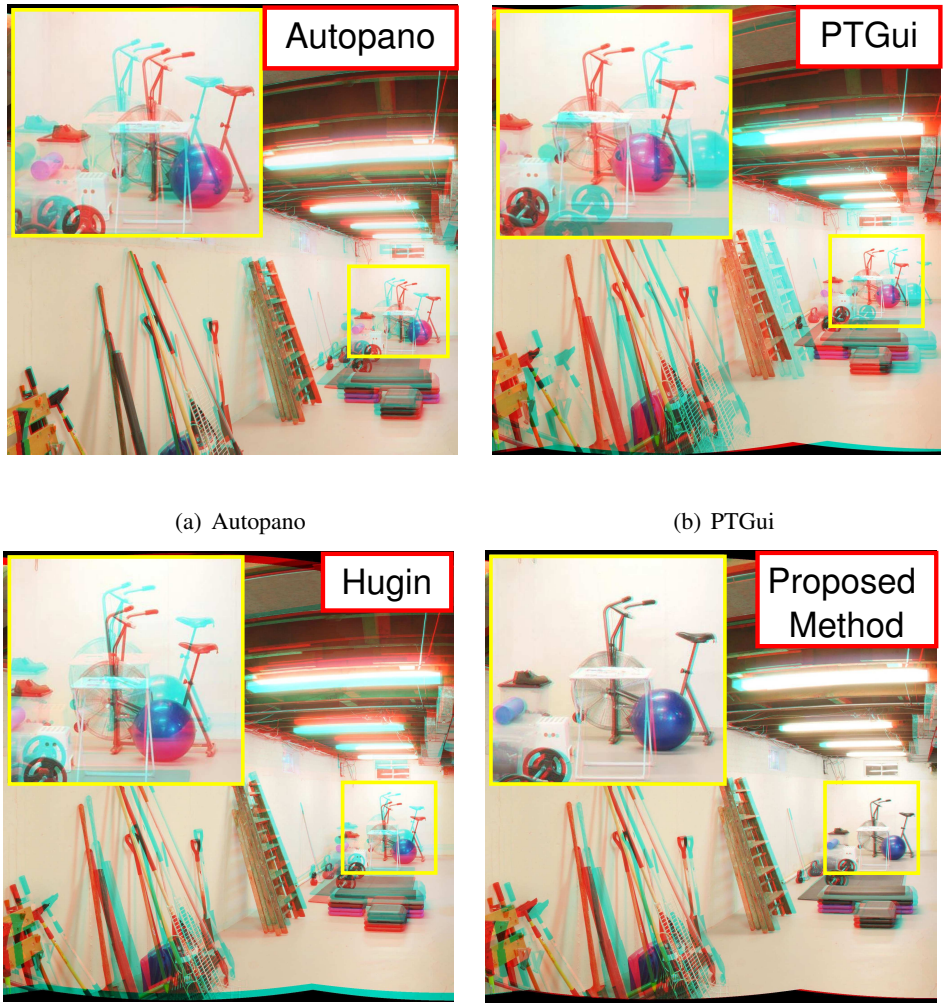
Figure 5. Comparison of monocular stitching result. Images from top to bottom, are left-view panorama stitched via AutoPano, PTGui, Hugin and our proposed method.

2.5.2 Binocular Stitching Comparison

Figure 6 shows the stereoscopic panoramas stitched by four different methods. In the areas marked by the yellow rectangles in the other three panoramas, the evident vertical and horizontal jumping of the bicycle results in severe viewing discomfort. The contradicted depth information carried with them can also confuse the viewer's perception of depth. In the panorama stitched by our proposed method, owing to the CIF set and transformation parameters that are as close as possible, the vertical disparity issue is barely detected, and the horizontal disparity of all objects is adjusted to one reasonable range that delivers correct depth information.

2.6 Conclusion

In this chapter, we explain the details of the proposed depth-constrained feature structure for the stereoscopic panorama stitching task. Compared to the standard 2D feature descriptor, the proposed CIF feature structure provides consistent points, edges, and areas to align the adjacent images at the left-view and right view simultaneously. Furthermore, the structure of CIF promises its easy access to the depth information of the described content. The carried extra depth information also contributes to the control matching accuracy before image alignment. Thus, the consistency between the binocular control point list cane be expected to maintain the stereo consistency between the generated binocular panoramas. In the next chapter, we discuss the proposed refinement techniques to these initially produced CIF and fit them into the video-stitching framework.



(a) Autopano (b) PTGui

(c) Hugin (d) Proposed

Figure 6. Comparison of stereoscopic stitching result. Images from left to right, top to bottom, are left-view panorama stitched via AutoPano, PTGui, Hugin and our proposed method.

CHAPTER 3

SALIENCY-BASED FEATURE SELECTION AND RE-DISTRIBUTION

The content of this chapter is based on our work that is published in [2] .©2018 IEEE. Reprinted with permission, from [2].

3.1 Background and Related Works

After we extract features from the input images, the standard step in the traditional panorama stitching framework is to send those control points into the feature tracker in the video sequence. However, in this section, we present one saliency-based feature selection strategy in the feature-based stereoscopic panoramic video generation system.

Note that both the features' quality and distribution play an essential role in the panorama stitching task. Once we obtain the CIF set for the four neighboring images in the first frame, we consider them as the control points for the current frame stitching. However, the initially produced control points are not always reasonably distributed because texture-rich regions bias the threshold. Whereas the number of detected features can be manipulated by altering the global threshold, owing to the equal weight of all the areas in the original images, fewer features are generated in the poor texture-rich region [20]. For example, in Figure 7, we can see most of the control points are clustered around the grassland, which indicates the latter image alignment will place more emphasis on the geometric correctness of it. The lack of enough control points distributed at the human's lower leg will likely cause the discontinues or misalignment at the human's lower leg in the stitching output. Thus, the first reason why we intend

to operate feature refinement is to relocate the position of all those CIF according to human visual interest. The second problem is over-sampled control points. To guarantee the detected features can

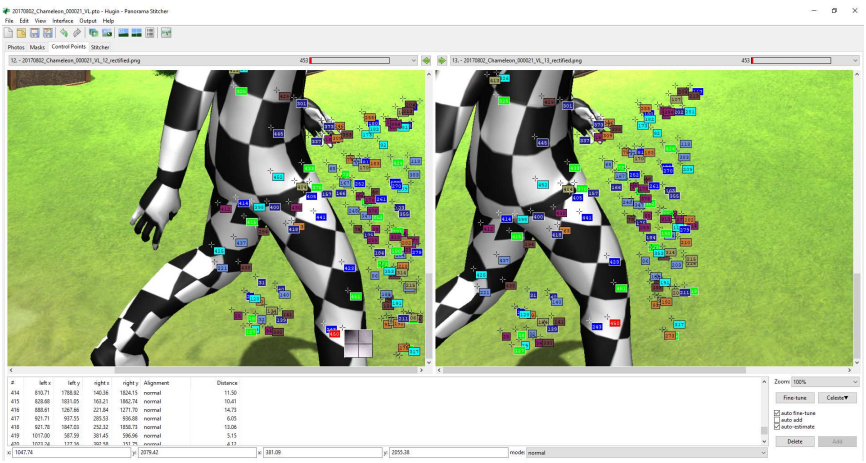
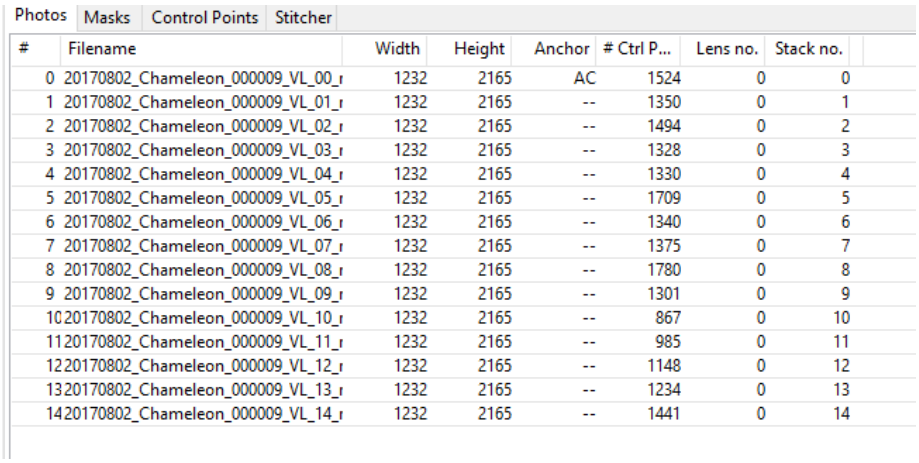


Figure 7. Example of CIF with unreasonable distribution.

be distributed across the whole image, direct utilization of the traditional feature detection algorithm (e.g., SIFT and speeded-up robust features [17, 21]) will always produce oversized control points. For example, in Figure 8, the CIF construction operation we described in the previous chapter produced 1000 features in the overlapping panel between each pair of neighboring camera views. If we accept all of them as the control points for the image alignment, there will be over 20,000 feature that needs to be tracked and aligned in later steps. A large number of control points always indicate high computation cost in feature tracking and image alignment step. The time complexity of KLT tracking and RANSAC



Photos | Masks | Control Points | Stitcher

#	Filename	Width	Height	Anchor	# Ctrl P...	Lens no.	Stack no.
0	20170802_Chameleon_000009_VL_00_r	1232	2165	AC	1524	0	0
1	20170802_Chameleon_000009_VL_01_r	1232	2165	--	1350	0	1
2	20170802_Chameleon_000009_VL_02_r	1232	2165	--	1494	0	2
3	20170802_Chameleon_000009_VL_03_r	1232	2165	--	1328	0	3
4	20170802_Chameleon_000009_VL_04_r	1232	2165	--	1330	0	4
5	20170802_Chameleon_000009_VL_05_r	1232	2165	--	1709	0	5
6	20170802_Chameleon_000009_VL_06_r	1232	2165	--	1340	0	6
7	20170802_Chameleon_000009_VL_07_r	1232	2165	--	1375	0	7
8	20170802_Chameleon_000009_VL_08_r	1232	2165	--	1780	0	8
9	20170802_Chameleon_000009_VL_09_r	1232	2165	--	1301	0	9
10	20170802_Chameleon_000009_VL_10_r	1232	2165	--	867	0	10
11	20170802_Chameleon_000009_VL_11_r	1232	2165	--	985	0	11
12	20170802_Chameleon_000009_VL_12_r	1232	2165	--	1148	0	12
13	20170802_Chameleon_000009_VL_13_r	1232	2165	--	1234	0	13
14	20170802_Chameleon_000009_VL_14_r	1232	2165	--	1441	0	14

Figure 8. Example of over-sized CIF.

algorithm is in $O(n)$ and $O(n^4)$. Thus, the size reduction of control points is one necessary operation for one efficient panorama stitching system. Besides, in the following homography estimation step, only a few of these initially detected control points would make contributions to the cameras pose estimation. Therefore, we wish to keep those control points that can give us insights for accurate image alignment and discard those convey redundant position information.

To solve these problems, we intend to propose one saliency-based and grid-based feature selection strategy to reduce the size of the control points and make them distribute across the image more uniformly and reasonably.

3.2 Saliency-based Feature Selection

Inspired by the image retargeting paper [22], which discusses the assignment of visual weight to different contents in the image, we propose a similar energy map that indicates the pixel-wise importance

of image alignment in our panorama stitching framework. The proposed energy is expected to assign weights to each pixel in the input image and guide the redistribution of the control points according to human visual attention. Furthermore, the idea of energy for the different regions is also used as the basic unit to predict the temporal change of saliency energy in the later tracking stage intuitively.

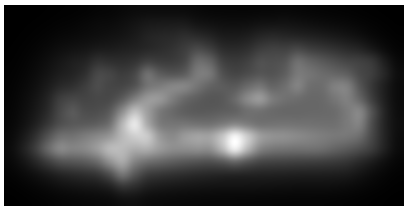
3.2.1 Saliency Map

In this part, we briefly discuss the saliency of the image before we move to the proposed feature selection part. In the neuroscience area, peripheral sensors generate afferent signals more or less continuously, and it would be computationally costly to process all this incoming information all the time. Thus, it is essential to make decisions on which part of the available information is to be selected for further, more detailed processing, and which sections are to be discarded. In the computer vision task, the conspicuity at different location in the visual field is called saliency and always represented by a scalar quantity. To facilitate the salient values assignments based on the spatial distribution, the saliency map always partitions an image into multiple segments (sets of pixels, also known as superpixels) and makes it easier to be analyzed. For instance, Figure 9 displays the input image, corresponding estimated saliency map, and the saliency map overlaid on the input image. According to the generated saliency map, we can quickly identify that the walking man area holds large saliency values, which means it will attract more attention from the viewers. In contrast, those regions assigned with lower saliency values usually stand for the areas without many visual essential objects, such as the ground or the sky.

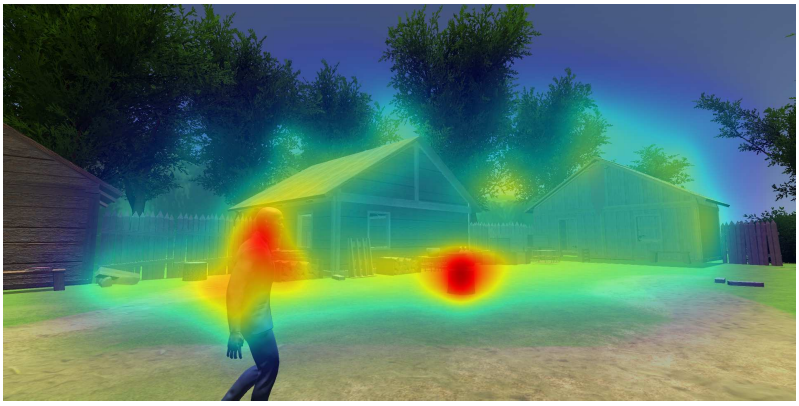
3.2.2 Energy Map Components

Given the concept of the saliency map, we propose one feature-based feature selection operation according to the human visual interest, which consists of three different maps:

(a) Input image



(b) Estimated saliency map



(c) Estimated saliency Map overlaid on Input Image

Figure 9. Example of saliency map.

1. Disparity Map
2. Saliency Map
3. Gradient Map

To generate a visual sensitivity map for the control points refinement, one energy fusion function [22] is used to combine the disparity map, gradient map and saliency map as:

$$e(i,j) = \alpha_1 \cdot Disp(i,j) + \alpha_2 \cdot Gradient(i,j) + \alpha_3 \cdot Sal(i,j). \tag{3.1}$$

In the above fusion function, (i,j) represents the pixel of the coordinate. Based on min-max normalization, the value of $Disp(i,j)$, $Gradient(i,j)$, and $Sal(i,j)$ are all normalized into [0,1]. In consequence, the intensity value of each pixel in the output energy map also ranges from [0,1].

Disparity Map Estimation

Disparity Mmap Disp(i,j) can provide the horizontal disparity value between two corresponded pixels in left-view and right-view panorama. Due to the equivalent triangles displayed in Figure 10, we can easily know the relationship between disparity and depth value:

$$disp(i,j) = x - x^{'} = \frac{B * f}{Z}. \tag{3.2}$$

In the above equation, B is the baseline distance between the optical center of the left-view and right-view camera, while f is the focal length of the stereo camera pair. Thus, the disparity value at position (i,j) is inversely proportional to the depth value at position (i,j). In other words, the pixel with a large

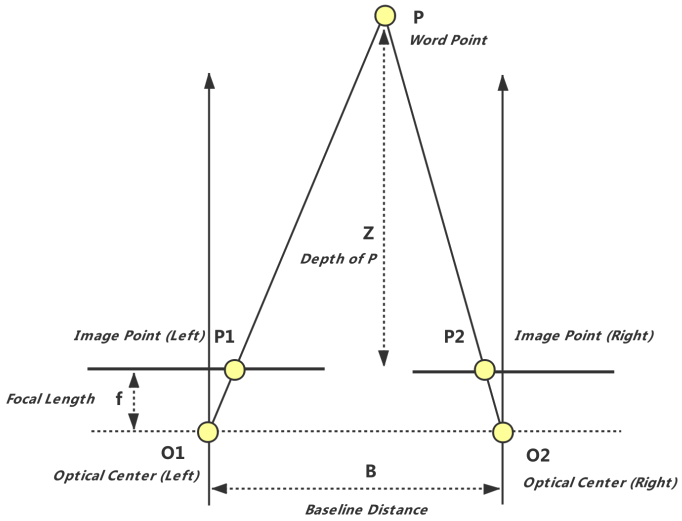


Figure 10. Depth from stereo.

disparity value is closer to the imaging plane of the capture equipment. In the panorama stitching task, those foreground objects usually attract more visual interest. Hence, we incorporate the disparity map into our proposed visual sensitivity map. Given the input image depicted in Figure 11, the corresponding estimated disparity map is depicted in Figure 12, and those areas with a whiter color imply closer pixels, which generally coincides with the shape of the walking man.

Gradient Map Estimation

The convolution to the original image with a filter can produce a gradient map. Each pixel of a gradient image measures the change in intensity of that same point in the original image, in a given direction. There are various gradient operators, such as Sobel, Prewitt, and Roberts. One example of

(a) Camera position 1 (b) Camera position 2

Figure 11. Input image for stitching

(a) Camera position 1 (b) Camera position 2

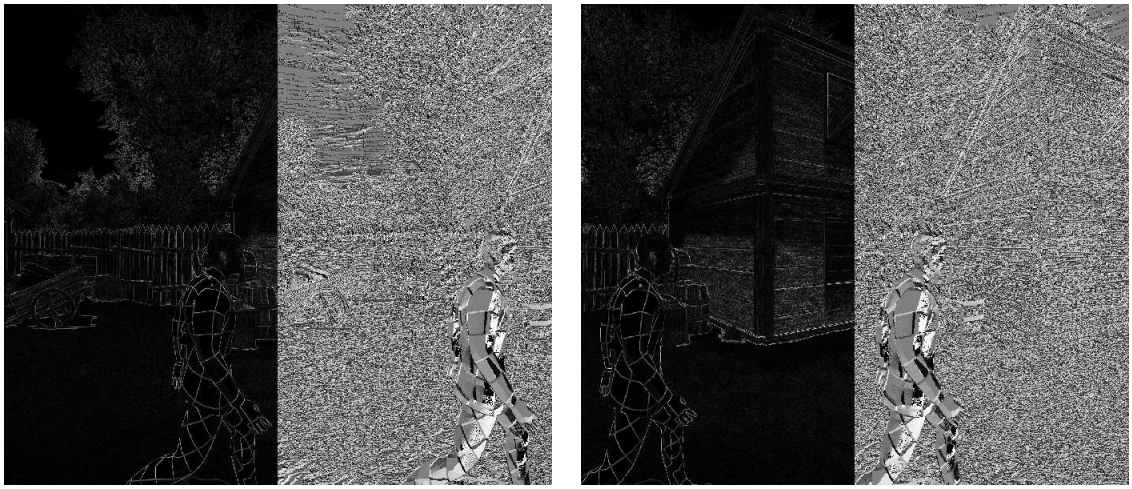
Figure 12. Disparity map

an estimated gradient map via the Sobel operator is depicted in Figure 13. Those edges, boundaries, and areas assigned with a whiter color implies larger intensity change compared to neighbors, and always attain more visual attention. Thus, the pixel-wise gradient information is also considered as one important component of the proposed visual sensitivity map.

Gradient image can be created by the convolution to the original image with a filter. Each pixel of a gradient image measures the change in intensity of that same point in the original image, in a given direction. There are various different gradient operators, such as sobel, prewitt and roberts. One example of generated Gradient Map via sobel operator is depicted in Figure 13. Those edges, boundaries and areas assigned with whiter color, which implies more intensity change compared to neighbors, always attain more visual attentions. Thus, the pixel-wise gradient information is also considered as one important component of the proposed visual sensitivity map. For more clarity, we temporally use the magnitude component of the gradient map for the final energy map composition.

Saliency Map Estimation

The algorithm we used to estimate the saliency map here is graph-based visual saliency (GBVS) [23]. It first computes the feature vectors at locations over the whole input image plane, analogous to the ITTI algorithm [24]. Then, one activation map is formed based on the generated feature vectors to highlight those significant pixels where the image carries some unusual information. In this step, the Markov chain is used to describe the dissimilarity between two potential pixels in the feature map. One example of the saliency map with black-white color code is depicted in Figure 14. It is noted that those regions with whiter pixels will attract more attention from viewers compared to the areas with darker pixels.



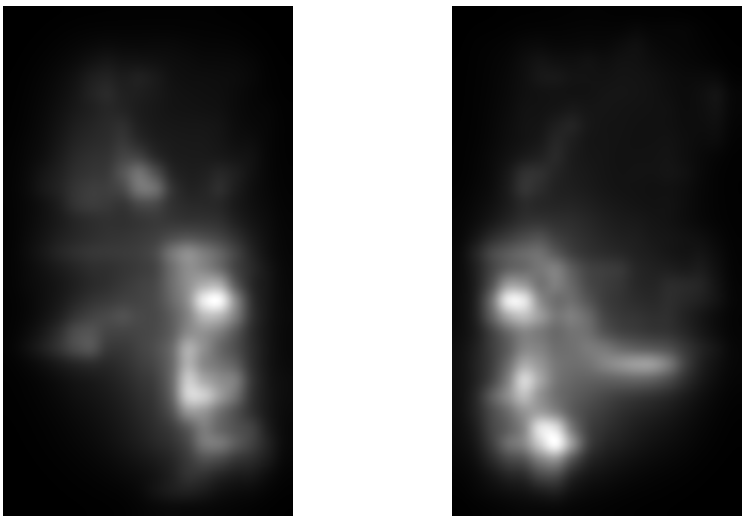
(a) Camera position 1 (b) Camera position 2

Figure 13. Gradient map

3.2.3 Energy Map Combination

Thus, the pixel-wise energy map for the human visual interest is the linear combination of the given three components. The generated map is shown in Figure 15. Here, we used the hot color code for better visualization of the difference in details between energy map and saliency map. It is noted that those regions with red color will be assigned a larger control point size in the later feature redistribution step.

In our panorama stitching task, for the four input images, we compute the corresponding energy maps of the overlapping region between four input images: E_{L1}, E_{L2}, E_{R1}, and E_{R2}. The matched CIF can help to estimate the boundary of overlapping regions. For example, the SIFT descriptor d_1 with the smallest x position in image I_{L1} will determine its left-most limit, and d_2 with the largest x position in image I_{L2} will determine the corresponding right-most boundary.



(a) Camera position 1 (b) Camera position 2

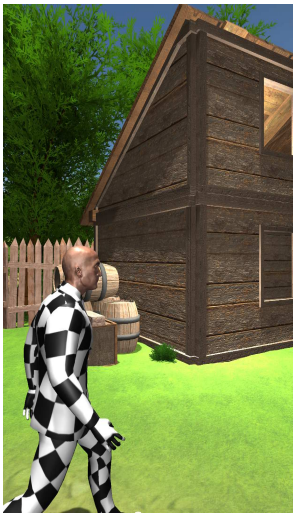
Figure 14. Saliency map with black-white color code

3.2.4 Feature Selection

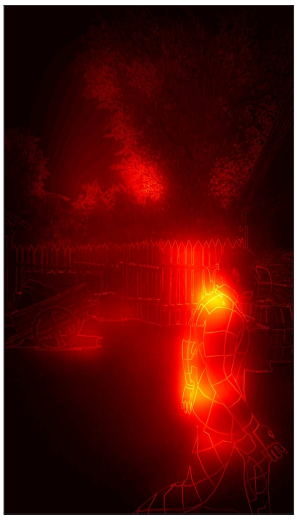
Then, for every single overlapping region, we fragment it into $M \times N$ grids. For instance, the overlapping region E_{L1} can be partitioned into: $\{G_{L1}^{p,q}, p \in \{1, 2, ...M\}, q \in \{1, 2, ...N\}\}$. For each grid at the p-the row and q-th column, its corresponding grid energy weight, $\hat{\omega}_{p,q}$, is defined as the normalized value of energy summation in the grid:

$$\omega_{p,q} = \sum_{(i,j) \in G_{p,q}} e(i,j) \tag{3.3}$$

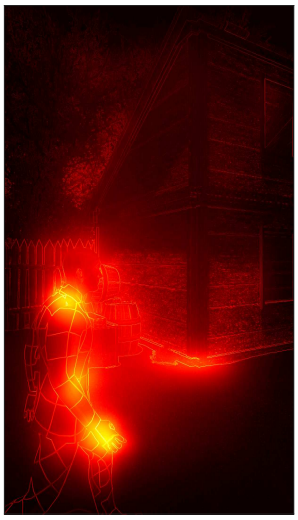
(a) Camera position 1



(b) Camera position 2



(c) Camera position 1



(d) Camera position 2

Figure 15. Combined energy map with hot color code. Sub-figure (a) and (b) are input image at camera position 1 and position 2. Sub-figure (c) and (d) are their corresponding energy maps.

$$\hat{\omega}_{p,q} = \frac{\omega_{p,q}}{\sum_{p,q} \omega_{p,q}} \tag{3.4}$$

The energy weight $\hat{\omega}_{p,q}$ represents the corresponding percentage of visual importance in the whole overlapping region. After the visual energy normalization in these four regions: E_{L1}, E_{L2}, E_{R1} and E_{R2}, we can use the average of them as the commonly-identified weight of all four corresponding grids:

$$\omega^c_{p,q} = (\hat{\omega}^{L1}_{p,q} + \hat{\omega}^{L2}_{M,q} + \hat{\omega}^{R1}_{p,q} + \hat{\omega}^{R2}_{M,q})/4. \tag{3.5}$$

Once we determine the total number we intend to track or the maximal limit of our computation power, we can compute the number of features we need to select in each grid:

$$B_{p,q} = T \times \omega^c_{p,q}. \tag{3.6}$$

In those texture-rich regions with over-sized features, we need to discard some less-reliable matched pairs or feature structures with redundant information feature squads based on our proposed ranking score [2]. The gradient difference, stereo-related term, and similarity penalty term are all taken into the construction for the proposed ranking score of one CIF structure:

$$R(d_1, d_2, d_3, d_4) = \beta_1 \cdot \epsilon(d_1, d_2, d_3, d_4) + \frac{\beta_2}{\epsilon(d_1, d'_1)}. \tag{3.7}$$

Because the small corresponding score between four pixels from Equation 2.6 usually implies high matching reliability of the feature, it is used to define the matching confidence term here. In the second

term, d_1' denotes the closest SIFT feature to the selected d_1 in the same grid. Thus, this redundant penalty term can inhibit the selection of two similar or identical features.

All commonly identified features in this grid are sorted in ascending order of the proposed ranking scores, R. The first $B_{p,q}$ commonly identified features in each grid are regarded as control points for image alignment. If there are not enough potential CIF in the grid, then we keep all of them. One example of the comparison between before and after the saliency-based selection is depicted in Figure 16. The features after the saliency-based selection, which are the sub-figure (c) and (d), are more even and reasonable across the whole overlapping region.

After we walk through all the girds, according to the different sources of energy maps, four candidates that refined the CIF set (i.e., C_{L1}^r, C_{L2}^r, C_{R1}^r, and C_{R2}^r) are generated. The final refined CIF set is then defined as the intersection of the above four candidates:

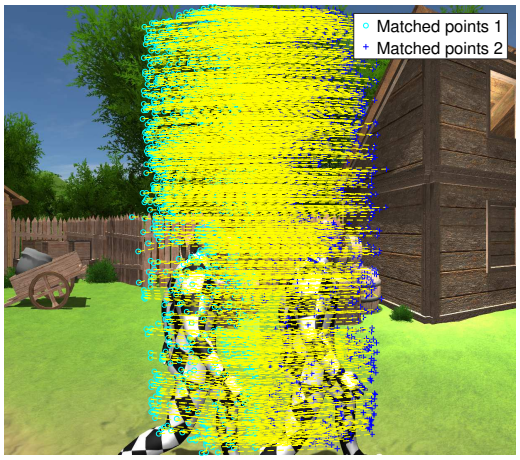
$$C_v^r = C_{L1}^r \cap C_{L2}^r \cap C_{R1}^r \cap C_{R2}^r. \tag{3.8}$$

3.3 Quantitative Analysis

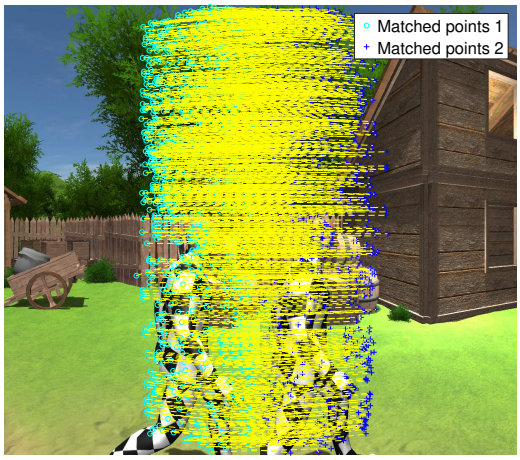
To test whether our proposed feature selection strategy contributes to stitching quality improvement in the stereoscopic panoramic video, we generally compare the stitching result before and after saliency-based feature selection:

1. No feature selection strategy (NFS)

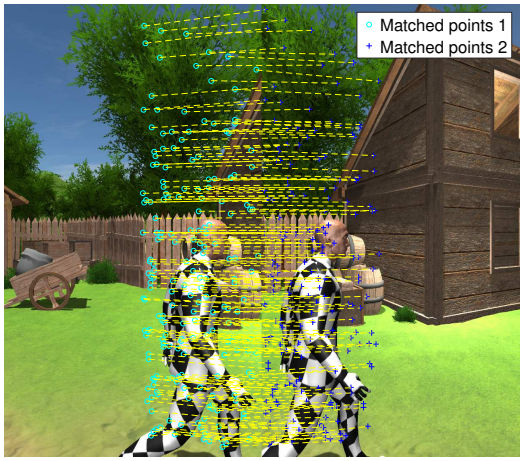
2. Saliency-based feature selection strategy (SFS)



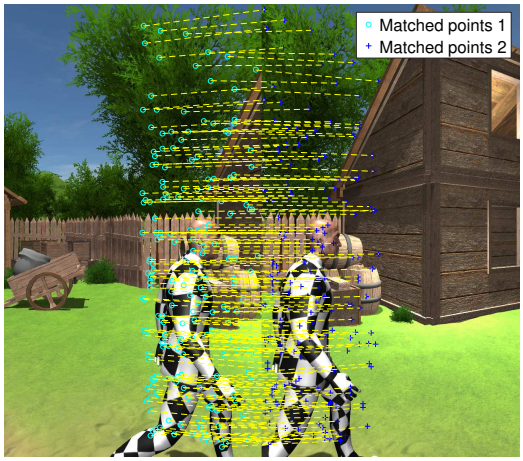
(a) Matched feautes before selection in left view



(b) Matched feautes before selection in right view



(c) Matched feautes after selection in left view



(d) Matched feautes after selection in right view

Figure 16. Matched features pairs before and after saliency-based feature selection.

The experiment's data consists of three synthesized outdoor videos, which describe one walking man in a circular path with a different radius. Every single video includes 30 frames. In our experiments, the overlapping regions of four neighboring images are all divided into square grids of 100 pixels wide by 100 pixels high. The number of the selected control points is 200 for each overlapping region. All output panoramas are scaled to $12,000$ by $3,000$ pixels for $360^\circ \times 90^\circ$. The input image usually hold 32.00 PPD ($1,920$ pixel for 60° FOV) and output panorama will hold 33.33 PPD ($12,000$ pixel for 360° FOV). Thus, the factor for panorama scaling is around 1.04. The three coefficients in the energy map generation are set as 0.33. The two coefficients β_1 and β_2 in Equation 3.7 are set as 0.70 and 0.30.

In the numerical comparison between our proposed method and the standard selection strategy, we mainly focus on the stitching quality of the single frame in the video. The feature-based projection error is used to evaluate the accuracy of alignment. For the quantitative analysis of the stereoscopic panoramic video in the vertical direction, we measured the average vertical disparity of all matched features between the left and right views. For the horizontal direction, we first consider the estimated depth from the original rectified image pair as the ground truth. The average distance of all matched features between the depth from stitched stereoscopic panoramas and the depth from the ground truth is then used as the metric to evaluate the performance of depth control.

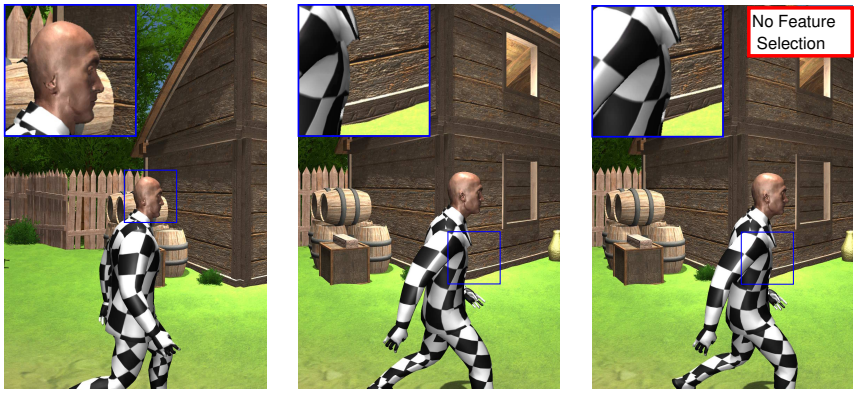
The numerical result of 30 frames of synthetic outdoor scenes in the different radii of the circular path is shown in Table III. It is noted that the overall stitching quality is improved as the distance between the principal moving object and image acquisition equipment decreases, in both the monocular and binocular sense. That trend indicates the high-quality stitching becomes a more challenging task when the major moving object is too close to the camera imaging planes. However, the CIF feature set after

the saliency-based selection could provide smaller projection distances between matched control points in the output stereoscopic panorama. For the stereo sense quality, saliency-based feature selection can efficiently reduce the vertical and horizontal depth error simultaneously. Thus, this quantitative analysis demonstrates that our proposed feature selection strategy can lessen the stitching misalignments in the output monocular-view panorama and obtain better stereo consistency between stereoscopic panoramas.

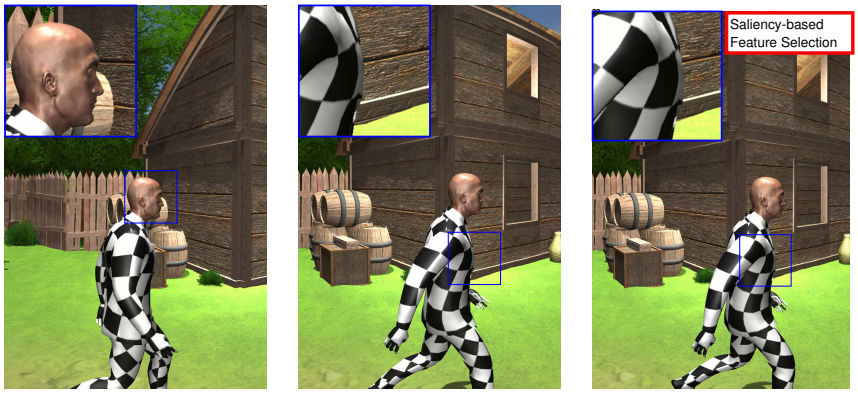
TABLE III

COMPARISON RESULT IN CIRCULAR PATH OF DIFFERENT RADIUS

	Mono Error	Vertical Error	Horizontal Error
NFS+1.3m	$9.73px$	$0.19°$	$0.94°$
NFS+2.0m	$2.46px$	$0.13°$	$0.63°$
NFS+3.3m	$1.81px$	$0.10°$	$0.35°$
SFS+1.3m	$8.45px$	$0.05°$	$0.12°$
SFS+2.0m	$2.06px$	$0.06°$	$0.12°$
SFS+3.3m	$1.03px$	$0.05°$	$0.11°$



(a) frame 9 in NFS strategy (b) frame 20 in NFS strategy (c) frame 24 in NFS strategy

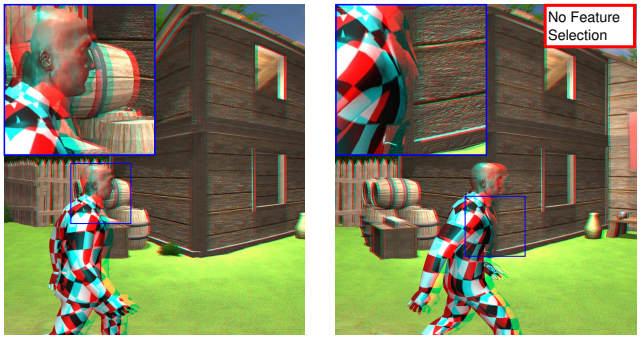


(d) frame 9 in SFS strategy (e) frame 20 in SFS strategy (f) frame 24 in SFS strategy

Figure 17. Comparison of left view video stitching result between NFS and SFS.

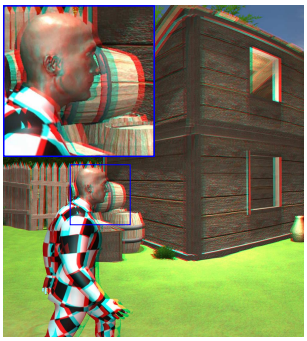
3.4 Visual Comparison

Figure 17 shows several left- view panoramas stitched by no feature selection strategy and our proposed strategy. In the top three panoramas, the walking man suffers from several visible stitching errors, such as the distortion of the head at the 9th frame, the discontinues, and cropping of the human's chest on the 20-th and 24-th frame. However, these stitching errors are barely detected in the output

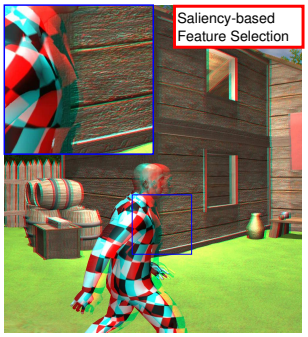


(a) frame 9 in NFS strategy

(b) frame 20 in NFS strategy



(c) frame 9 in SFS strategy



(d) frame 20 in SFS strategy

Figure 18. Comparison of stereoscopic stitching result between NFS and SFS in red-cyan anaglyph version.

panorama after feature selection, which implies more reasonable alignment under CIF after our proposed feature refinement strategy

Figure 18 shows several stereoscopic panoramas stitched by no feature selection strategy and our proposed strategy in the red-cyan anaglyph version. In the top row, these obvious monocular stitching errors deliver contradicted depth information of the human head and chest in the stereoscopic video and

result in severe viewing discomfort. The visible stitching error is caused by the inadequate sampled control points in the suit with a chessboard texture. However, our proposed method can handle this problem and produce the close walking man as smoothly stitched. Based on the optimized distribution of control points that assign an adequate percentage of features to those visual sensitive regions, the homography estimation will be operated under the guidance of human attention. Thus, better monocular stitching quality and correctly delivered depth information in those visually sensitive regions are expected.

3.5 Conclusion

In this chapter, we present one feature-based selection strategy that reduces the control point size and improves the distribution of control points in the panoramic video generation system. For this goal, we utilize the energy map that consists of a disparity map, saliency map and gradient map to compute the visual importance of each pixel. Under the guidance of these pixel-wise maps, we divide the overlapping region into grids and decompose the control points optimization problem into multiple ranking problems in each grid according to our proposed matching score. In every single grid, CIF with a higher matching reliability term and smaller similarity penalty term will be selected as the control points for the image alignment. To achieve the well-stitched output successfully and efficiently, the number of the necessary control points, grid window size, and coefficients of the energy map generation need to be carefully specified. In the next chapter, we discuss the corresponding tracking strategy to these refined control points in the video sequence.

CHAPTER 4

DEPTH-CONSTRAINED FEATURE TRACKING

The content of this chapter is based on our work that is published in [2] .©2018 IEEE. Reprinted with permission, from [2].

4.1 Background and Related Works

Compared to image stitching, panoramic video-stitching has received far less attention. Many video-stitching methods either adopted a fixed dominant homography from one selected frame to align the whole video [25, 26] or conducted frame-stitching independently [27]. Shimizu et al. proposed a video-stitching scheme that used pure translation motion for sporting events [28], which can only deal with the simple translation case. Xu and Mulliga used a multi-grid scale-invariant feature transform (SIFT) to show acceleration and combined SIFT feature sets from randomly selected frames to one common homography [29]. Recently, Jiang and Gu proposed an algorithm to utilize spatial-temporal content-preserving warping [30] to composite one output panoramic video from multiple synchronized input video streams. However, they only used the matched features in the first K frames to define one commonly-used homography for all remaining video frames and never update features later.

Regarding video stitching from free-moving mobile devices, other works consider stabilization for better stitching [31–36]. These algorithms usually only utilize the estimated camera path to combine different camera views and did not take advantage of the calibrated geometries between adjacent cam-

eras. Hence, one generalized control point sampling strategy is needed in our proposed video-stitching framework for good quality of output video.

In previous chapters, we present the proposed optimization and extension to the standard stereoscopic panorama generation system. While our ultimate goal is the stereoscopic 360 video, we will discuss the corresponding strategy to deal with the panorama stitching in the video case. To follow with the human visual interest idea in Chapter 3, we present one saliency-based feature update strategy in the feature-based stereoscopic panoramic video generation system. Based on the CIF detection and matching we introduced in Chapter 2, we still utilize it as the basic stitching unit in later frame composition to achieve stable stitching output in the video task. The temporal consistency between consecutive frames can be interpreted as consistencies in geometry, vertical disparities, and horizontal disparities. For more detail, the geometric consistency indicates the shape, size, and relative location of objects that should remain identical between the prior and consecutive frames. The temporal consistency in vertical and horizontal directions guarantees there will be no abrupt changes in the perceived depth from the same object. We employ human visual sensitivities to generate a grid- and saliency-based energy map to indicate the visual importance of pixels. Then, the global temporal feature- tracking can be decomposed into several grid-based local tracking tasks according to changes in pixel energy. To further improve the accuracy of commonly identified features, we extend the underlying assumption of small-displacement into the depth domain, removing the falsely tracked control points. Moreover, to compensate removed tracked control points that violate the commonly-identified property, we can detect new CIF from the new frame and incorporate the position data from the previous rejected CIF into the new CIF selection.

4.2 Feature Update Strategy

Given the refined CIF set, we start to deal with the tracking operation of those control points in later frames. One straightforward way is directly utilizing the tracked feature points in the next frame via different objective tracking algorithms. However, the pure tracking-based feature selection is always affected by occlusion, drift, and loss in the long-term video. It is nearly impossible to generate one well-stitched stereoscopic panoramic video based on the initialized feature set.

Another promising method uses the newly detected commonly identified feature set for the current frame, stitching each frame independently. This method, based on detection, can mostly avoid the monocular geometrical stitching errors of each frame. However, it always causes poor temporal consistency. Another reason why the detection-based feature selection is not a good choice for video stitching is its inefficiency, compared to the tracking-based method.

To avoid the problems mentioned above, we propose a local saliency-based feature tracking strategy that focuses on the temporal energy change of each grid. Thus, instead of stitching based on tracked features or newly detected features, we update the control points in the necessary grid and operate image alignment based on these local hybrid feature sets. The proposed feature-tracking strategy can be divided into three parts. Grid update determination tells us which grid needs a control-point update. Depth-constrained feature tracking refines the KLT tracking result with the depth-related conditions. Feature compensation detects new features, making up for invalid CIF at the new frame.

4.2.1 Grid Update Determination

Before we send initially selected features, $\hat{S}_{i,j}$, to the KLT tracker, we should determine each divided grid, based on its energy change. As in the generation of energy maps from the first frame, we compute

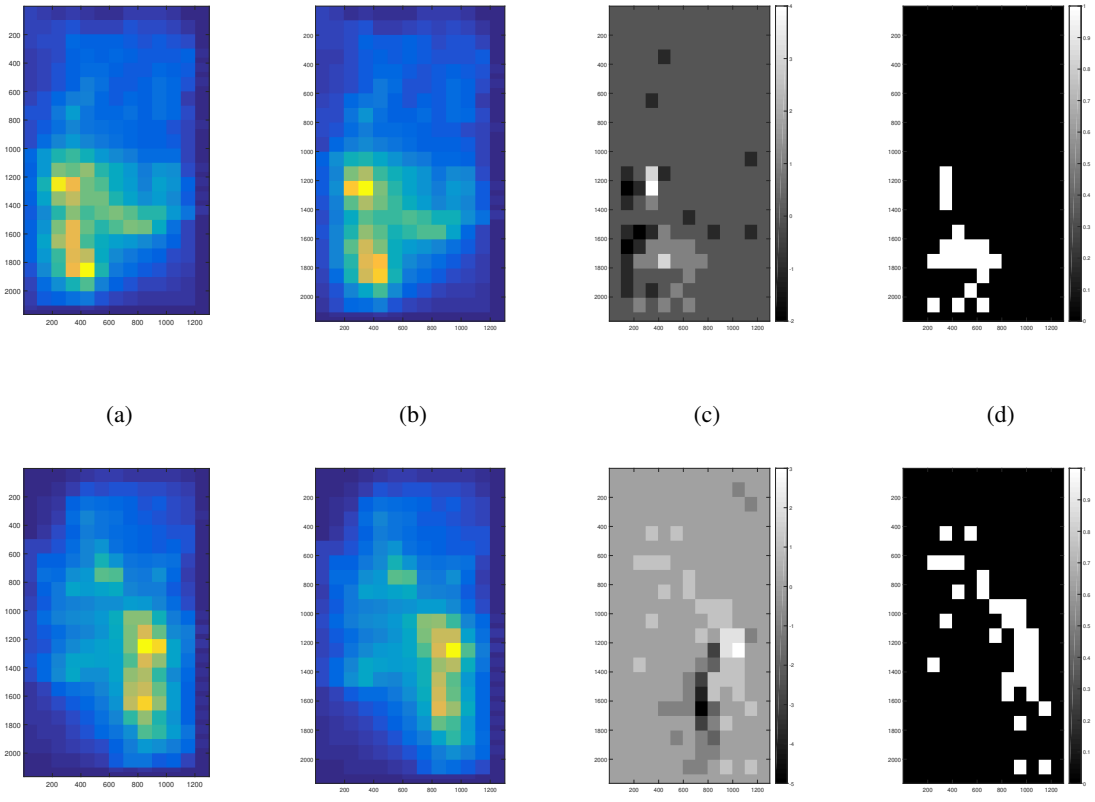
pixel-based energy changes, $\delta(i,j)^{n+1} = e(i,j)^{n+1} - e(i,j)^{n+1}$, for pixels in all grids. The grid-based energy change for the grid at the row p and column q, between frame n and $n+1$, is defined as:

$$D_{p,q} = Mean(\delta(i,j); (i,j) \in G_{p,q}). \tag{4.1}$$

Those grids, associated with tiny energy changes, are considered as still regions (e.g., background). Generally, no object enters or leaves those grids. Therefore, it is unnecessary to operate tracking processes, because those control points inherited from a previous frame can describe the current frame quite well. Thus, we only focus on control points in those grids having energy changes. The original grid is assigned to two groups, still and dynamic, according to their grid-based energy changes. Then, we pass the control point information of the dynamic grid to the KLT tracker in a later step. An example of the grid after finishing the update determination is shown in Figure 19. Sub-figure(a) and sub-figure(b) are the grid-wise energy map for camera position 1 at two consecutive frames. Sub-figure(c) is the difference energy map between two consecutive frames. Sub-figure(d) is the difference energy map after we convert Sub-figure(c) into the black-white map. Grids with black color mean there is no need to update control points, and grids with white color imply the features update is necessary. Sub-figure(e) to sub-figure(h) are the corresponding maps for camera position 2. In Figure 19, the jet and black-white color maps are only used for visualization, while no color information used in the grid update determination.

4.2.2 Depth-constrained Feature Tracking

After we determine whether each grid needs an update or not, we update the control points and utilize the traditional KLT to track the control point from the previous frame. In the standard formulation statement for the KLT tracking problem, I and J refer to the previous and current images. $I(x,y)$



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 19. Left part: (a) Grid-based energy at 1st nd 2nd right camera view at frame n; (b) Grid-based energy at 1st nd 2nd right camera view at frame n+1; (c) Grid-based energy difference; (d) Grid-based indicator map for feature update.

represents the gray value of the pixel at $(x, y)^T$. Let $\mathbf{u} = [u_x, u_y]^T$ be one pixel of the previous image, I. The goal is to find one pixel, $\mathbf{v} = \mathbf{u} + \mathbf{f}$, on the current image, J, where the intensity distance

between two integration windows around $I(\mathbf{u})$ and $J(\mathbf{v})$ is minimized. With the integration window size parameter, w_x and w_y, the residual function, $\epsilon(\mathbf{f})$, between two integration windows is defined as:

$$\epsilon(\mathbf{f}, \mathbf{I}, \mathbf{J}, \mathbf{u}) = \sum_{x=u_x-w_x}^{x=u_x+w_x} \sum_{y=u_y-w_y}^{y=u_y+w_y} (I(x,y) - J(x+f_x, y+f_y)). \tag{4.2}$$

In our case, we need to operate a tracking process for the CIF set, $\{d_1, d_2, d_3, d_4\}$, simultaneously between consecutive frames from four neighboring cameras views. Note that the target pixel, **u**, in the residual function, refers to the key point position of the target feature. Thus, we obtain four motion vectors:

$$\begin{aligned}
\boldsymbol{f}_1 &= arg\min_{\boldsymbol{f}} \epsilon(\boldsymbol{f}, I_{L1}^n, I_{L1}^{n+1}, d_1) \\
\boldsymbol{f}_2 &= arg\min_{\boldsymbol{f}} \epsilon(\boldsymbol{f}, I_{L2}^n, I_{L2}^{n+1}, d_2) \\
\boldsymbol{f}_3 &= arg\min_{\boldsymbol{f}} \epsilon(\boldsymbol{f}, I_{R1}^n, I_{R1}^{n+1}, d_3) \\
\boldsymbol{f}_4 &= arg\min_{\boldsymbol{f}} \epsilon(\boldsymbol{f}, I_{R1}^n, I_{R2}^{n+1}, d_4).
\end{aligned} \tag{4.3}$$

However, the straightforward independent tracking of these four images cannot ensure good updated positions for the initial feature descriptors. Any mistakenly tracked result or drift in these four KLT tracking processes will lead to the failure of new CIF constructions in the current frame.

To obtain more reliable tracked results of control points in the current frame, we utilize the depth-based constraint to qualify those tracking results strictly. The computed motion vectors for the four feature descriptors are $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \text{and } \mathbf{f}_4\}$, based on the assumption of the target pixel's small-displacement

between consecutive frames in the traditional KLT tracking algorithm. We extend this assumption from the 2D to 3D case. Hence, we formulate two depth-based criteria to qualify these independent tracking features:

1. The difference of horizontal disparity between previous and current frame should be small;
2. There is no visible vertical disparity between left and right views for tracked, commonly identified features in the current frame.

Given that the CIF at the frame n is:

$$CIF_n = \{d_1^n, d_2^n, d_3^n, d_4^n\}, \tag{4.4}$$

and the tracked feature at frame $n+1$ is:

$$CIF_{n+1} = \{d_1^{n+1}, d_2^{n+1}, d_3^{n+1}, d_4^{n+1}\}, \tag{4.5}$$

the horizontal temporal disparity between two frames and the vertical disparity of the new frame can be represented as:

$$
\begin{aligned}
\epsilon_h(n,1) &= \|(d_1^n.x - d_3^n.x) - (d_1^{n+1}.x - d_3^{n+1}.x)\| \\
\epsilon_h(n,2) &= \|(d_2^n.x - d_4^n.x) - (d_2^{n+1}.x - d_4^{n+1}.x)\| \\
\epsilon_v(n,1) &= \|(d_1^{n+1}.y - d_3^{n+1}.y)\| \\
\epsilon_v(n,2) &= \|(d_2^{n+1}.y - d_4^{n+1}.y)\|
\end{aligned}
\tag{4.6}
$$

Based on our proposed two depth-constrained qualifying conditions for the commonly-identified property, we filter all tracked CIFs with $\widetilde{\epsilon_h}$ and $\widetilde{\epsilon_v}$. The tracked control point at the new frame with disparity values larger than two tolerance values will be rejected as the falsely tracked CIF. Only those features having small depth changes and with invisible vertical disparities can be regarded as reliable control points and pushed into the list for camera pose estimation. The proposed small-assumption in the depth domain is usually set as 10 pixels in our experiments.

4.2.3 Feature Compensation

In the traditional feature update strategy, combined with tracking and detection, one new feature detection will be operated when the successfully tracked features drop below a given threshold [35]. However, the size of the control point list remains shrunk until the next feature detection operation. Thus, newly feature detection will not be conducted when the number of successfully tracked features remains in the accepted range. Therefore, we prefer to operate the feature update step wise and to decompose changes of control point lists into tiny updates between frames.

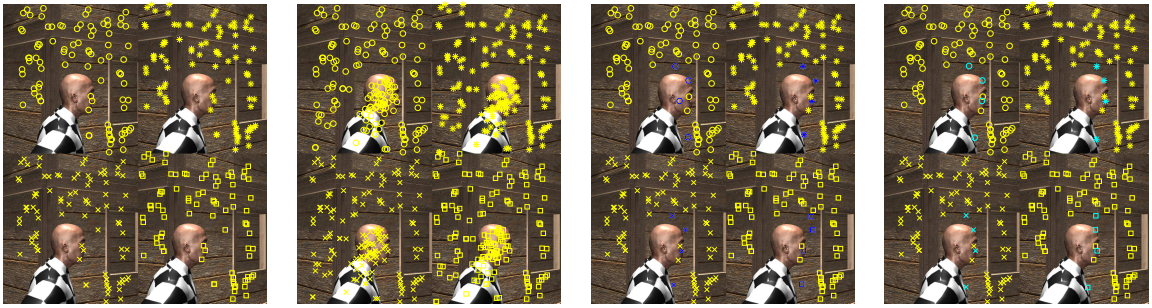
According to the depth-constrained qualifying condition, we identify invalid tracked features and remove them from the control point list. Thus, to obtain the size consistency of control points in each

grid, we fill missing slots with newly detected features. Normally, we select CIF with the highest matching scores. In our proposed matching score, we also incorporate the position information of commonly identified features of the normal ranking score so that the temporal consistency of the selected control points can be obtained:

$$R_2(d_1^{n+1}, d_2^{n+1}, d_3^{n+1}, d_4^{n+1}) = R(d_1^{n+1}, d_2^{n+1}, d_3^{n+1}, d_4^{n+1}) + \frac{\beta_3}{\sum_{i=1}^{4}[\sqrt{(d_i^{n+1}.x - d_i^n.x)^2 + (d_i^{n+1}.y - d_i^n.y)^2}]}. \tag{4.7}$$

In this modified score for compensated feature selection, the second term is used to describe the position distance between newly detected features at frame n+1 and old commonly identified features at frame n. The small distance indicates less of a positional change in the control point list and more consistently estimated camera poses

In Figure 20, we see a comparison between different feature update strategies. Sub-figure (a) is the initially detected CIF in the previous frame. Sub-figure (b) is the update result based on pure detection in the next frame. Sub-figure (c) is the update result based on pure tracking at the next frame, and those key points marked with blue colors are invalid control points that fail to fulfill the proposed depth-constrained qualifying condition. Sub-figure (d) is the update result based on our proposed update strategy, and the key points marked with cyan colors are the corresponding compensated features to those invalid control points in Sub-figure (c). The tracking result displayed in sub-figure (b) shows more newly detected CIFs clustered around the human's head, which will break the control points consistency in the temporal domain. The tracking result displayed in sub-figure (c) contains that control



(a) (b) (c) (d)

Figure 20. Comparison of stitched panoramas between different softwares and proposed method. Sub-figure (a) is the initially detected CIF in the previous frame. Sub-figure (b) is the update result based on pure detection in the next frame. Sub-figure (c) is the update result based on pure tracking at the next frame, and those key points marked with blue colors are invalid control points that fail to fulfill the proposed depth-constrained qualifying condition. Sub-figure (d) is the update result based on our proposed update strategy, and the key points marked with cyan colors are the corresponding compensated features to those invalid control points in Sub-figure (c).

points without commonly-identified property that may impair the stereo consistency between left-view and right-view output panorama.

4.3 Conclusion

In this section, we propose one saliency-based control point update strategy to deal with the 360 video-stitching task. Since the first frame in the video sequence has been aligned under well-matched control points and stereo-constrained homography, the following frames should be processed with the

same routines. However, due to the spatial position change of objects in different frames, one highly reliable CIF in the previous frame may become the POF in the next frame. To maintain the temporal consistency in the stitched 3D 360 video, we operate one grid-based feature update strategy based on the human visual interest difference between two successive frames. Those areas with an abrupt change of human visual interest indicate a significant change of content. Hence, we adjust the size of control points in all the updated grids via CIF pruning or CIF compensation for more reasonable image alignment. In the next chapter, we discuss the details of image alignment based on the control point we generate for each frame in the video sequence.

CHAPTER 5

STEREO-CONSTRAINED IMAGE ALIGNMENT

The content of this chapter is based on our work that is published in [1] .©2017 IEEE. Reprinted with permission, from [1].

5.1 Background

After we extract a well-matched CIF feature set from the input images and conduct saliency-based feature selection, the next step in the standard panorama stitching framework is called image alignment, which intends to estimate the global homography with these control points. In the monocular stitching pipeline, the RANSAC algorithm of Fischler and Bolles [37] is usually adopted for homography estimation for the adjacent camera pose computation. The RANSAC algorithm is a robust estimator for mathematical model parameters based on a set of observed data containing outliers. For more details about its application in the panorama stitching case, the model presents a planar homography between two camera views, and the observed data corresponds to the detection of 2D-point correspondences. Without considering the stereo constraint, the standard RANSAC algorithm [38] can fit the best transformation matrix and generate a high-quality monocular panorama for the left and right views. However, the two independently well-stitched monocular panoramas cannot guarantee correct delivered depth information and an excellent 3D viewing experience. Thus, we proposed a modified RANSAC algorithm incorporated with a stereo constraint to maintain consistency between the left and right views. In this

chapter, we mainly discuss the original RANSAC in the homography estimation task and its adapted version based on the CIF for the stereoscopic panoramic stitching case.

5.2 Standard Image Alignment

The application of original RANSAC to homography estimation is described in Algorithm 1.

The standard RANSAC-based homography estimation starts with the random selection from the given input dataset. Let us assume we already produced one well-matched pair of control points $\mathcal{T}$ between two neighboring images, L_1 and L_2. Then, we can randomly draw four pairs of corresponded control points from the data-set: $\{(f_j, f_j^{'}), j = 1:4\}$. Based on the eight equations from 4 pairs of matched points, we can fit the projection matrix H via normalized direct linear transform. Then, those unselected control points will be used as the test set to evaluate the fitness of this estimated H to all control points. The projection error of each corresponded pair from the testing set will be computed. If the error is within the accepted range, there will be one consistent count added to this estimated homography. After we go through all the possibilities of the random selection or the predefined iteration number runs out, the homography with the largest percentage of the consistent count will be considered as the best-fitted projection matrix H. For more details about parameters set-up, the corresponding data set $\mathcal{T}$ in Algorithm 1 is defined as one well-matched pair of control points. The inlier percentage p is always set as 95%. The distance threshold σ need to be specified under different input datasets; we usually set them as 10 pixels in our experiments. The max number of iteration N is determined by the trade-off between the estimation accuracy and computation time.

5.3 Modified Image Alignment

The application of modified RANSAC to homography estimation is described in Algorithm 2.

In the modified RANSAC, we randomly select corresponding points from the commonly identified feature set, so that the sampled subset data for model-fitting and the remaining testing data remain consistent. Then, we solved the inhomogeneous linear least squares problem from the overdetermined system and fit one identical homography for the left-view and right-view. For each unselected correspondence, we utilized the combined distance error to determine whether it is consistent under the fitted homography. The identical homography here can ensure the output panorama will carry more consistent stereo information and avoid undesired disparity issues. For more details about parameters set-up, the corresponding data set $\mathcal{T}$ in Algorithm 2 is defined as CIF set instead of monocular-view matched control points. The inlier percentage p is always set as 90%. The distance threshold σ also need to be specified under different input datasets; we usually set them as 10 pixels in our experiments. The max number of iteration N is determined by the trade-off between the estimation accuracy and computation time.

5.4 Visual Comparison

An example of input control points grouping result under the standard and proposed RANSAC algorithms are depicted in Figure 21. All the input control points into RANSAC are divided into two groups according to whether the control point pairs are consistent under the estimated projection matrix. Those control points pairs that carry the tolerated projection error are regarded as inliers, while the remaining one is called an outlier. For more details, those inliers are marked as green lines and outliers are marked as blue lines in Figure 21. The first row represents the division result of the standard RANSAC algorithm into the left-view and right-view control points independently. Though we observe that several falsely-matched control points are identified as outliers successfully in both left-view and right-view panorama,

the inliers identification between left-view and right-view images is not consistent. In the second row of Figure 21, we can see the inliers and outliers are selected based on the CIF sense according to our proposed adapted RANSAC algorithm. The stereo-constrained RANSAC will output the left-view and right-view homography matrix that can fit the alignment under the same points, edges, and areas in the binocular views. Thus, the stereo-contained RANSAC is expected to provide a more similar camera pose estimation result between the left-view and right-view output panorama.

5.5 Conclusion

In this section, we discussed the adapted version of the RANSAC algorithm for the homography estimation based on the CIF feature set. Given the stereo consistency of the CIF set in the feature matching step, we wish to keep this good property in the later system operations. To achieve that goal, we randomly pick the CIF feature instead of standard 2D features for projection matrix fitting. The best-fitted homography should obtain the largest consistent percentage at the left-view and right-view simultaneously to guarantee the majority of the control points can be correctly aligned in the output panorama. Hence, the finally fitted homography in the left-view and right-view can prevent most of the stereo inconsistency in output stereoscopic panorama.

(a) Inliers and outliers determined by standard RANSAC

(b) Inliers and outliers determined by proposed RANSAC

Figure 21. Inliers and outliers determined by proposed RANSAC. Green lines connect those corresponding control points that are considered as inliers under y the fitted homography. Blues lines connect those corresponding control points that are considered as outliers under the fitted homography.

Algorithm 1 RANSAC-based homography estimation

Input: Corresponding data set $\mathcal{T}$, distance threshold σ, inlier percentage p, max number of iteration N, max number of inlier M

Output: 2D homography $\hat{H}$

Initialize N, M, σ, p

for all for ith ($i = 1 : N$) estimation **do**

  Randomly select 4 corresponding feature pair from $\mathcal{T}$: $\{(f_j, f_j^{'}), j = 1 : 4\}$

  Compute transformation matrix H by normalized DLT from 4 selected corresponding feature pairs

  For each unselected putative correspondence, calculate distance $d_j = d(f_j^{'}, Hf_j) + d(f_j, H^{-1}f_j^{'})$

  Count the number of inlier m which has the distance $d_j < \sigma$

  if $m_i > M$ **then**

    Update best fitted homography $\hat{H} = H$

    Update largest number of consistent inlier $M = m_i$

    Record all the inlier set as S

  end if

end for

return $\hat{H}$

Algorithm 2 Stereo Constrained RANSAC-based homography estimation

Input: Commonly-identified feature set $\mathfrak{T}$, distance threshold σ, inlier percentage p, max number of iteration N, max number of inlier M

Output: 2D homography in left and right view $\hat{H}_L$, $\hat{H}_R$

Initialize N, M, σ, p

for all for ith ($i = 1 : N$) estimation **do**

  Randomly select 4 commonly-identified features from $\mathfrak{T}$: $\{(f_{L,j}, f'_{L,j}, f_{R,j}, f'_{R,j}), j = 1 : 4\}$

  Compute transformation matrix H_L and H_R by normalized DLT independently

  For each unselected putative correspondence in left view, calculate distance $d_{L,j} = d(f'_{L,j}, Hf_{L,j}) + d(f_{L,j}, H^{-1}f'_{L,j})$

  For each unselected putative correspondence in right view, calculate distance $d_{R,j} = d(f'_{L,j}, Hf_{L,j}) + d(f_{L,j}, H^{-1}f'_{L,j})$

  For each unselected putative correspondence, calculate the distance between two fitted homography $d_{H,j} = \|H_L - H_R\|$

  Count the number of inlier m_i when the combined distance $d_j = d_{L,j} + d_{R,j} + \gamma d_{H,j} < \sigma$

  if $m_i > M$ **then**

    Update best estimated homogrphy $\hat{H}_L = H_L$ and $\hat{H}_R = H_R$

    Update largest number of consistent inlier $M = m_i$

    Record consistent inlier sets $\hat{S}_L$ and $\hat{S}_R$

  end if

end for

return $\hat{H}_L$, $\hat{H}_R$

CHAPTER 6

EXPERIMENT AND SIMULATION

6.1 Introduction

In this section, we present the experiments and simulation we conducted to validate the stitching feasibility and robustness of the proposed stitching framework. We will introduce several image acquisition equipment used in our stitching experiments. Then, some basic information about the real-captured datasets and synthetic data will be presented. After that, we define the numerical metric we used in performance evaluation in the later section of this chapter. In the following, we demonstrate the visual comparison between our proposed stitching framework and other stitching solutions in different stitching tasks in the monocular sense, stereoscopic sense, and video sense. Then, the corresponding numerical analysis of those comparisons will be stated in the last part of this chapter.

6.2 Image Acquisition Equipment

In the generation process of the stereoscopic panoramic video task, multiple images from different perspectives are needed to provide input images with overlapping files. In our experiments and simulations, three types of image acquisition equipment are used to provide raw data from real-world scenes or synthetic scenarios.

6.2.1 SENSEICam Simulator

SENSEICam Simulator is a pair of stereo cameras mounted on one spinner. Its mechanical system can freely rotate the dual camera around the optical center without the introduction of parallax. Hence,

this simulator is widely employed in our experiments to set up many different camera array arrangements. However, since it cannot shoot shoot images from different capture positions simultaneously, it is usually employed as the image acquisition equipment for the still panorama rather than the panorama with moving objects or the panoramic video. One possible camera's configuration via SENSEICam Simulator is depicted in Figure 22.

Figure 22. SENSEICam Simulator. (Photograph by Lance Long, *SENSEICam*, September 30, 2018, Electronic Visualization Laboratory, University of Illinois at Chicago).

6.2.2 StarCam

The second camera design is called StarCam, which is one typical interleaved parallel camera. The interleaved design indicates the stereo camera pair are not placed as neighbors along the circle. For each pair of the stereo camera pair, there are four more cameras mounted between the left and right camera perspective. This special camera structure is designed for a more compact camera package and

larger overlapping fields for stitching. Additionally, it will provide a smaller position difference between adjacent cameras for stitching with less parallax. The prototype of StarCam is depicted in Figure 23 and more detailed information can be found in [39].

Figure 23. StarCam prototype. (Photograph by Dominique Meyer, *StarCam*, September 30, 2018, Qualcomm Institute, University of California at San Diego).

6.2.3 Chameleon

The third camera design is called Chameleon, which is one typical radial design. It consists of a monocular camera array mounted in one circle, and there is no standard left or right camera view as in other dual-camera designs. Thus, before we process the image captured by Chameleon design with our standard stitching framework, it requires extra rectification to construct the virtual left and right view

based on captured images. Moreover, the radial design is one of the most popular camera arrangement solutions on the commercial market. One 5-camera prototype of Chameleon is depicted in Figure 24.

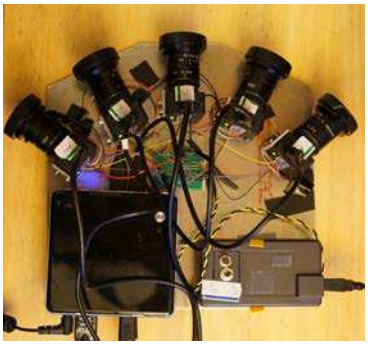


Figure 24. Chameleon prototype. (Photograph by Daniel Sandin, *Chameleon*, September 30, 2018, Electronic Visualization Laboratory, University of Illinois at Chicago).

6.3 Experiment Setup

We conducted experiments with unoptimized MATLAB and *PanoTools* [7] on a PC with an Intel Core i7-3770s 3.1-GHz CPU and 16-GB memory. The *vlfeat* library [40] provided the original SIFT feature detection to produce basic features for CIF construction. During energy map generation, the depth map was generated by basic semiglobal block matching algorithms. The gradient map was generated by the Sobel gradient operator. The saliency map was generated by the GBVS. All three maps were implemented in MATLAB. Regarding feature tracking in the video, we utilized MATLAB's built-in KLT

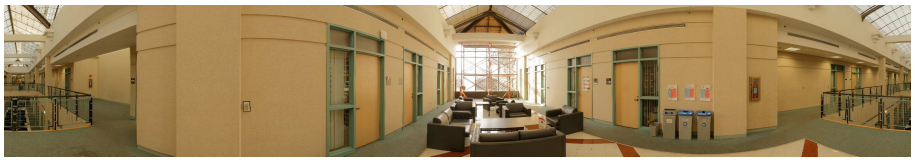
tracker. For the image alignment and blending steps, we estimated the cameras poses and combined individual camera views into the final panoramic video using *PanoTools* framework and *Enblend* [8].

To verify whether our proposed stitching system was capable of stitching under different scenes and camera setups, we tested our proposed stitching strategy on several data-sets. Generally, we divided all the testing data-set into two groups: one for the real-captured scenes from the camera prototype and another group is the synthetic data-sets rendered in simulation software. The distance between stereo camera pair is 75 mm for real data-sets and set as 80 mm in the synthetic data-sets.

In Figure 25, we observed six different datasets for camera-based capture. Among them, Cases (a) and (b) were captured by a pair of stereo cameras mounted on a spinner for indoor and outdoor scenarios. The StarCam prototype [39] is used to capture cases (c) and (d). Cases (e) and (f) were captured by CAVECam [41].

In Figure 26, we observed six different datasets from the simulation software. Among them, Cases (a) and (b) were rendered with 3D software, *Blender*, for outdoor and indoor scenarios. Cases (c) and (d) are rendered by *Unity* for the indoor walking man in a circular path with various radii. Cases (e) and (f) are rendered by *Unity* for the outdoor walking man in a circular path with two different radii.

To clarify the technical details of input data used in the experiment and simulation, we also state the resolution of the input image, number of cameras, and the field of view for each dataset. All the basic information for static scenario dataset-sets and video dataset-sets is recorded in Table IV and Table V, respectively.



(a) Atrimum

(b) Basement

(c) Bearstone

(d) Courtyard

(e) Campus

(f) Rampart

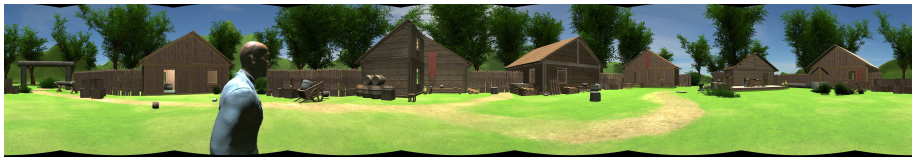
Figure 25. Real data used in the experiments, numbered from 1 to 6.

(a) Pavalion

(b) Classroom

(c) Lving-Room 1

(d) Lving-Room 2



(e) Village 1

(f) Village 2

Figure 26. Synthetic data used in the experiments, numbered from 7 to 14.

TABLE IV

BASIC INFORMATION OF STATIC DATASET

Static Scene	Atrium	Basement	Campus	Rampart	Pavilion	Classroom
Number of Cameras	10x1	10x1	12x3	12x3	10x1	10x1
Number of Rings	1	1	3	3	1	1
Resolution	3448x4729	3448x4729	2992x2992	2992x2992	1500x2000	1500x2000
FOV	47x64	47x64	36x36	36x36	46x69	46x69

6.4 Evaluation Metric Definition

In this part, we define three pixel-based and one feature-based evaluation metrics for a fair comparison with other different panorama stitching solutions. Compared to the feature-based quality metric we used in 2, the pixel-based metric for the single frame can characterize the final stitching quality for all overlapping regions rather than the sparsely distributed features. The performance evaluation based on pixel-wise alignment and disparity distance can better define the quality of stitching output in both the monocular sense and stereo sense. As for the evaluation of panoramic video, we believe the feature-based metric is more suitable rather than the pixel-based one because the long-term tracking for the dense pixels is a difficult task.

TABLE V

BASIC INFORMATION OF VIDEO DATASET

Dynamic Scene	Village	Living room	Courtyard	Bearstone
Video Number	8	8	8	8
Resolution	1920x1920	1920x1920	1920x1080	1920x1080
FOV	60x60	60x60	60x33	60x33
Number of Frames in Video	100	100	500	500
Synthesized or Real	Synthesized	Synthesized	Real-captured	Real-captured

6.4.1 Pixel-based Alignment Accuracy

To quantify the effect of our proposed strategy on various monocular stitching pipelines, we use the projection position distance as the metric to evaluate the image alignment accuracy. After the warping of two neighboring camera views (e.g., L_1 and L_2) into the output canvas, each pixel within overlapping region always corresponds to two source pixels from warped L_1 and L_2 respectively. In the ideal situation, the two source pixels on the output canvas are expected to have zero position displacement. Therefore, given the dense optical flow map $Flow$ between warped L_1 and L_2, we define the monocular-view stitching error as the average position displacement of all the pixels in the overlapping region:

$$S_l = \frac{1}{M * N} \sum_{i=1}^{M} \sum_{j=1}^{N} \sqrt{Flow_x^2(i,j) + Flow_y^2(i,j)}. \tag{6.1}$$

$Flow_x(i,j)$ and $Flow_y(i,j)$ indicates the horizontal and vertical component of the dense optical flow map between warped L_1 and L_2 at position (i,j). The algorithm we used to compute the dense correspondence is called SIFT Flow [42]. Regarding similarity, the right- view monocular stitching error can be defined as a similar pattern. The monocular stitching error stated in later figures is the average error from left-view and right-view panoramas. Usually, the smaller error implies fewer misalignment and discontinuities; in other words, a better stitching quality in the monocular sense.

6.4.2 Pixel-based Vertical Depth Accuracy

To quantify the depth information error, we use the pixel-wise depth accuracy to evaluate whether the output stereoscopic panorama carries the correct disparity value. In the ideal stitching result, the vertical disparity value between each of the two corresponded pixels in binocular view is expected to be zero. Thus, the vertical component of the dense optical flow map implies the vertical depth accuracy of the output stereoscopic panorama. Given the dense optical flow map $Flow$, the vertical depth error is defined as:

$$S_v = \frac{1}{M * N} \sum_{i=1}^{M} \sum_{j=1}^{N} |Flow_y(i,j)|. \tag{6.2}$$

$Flow_y^{GT}(i,j)$ indicates the vertical component of the pixel displacement at position (i,j). A small vertical disparity error between two final stitched panorama canvas indicates good stereo consistency.

6.4.3 Pixel-based Horizontal Depth Accuracy

To define the depth error along the horizontal direction, we first assume that all the input image pairs can provide correct depth information for each pixel in the original spatial coordinate. Then,

we construct a dense correspondence between the input image and output panorama to identify the transformation via optical flow. Due to the diversity of existing stitching algorithms, the displacement between the input image and output panorama images might vary across spatial dimensions. Thus, SIFT Flow [42] is adopted to calculate point correspondence. Given the disparity value from the input image and the dense pixel correspondence, we construct the expected disparity map of the output stereoscopic panorama. Thus, the pixel-based depth difference between the expected disparity map and the measured disparity map can be used to characterize the depth accuracy of output panorama. Given the ground truth disparity map $Disp^{GT}$ and measured disparity Map $Disp^{M}$, the horizontal disparity score is defined as:

$$S_h = \frac{1}{M * N} \sum_{i=1}^{M} \sum_{j=1}^{N} |Disp^{GT}(i,j) - Disp^{M}(i,j)|. \tag{6.3}$$

A small horizontal depth error indicates the delivered depth in the stitched panoramas are close to the depth perceived from the input camera views.

6.4.4 Temporal Consistency

Giving one reasonable evaluation of temporal consistency for different algorithms is rather difficult because there is no one commonly- used benchmark to evaluate the quality of stereo 360 videos. Liu et al. [43] and Guo et al. [35] have presented one metric based on the frequency component of feature trajectories. However, the metric in the frequency domain is not reliable sometimes due to its short length. Nie et al. [36] proposed another stability metric, which is defined as the ratio between the length of the feature trajectory and the length of the virtual straight line from the feature position at the first frame to the last frame. Based on the idea from [35], we proposed one stability score based on CIF.

Formally, a feature-based trajectory is composed of a set of CIFs $\{d_{i,1}^t; i \in \{1,2,3,4\}; t_s \leq t \leq t_e\}$, and t_s, t_e indicate the position of feature at the first and last frame. Then, the stability score of the CIF in this stitched video is defined as:

$$S_t = \frac{\sum_{i=1}^{4} \|d_i^t - d_i^{t-1}\|^2}{\sum_{t=t_s+1}^{t_e} \sum_{i=1}^{4} \|d_i^t - d_i^{t-1}\|^2}. \tag{6.4}$$

If the trajectories of selected control points can perfectly match the virtual straight line between the initial and end position, the stability score will be 1, which means the most stable case as expected. To evaluate the stability of stitched video, we extract all the CIF's trajectories in some selected frames from the stitched video. In our experiments, we don't place too much emphasis on these frames when no moving objects in the overlapping field and wish to analyze the stability between consecutive frames with significant content change. The whole tracked trajectories are cut into small segments with a length of 25 frames. Then, we compute stability scores for all the sections via the previously defined stability score and use the average of them as the temporal metric for video stitching quality comparison.

6.5 Application of CIF in Standard Monocular Stitching Algorithm

Given the definition of the stitching quality metric, we present several different experiments to compare our proposed system with monocular stitching algorithms, stereoscopic panorama stitching algorithms, and stereoscopic panoramic video solutions. First, to demonstrate the improvement brought with our proposed feature structure CIF, in the following, we investigate the application of the proposed strategy in different standard monocular stitching frameworks, such as Simple Homography, AANAP [44], Hugin [7], and SPHP [45].

6.5.1 Simple Homography

Figure 27 demonstrates one example of visual comparison before and after the application of CIF to simple homography. Simple homography refers to one straightforward panorama stitching solution. In a simple homography stitching process, one input image will be selected as the reference plane and all the other input images will be projected into it under the guidance of the fitted projection matrices. In Figure 27, there is no visible misalignment or stitching errors in the amplified ROI, which indicates both the left-view and right-view panorama enjoy good monocular stitching quality. However, the object size consistency in the stereo sense is violated. Given the same resolution of ROI selected from binocular views and left-bottom pixels as the anchor, we can find that the top beam in the second sub-figure has a different height compared to the third sub-figure, which implies the different scaling factors between left-view and right-view panoramas. However, the stitching result under the guidance of CIF, without any visible artifacts and stereo rivalry, can achieve the good monocular stitching quality and stereo consistency simultaneously.

6.5.2 AANAP

Figure 28 demonstrates one example of visual comparison before and after the application of CIF to AANAP [44]. In Figure 28, we can see the stitching output without CIF suffers from severe misalignment around the beam area. The output with a broken line and cropped objects fail to describe the correct geometric structure of the scenario, much less the artifacts-free depth distribution. Thanks to the rejection of the POF set and abandoning of redundant features, the proposed method can fix all these errors under the CIF-based image alignment and produce well-stitched stereoscopic panoramas.

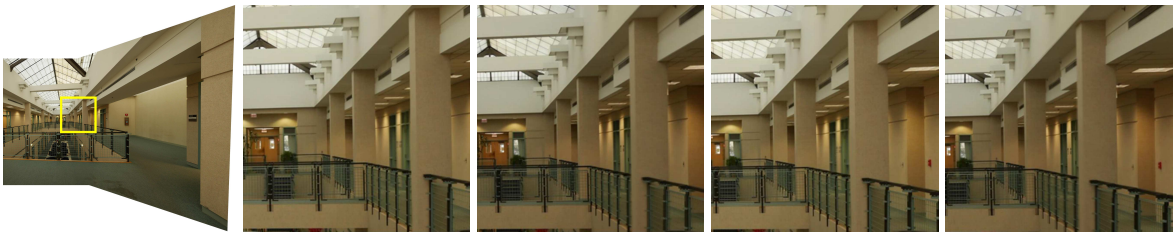


Figure 27. From the left to the right: left view stitched by simple Homography; left view ROI without CIF; right view ROI without CIF; left view ROI with CIF; right view ROI with CIF. Given the same resolution of ROI selected from binocular views and left-bottom pixels as the anchor, we can find that the top beam in the second sub-figure has a different height compared to the third sub-figure, which implies the different scaling factors between left-view and right-view panoramas.

6.5.3 Hugin

Figure 29 demonstrates one example of visual comparison before and after the application of CIF to Hugin [7]. Similar to what we see in the AANAP stitching output, there is also severe distortion detected in the stitching output of Hugin without the operation of CIF. It is noted that the stitching error is corrected in the 4th and 5th sub-figures, which indicates the improvement of proposed CIF to the original monocular stitching algorithms.

6.5.4 SPHP

Figure 30 demonstrates one example of visual comparison before and after the application of CIF to SPHP [45]. It is evident that the major beam is slightly curved in the left-view output without CIF, but turns out to be straight in the corresponding right-view output. Though the viewers may not feel any

Figure 28. From the left to the right: left view stitched by simple AANAP; left view ROI without CIF; right view ROI without CIF; left view ROI with CIF; right view ROI with CIF. We can see obvious misalignment around the beam area in the second and third sub-figure.

Figure 29. From the left to the right: left view stitched by Hugin; left view ROI without CIF; right view ROI without CIF; left view ROI with CIF; right view ROI with CIF. Similar to what we see in the AANAP stitching output, there is also severe distortion detected in the second and third sub-figure.

discomfort when they only observe the left-view or right-view panorama, the delivered depth information around this beam area is corrupted. However, the stitching result under the guidance of CIF can maintain good consistency between left-view and right-view output. The shape and curve of the beams

at two monocular views remain the same, which ensures the correct carried depth information in this area.

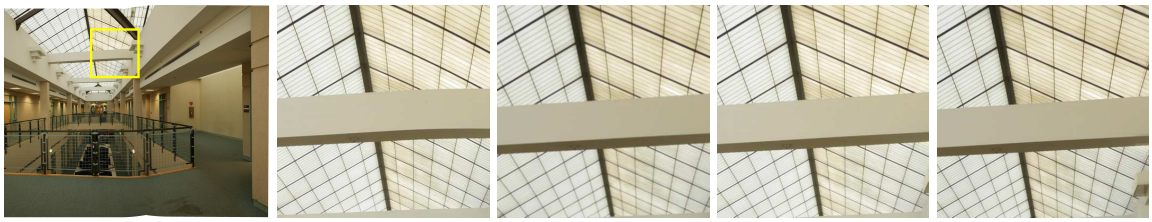


Figure 30. From the left to the right: left view stitched by SPHP; left view ROI without CIF; right view ROI without CIF; left view ROI with CIF; right view ROI with CIF. It is evident that the major beam is slightly curved in the second sub-figure without CIF, but turns out to be straight in the third sub-figure. Therefore, the delivered depth information around this beam area is corrupted.

6.5.5 Quantitative Analysis

Figure 31 summarizes the monocular stitching error of all four algorithms before and after the application strategy. As indicated by these numbers, the monocular image alignment accuracy is slightly improved in all examples. Therefore, given the rejection of the POF set and abandonment of redundant features, the proposed refinement, and the selection strategy to independently detection result will not bring impairment to the original well-stitched panorama in the monocular sense.

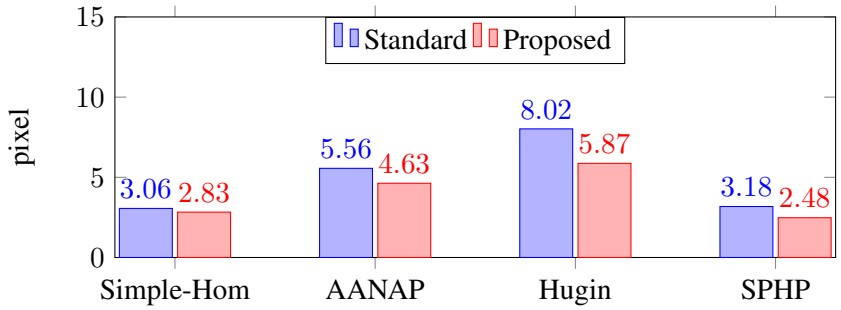


Figure 31. Monocular stitching error comparison

On the other hand, Figure 32 and Figure 33 conclude the average depth error before and after the introduction of CIF. Those monocular stitching algorithms all suffer from significant depth error because they lack the necessary mechanism to deal with depth control. After we use the CIF to replace the original 2D SIFT feature in these stitching frameworks, the pixel-based depth error is largely mitigated. The better vertical and depth accuracy after the introduction of our proposed feature structure demonstrates the improvement in the depth control as we expect intuitively.

6.6 Comparison with Other Featured-based Stereoscopic Panorama Stitching Solutions

To compare with other stereo panorama stitching algorithms, we consider the standard feature-based panorama generation framework as the baseline and utilize the stitching result from Casual Stereoscopic Panorama Stitching [4] and AutoPano Pro [17] as another two competitors.

6.6.1 Hugin

Figure 34 shows the left-view, right-view, and dense disparity map of the stereoscopic panoramas stitched by the Hugin. In the second sub-figure, the right-view output panorama, we can see the area

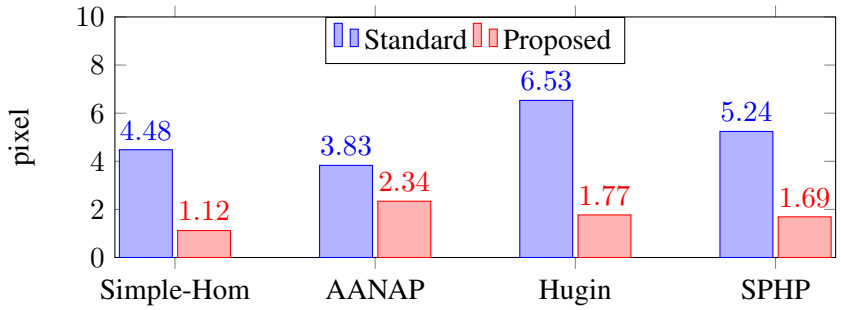


Figure 32. Horizontal depth error comparison

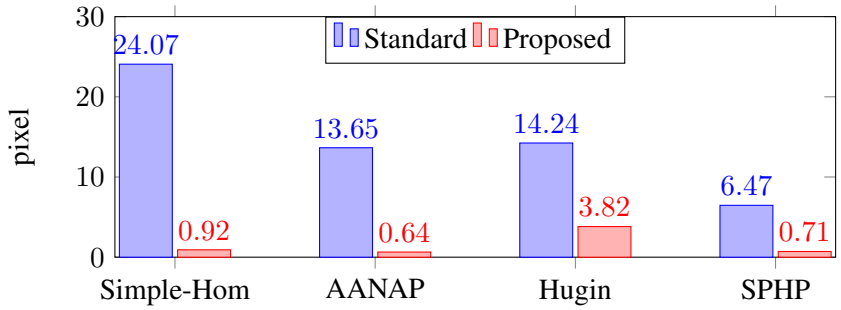


Figure 33. Vertical depth error comparison

marked with the yellow rectangle has noticeable shape distortion between the pillar bottom and the ground. Moreover, the estimated dense map can only provide incomplete depth information at a very limited area in the output panorama. Most of the background is marked with black, which indicates no corresponding matches between left-view and right-view panoramas.

6.6.2 AutoPano Pro

Figure 35 shows the left-view, right-view, and dense disparity map of the stereoscopic panoramas stitched by the AutoPano Pro. In general, the output panoramas at the left-view and right-view enjoy good monocular stitching quality except one tiny discontinuity at the bottom of the cart. The majority area of dense disparity map is smooth and accurate. The only erroneousness area is around the bottom of the cart, where the right-view stitching error corrupts the depth information.

6.6.3 CSPS

Figure 36 shows the left-view, right-view, and dense disparity map of the stereoscopic panoramas stitched by the CSPS. Due to the parallax-tolerant stitching algorithm [46], we can barely detect any geometric errors or misalignments in the monocular-view output panorama via CSPS. For the corresponding disparity map, we can also identify the correct horizontal disparity value of most pixels in this panorama and understand the overall depth distribution in this scenario. However, there are also many local regions without valid depth information. One of these problems can be found at the left-bottom corner, which slightly impairs the smoothness of the whole panorama depth map.

6.6.4 Proposed System

Figure 37 shows the left-view, right-view, and dense disparity map of the stereoscopic panoramas stitched by the proposed stitching system. From the monocular stitching perspective, there are no visible stitching errors or misalignment in left-view or right-view panorama. From the stereoscopic stitching perspective, the estimated disparity map can deliver complete and smooth depth distribution to the viewers without any invalid patch or ambiguity. In summary, compared to the other three stitching

solutions, the output panoramas from our proposed system can achieve good monocular quality and excellent depth control simultaneously.

6.6.5 Stereoscopic 360 Panorama Comparison

Figure 38 shows another example of 360 panorama generation. The first row is the stereoscopic panorama from Hugin with CIF in red-cyan anaglyph. The following rows are dense disparity maps from four different stereoscopic panorama stitching solutions we discussed above. For the original Hugin and AutoPano Pro, although we manually shifted the left-view and right-view panorama to ensure they can deliver some useful depth information at some areas of the output, significant vertical disparities and incorrect horizontal disparities still exist. In their disparity map, we can only perceive quite limited useful depth information. For the stitching result of CSPS, there is no noticeable parallax in the vertical direction. However, for the horizontal disparity distribution over the whole 360 panoramas, the output of CSPS suffers from serious depth contraindication. Those objects on the right side usually have positive horizontal disparity, which can be identified due to their reasonable depth. However, the objects that should also have a positive disparity value in the middle and left side are found to carry negative horizontal disparity. No matter how we align the left-view and right-view panorama, their perceived depth information in this scenario is always confused and makes it difficult for viewers to understand the accurate distance of the surrounding objects. In this case, only the output panorama of our proposed method can provide viewers a reasonable depth distribution of all objects without any ambiguity or misunderstanding.

6.6.6 Quantitative Analysis

In Figure 39, we list the depth error of four stereoscopic panorama generation solutions. Hugin shows the worst performance in the depth control as we expected because it is not designed for stereoscopic panorama stitching tasks. The other three stereo stitching solutions obtain much better depth control results, especially in the vertical direction. Our proposed strategy achieves the minimal horizontal depth error and the second smallest vertical depth error. Though the CSPS can perform slightly better in the vertical depth control, the difference between the CSPS and our proposed system is rather small. Finally, it is noted that the latter two solutions outperform Hugin and AutoPano Pro in the depth accuracy control.

6.7 Comparison with Other Featured-based Stereoscopic Panoramic Video Stitching Solution

In addition to academic research, many commercial companies are producing the stereoscopic 360-degree video. Some products, such as the Ricoh Theta or the Gear360, capture monoscopic video via a monocular radial camera arrangement. That camera design allows for compact cameras but does not meet our target since the stereo cue is missing. Some software is producing stereoscopic 360-degree cameras such as Google Jump [5] and Facebook Surround 360 [47], however, it is challenging to evaluate these systems because some of them did not employ the feature-based stitching framework and others use proprietary stitching methods. For a fairer performance evaluation, in this section, we mainly compare our method with Autopano Pro [48] stitching result, which is one feature-based 360 video generation software. We consider it as the standard feature-based panorama generation framework without any optimization to the original detected SIFT features.

6.7.1 Monocular Panoramic Video

Figure 40 and Figure 41 depict the ROIs from stitched monocular panorama in the indoor and outdoor scenarios respectively. The first row is the stitching output form AutoPano Pro and the second row displays the stitching result from the proposed stitching system. In Figure 40, we can find shape distortion of the human head in sub-figure(b). Compared to the correctly stitched human head at the previous and consecutive frames, discontinuities of the object shape will lead to viewing discomfort and will ruin the depth perception of the human head. Similar errors can also be identified in sub-figure(b) in Figure 41. In our proposed stitching result, those unpleasant artifacts are fixed under more accurate image alignment, which ensures the shape consistency of the moving human head between neighboring frames.

6.7.2 Stereoscopic Panoramic Video

Figure 42 and Figure 43 demonstrate the visual comparison of stereoscopic panorama for video datasets in red-cyan anaglyph. In those stitching results from Autopano Pro, we can always find some objects with incorrect depth. For example, in the top row of sub-figure (c) in Figure 42, we can barely detect any horizontal disparity of the stake between red and cyan views. Thus, the zero horizontal disparity implies the stake located at a distance from the shooting position. However, the visual understanding of the whole scene tells us that the stake should be located between the close walking man and the house in the background. The contradicted depth distribution in this scenario makes it difficult for the viewer to know the true position of those objects in the 3D coordinate. In the bottom row of sub-figure (c) in Figure 42, the horizontal disparity of the walking man, the stake and house are all adjusted into a

reasonable range so that their depth can be correctly perceived now. More similar situations can also be observed under other scenarios in Figure 42 and Figure 43.

6.7.3 Quantitative Analysis

Table VI summarizes the video stability score of AutoPano Pro and our proposed stitching system. As indicated by these scores, the stability has been improved in our proposed stitching system on all examples. Our proposed stitching system can achieve a panoramic video with a better viewing experience.

6.8 Conclusion

In this chapter, we explain the details of our experiments and simulations for the system performance evaluation. In the image acquisition part, we present three different camera array designs and corresponding synthetic/camera-based data-sets. Moreover, to conduct one fair performance evaluation, we divide the all of the experiments into three parts: application of CIF in other monocular stitching algorithms, comparison with other stereoscopic panorama stitching algorithms, and comparison with other stereoscopic panoramic video stitching solutions. These three different levels of comparisons focus on the monocular alignment accuracy, stereoscopic depth control, and video stability respectively. To quantify these stitching quality improvements, we also proposed one pixel-based metric to characterize the stitching result from the geometric correctness, depth distance error, and temporal consistency.

Figure 34. Left-view, right-view panorama and dense disparity map via Hugin. We can see the area marked with the yellow rectangle has noticeable shape distortion between the pillar bottom and the ground. The estimated dense map can only provide incomplete depth information at a very limited area in the output panorama. Most of the background is marked with black, which indicates no corresponding matches between left-view and right-view panoramas.

Figure 35. Left-view, right-view panorama and dense disparity map via AutoPano Pro. The output panoramas at the left-view and right-view enjoy good monocular stitching quality except one tiny discontinuity at the bottom of the cart. The majority area of dense disparity map is smooth and accurate. The only erroneousness area is around the bottom of the cart, where the right-view stitching error corrupts the depth information.

Figure 36. Left-view, right-view panorama and dense disparity map via CSPS. There are also many local regions without valid depth information. One of these problems can be found at the left-bottom corner, which slightly impairs the smoothness of the whole panorama depth map.

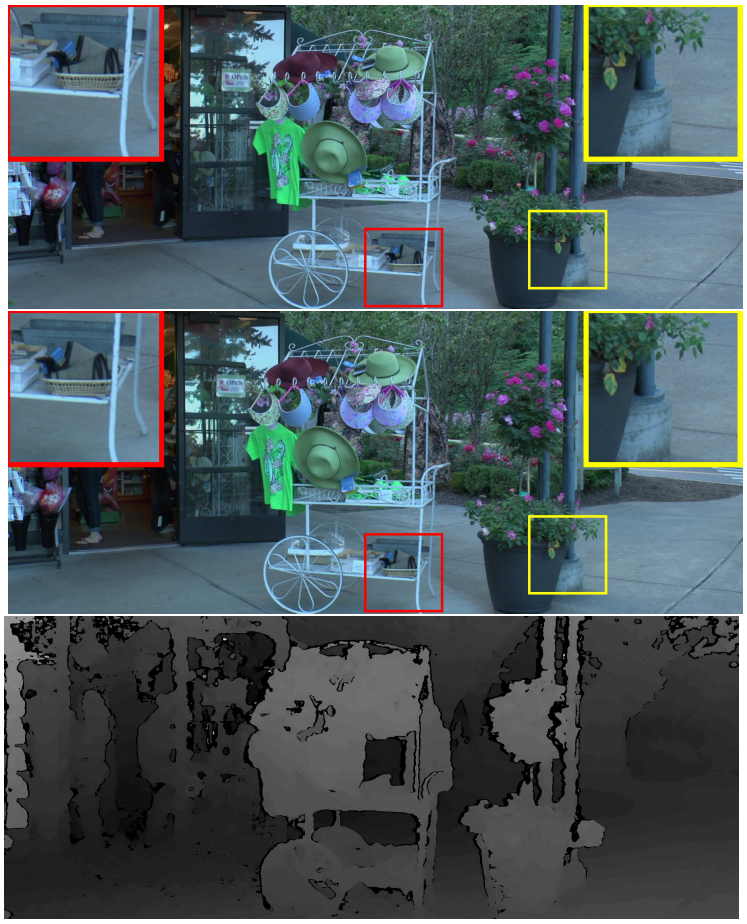


Figure 37. Left-view, right-view panorama and dense disparity map via proposed system. There are no visible stitching errors or misalignment in left-view or right-view panorama. From the stereoscopic stitching perspective, the estimated disparity map can deliver complete and smooth depth distribution to the viewers without any invalid patch or ambiguity.

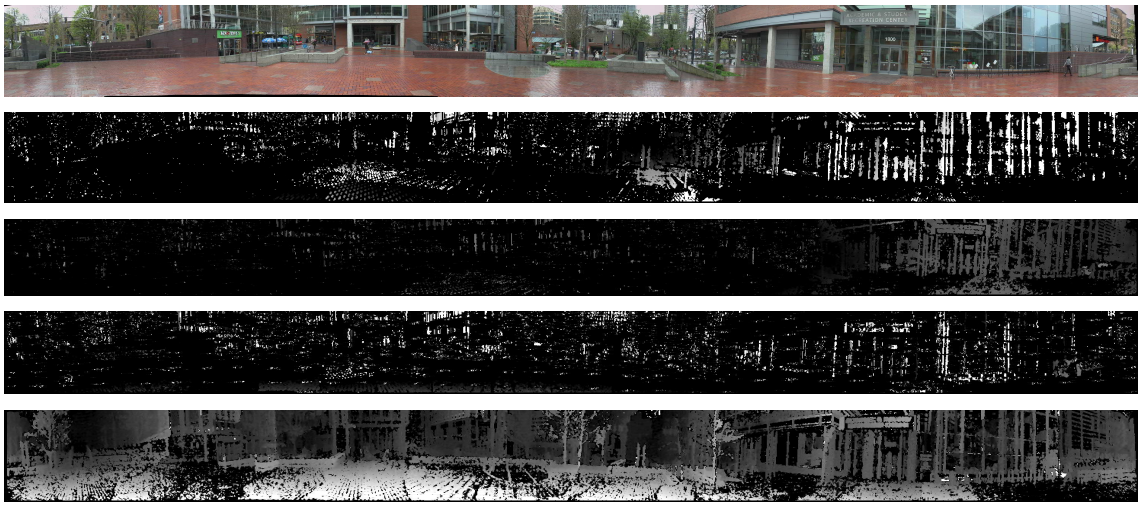


Figure 38. Comparison between different stereoscopic panorama solutions under 360 case. The first row shows the stitched 360 panorama via proposed method in red-cyan anaglyph version. The lower four rows display the estimated dense depth map from original Hugin, CSPS, Autopano Pro and our proposed method respectively.

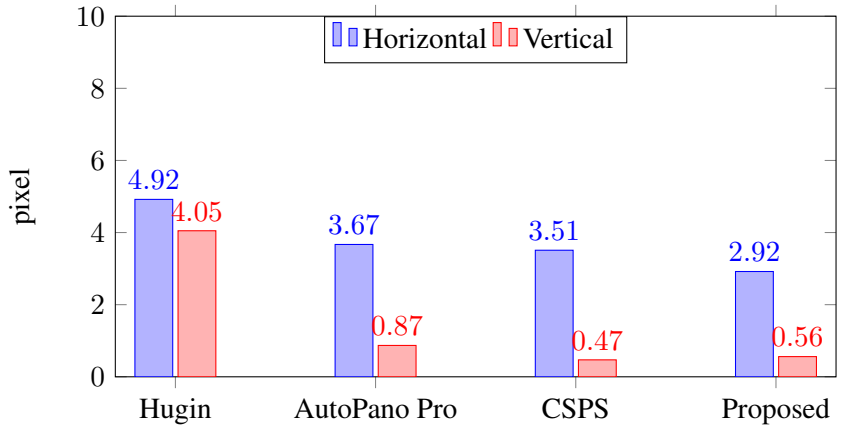
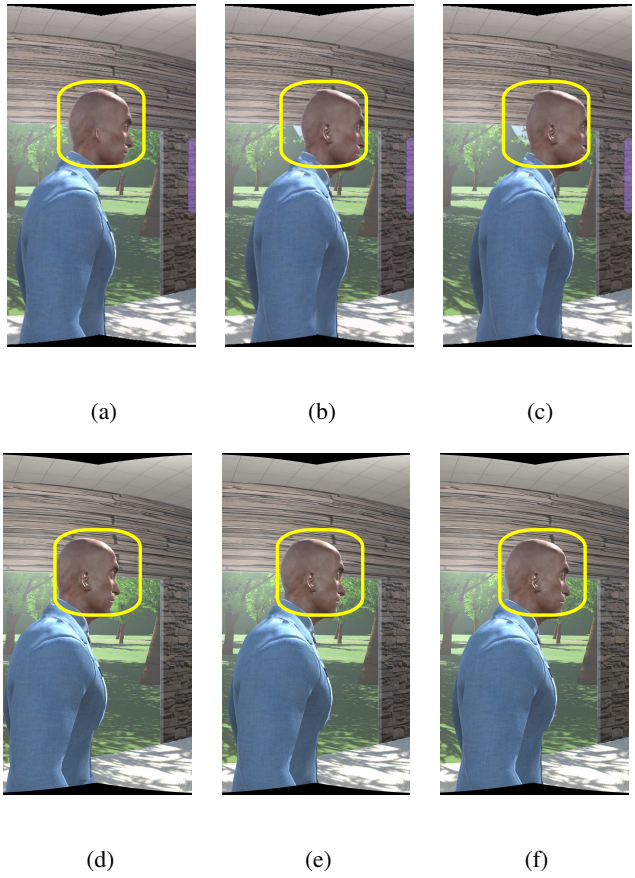


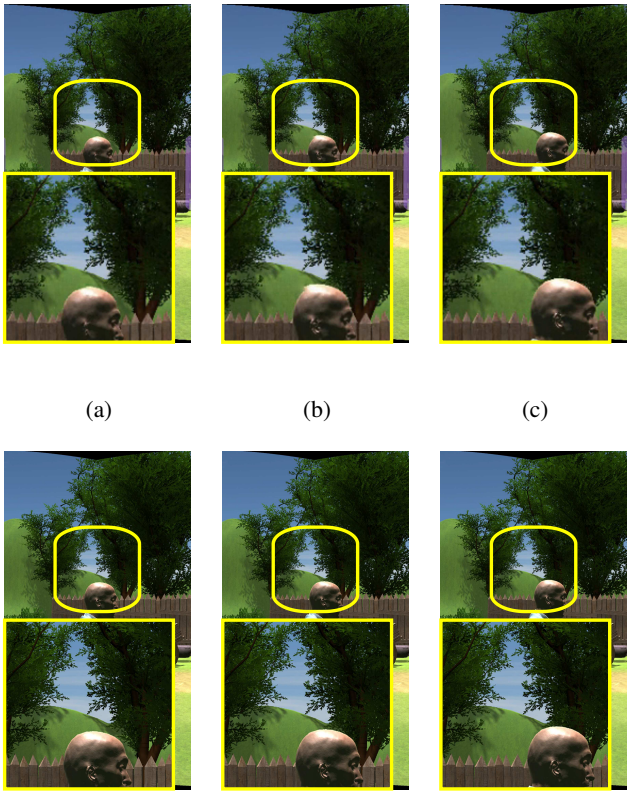
Figure 39. Depth error comparison between stereoscopic stitching solutions



(a) (b) (c)

(d) (e) (f)

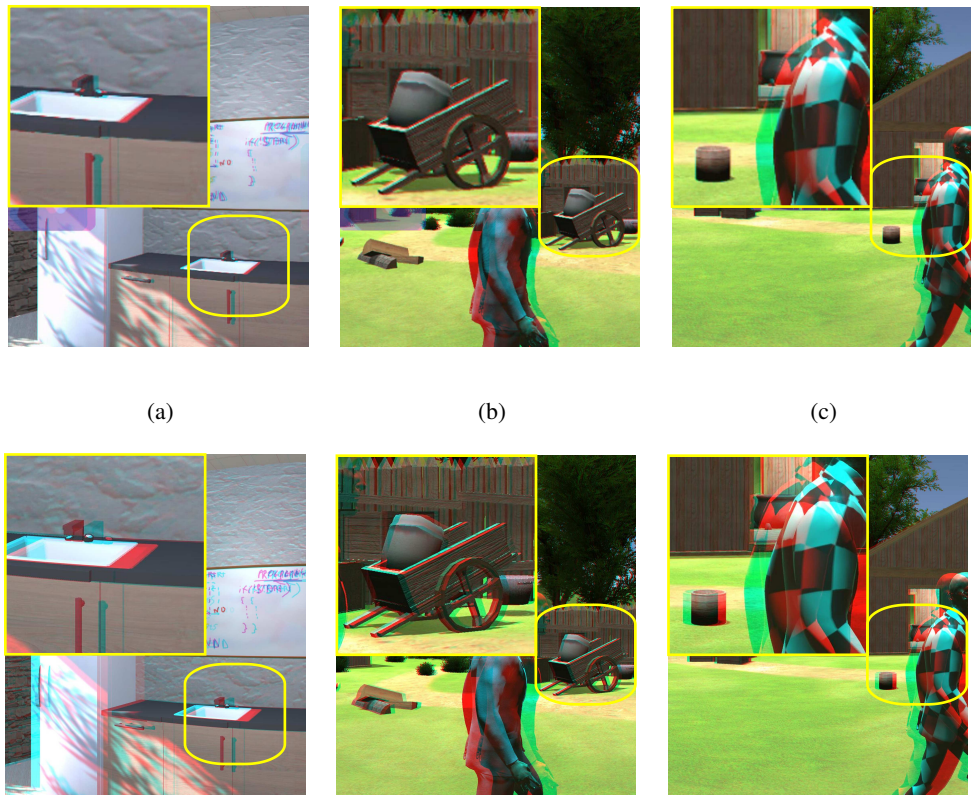
Figure 40. Visual comparison of monocular panorama for synthetic indoor dataset. The first and second row shows three consecutive stitching results of AutoPano Pro and our proposed method respectively.



(a) (b) (c)

(d) (e) (f)

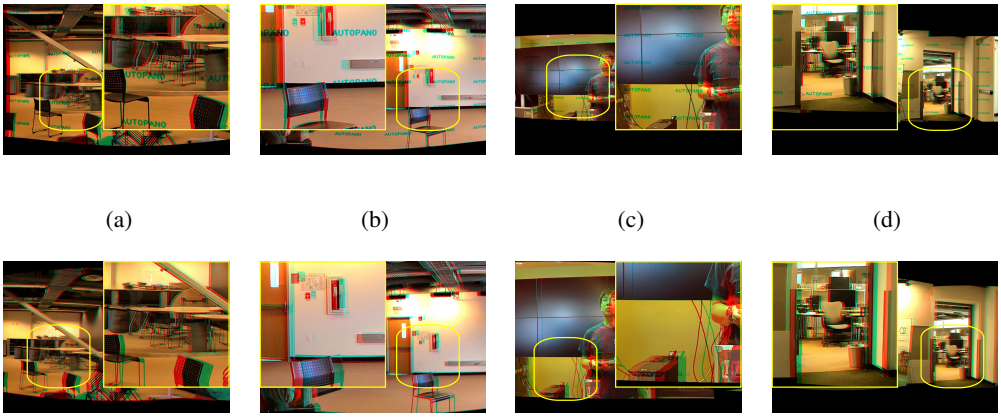
Figure 41. Visual comparison of monocular panorama for synthetic outdoor dataset. The first and second row shows three consecutive stitching results of AutoPano Pro and our proposed method respectively.



(a) (b) (c)

(d) (e) (f)

Figure 42. Visual comparison of stereoscopic panorama for synthetic dynamic dataset in red cyan anaglyph. The first and second row shows stitching results of Autopano and our proposed method respectively.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 43. Visual comparison of stereoscopic panorama for real dynamic dataset in red cyan anaglyph. The first and second row shows stitching results of AutoPano Pro and our proposed method respectively.

TABLE VI

STABILITY COMPARISON OF DYNAMIC DATASET

	Autopano	Proposed
Synthetic Outdoor 1(2m)	0.7157	0.7677
Synthetic Outdoor 2(5m)	0.7186	0.7685
Synthetic Outdoor 3(Line)	0.7549	0.8246
Synthetic Outdoor 4(Cross)	0.7163	0.7523
Synthetic Indoor 1(1m)	0.6976	0.7429
Synthetic Indoor 2(2m)	0.6169	0.6839
Real Indoor 1(Real1)	0.7099	0.7798
Real Indoor 2(Real2)	0.7131	0.7201

CHAPTER 7

POST-STITCHING FLOW MAP GUIDED PANORAMA CORRECTION

7.1 Introduction

Most of the existing stereoscopic panorama generation algorithms try to achieve better stitching output via improving the corresponding accuracy of control points or fitting one optimized blending mask to avoid the possible discontinuities. However, these efforts are all made before the generation of the final output, which means they lack the mechanism to adjust or update the final output according to the feedback from the viewers. The refinement and optimization during image alignment and image blending step can generally produce acceptable output, but they cannot guarantee the stitched panorama is free of any stitching errors. Thus, in this chapter, we propose one post-stitching correction to remove those visible artifacts in the original output panorama and bring a better viewing experience.

7.2 Objective

One good stereoscopic panorama is always expected to have artifact-free monocular stitching results in both left-view and right-view output. Additionally, the stereo consistency between binocular view panorama and the reasonable depth distribution are our purposes in the stereoscopic panorama stitching task. Since there is no ground truth for the final stitching output, we intend to utilize the information from the input image to adjust or correct the output panorama for satisfactory viewing. For instance, we can see the stitched panorama with discontinuities in Figure 44 and one artifact-free input image in Figure 45.

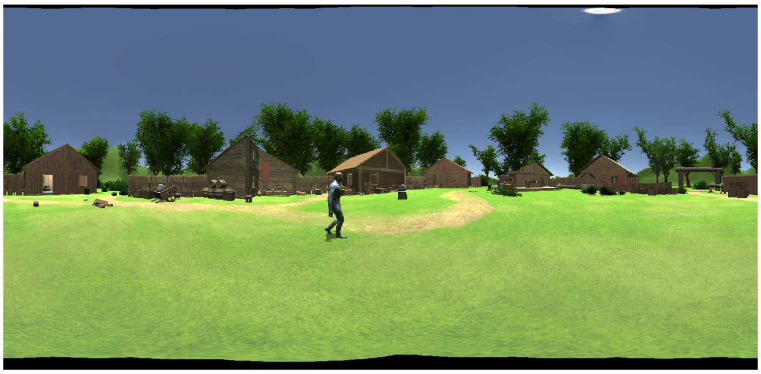


Figure 44. Left-view output panorama from 48 images. One discontinuity is detected around the overlapping region near the human's leg.

7.3 Proposed Monocular Panorama Correction

The proposed stereoscopic panorama correction mainly consists of four steps: ROI Registration, Flow-map Estimation, Flow-map Correction, and ROI Reconstruction. In the following, we will discuss the details of each step.

7.3.1 ROI Registration

To correct the visible artifacts in the output panorama, we need to manually identify the ROI that need correction. Since the stitching operation only occurs at the area that can be viewed by neighboring cameras, we can leave those regions from unique input image and focus on the details in each overlapping window. Given the identified ROI that needs correction, we employ it as the template and operate the normalized cross-correlation (NCC) method to search the best-matched region in the input image.

Figure 45. Left-view major input image selected from four input images,

In this correspondence problem, the appearance difference function needs to be defined for search the patch with the highest similarity. One of the popular searching options, called correlation, is defined as:

$$C_{f,g} = \sum_{[i,j] \in R} f(i,j) * g(i,j), \tag{7.1}$$

where f is the template, and g is one of the potentially matched windows.

However, the straightforward correlation may cause one problem in that brighter regions always generate higher similarity scores, regardless of details in the template. Thus, the solution to this problem is to subtract the mean value of the template before we operate the correlation computation. Thus, the correlation result will not be biased by the significant difference between the average intensity values of the two windows. Another problem of the straightforward correlation is its failure in intensity change control between the reference and template. Those two images we intend to match may have different intensity response characteristics due to the natural illumination change, capture position movement

or the autogain control set from the camera. The corresponding solution is to normalize the pixels by subtracting the mean of patch intensities and dividing by the standard deviation. Thus, one improved correlation function, the NCC score between reference f and potentially matched template g can be defined as:

$$NCC_{f,g} = \sum_{[i,j]\in R} \hat{f}(i,j) * \hat{g}(i,j), \tag{7.2}$$

$$\hat{f} = \frac{f - \bar{f}}{\sqrt{\sum (f - \bar{f})^2}}, \tag{7.3}$$

$$\hat{g} = \frac{g - \bar{g}}{\sqrt{\sum (g - \bar{g})^2}}. \tag{7.4}$$

After the search for all the possible windows with the NCC function in the input image, we will identify the patch with the largest score as the best-matched ROI.

In Figure 46, we can see the target ROI from output panorama and its matched ROI from the input image. It is noted that the two registered patches are similar. However, we cannot directly use the matched ROI to replace the target ROI due to the many tiny differences and mismatches. To reconstruct one satisfied ROI, we need more specific corresponding information rather than the patch-wise registration.

7.3.2 Flow-map Estimation

In this section, we establish one dense pixel-based correspondence between the panorama ROI and image ROI we registered in the previous step. Due to the diversity of existing stitching algorithms, the displacement between the input image and output panorama might vary across spatial dimensions.

Figure 46. Target ROI from output panorama and matched ROI from input image.

Thus, SIFT Flow is adopted to calculate point correspondence. The SIFT Flow is formulated almost the same as the standard optical flow framework. The difference between them is that the matching function in the SIFT Flow framework is based on the SIFT descriptors instead of RGB intensity values in the optical flow framework.

For each pixel $p = (x, y)$ in target ROI, the SIFT Flow framework tries to find the closest pixel in reference ROI and return its position displacement $w(p) = (u(p), v(p))$. For two images I_1 and I_2 we intend to match, the SIFT Flow framework utilizes the standard SIFT detection algorithm to generate the dense SIFT representation for every pixel, which is called SIFT image. Let s_1 and s_2 be two SIFT images and set ϵ a four-neighbor system that contains all the spatial neighborhoods, the energy function in the SIFT Flow framework can be defined as:

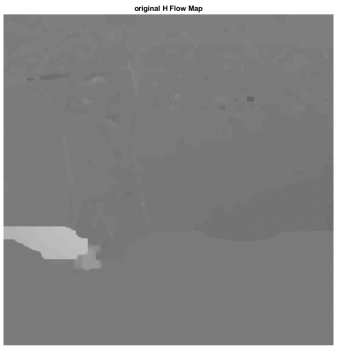
$$E(w) = \sum_{p} min(\|s_1(p) - s_2(p + w(p))\|_1, t) + \sum_{p} \eta(u(p) + v(p)) + \sum_{(p,q)\in\epsilon} min(\alpha(u(p) - u(q)), d) + min(\alpha(v(p) - v(q)), d). \tag{7.5}$$

In Equation 7.5, the formulated energy function contains three terms: data term, small displacement and smoothness term. The data term describes the gradient difference between two SIFT descriptors. The small displacement term then constrains the estimated position displacement vector to be as small as possible. The third term, which is the smoothness penalty, forces those adjacent pixels to have similar position displacement values.

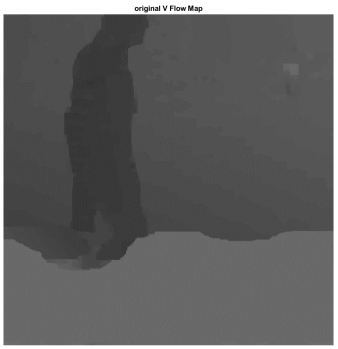
One example of horizontal and vertical flow maps between panorama ROI and image ROI are depicted in Figure 47. We can observe that there are some un-smooth pixels in the area we labeled as the target ROI in the last step. Those pixels with the sudden change of intensity value because they got different position displacement compared to their neighbors. We believe these inconsistent position displacements are the primary reason why we can see these discontinuities in the output panorama. To achieve the consistency of the corresponding from reference coordinate into the target coordinate, we need to adjust the pixel intensity values in these two estimated maps. The object of this intensity value adjustment is to make sure all the pixels of ROI at the reference coordinate to hold close position displacement.

(a) Target ROI

(b) Reference ROI



(c) Original horizontal map



(d) Original vertical map

Figure 47. Horizontal and vertical flow map before correction.

7.3.3 Flow-map Correction

In this section, we explain the details to adjust the position displacement value in the original estimated flow map. In Figure 48, we first remove all the pixels in the labeled ROI and interpolate them based on the pixels on outer boundaries. To make the nearby pixels can be warped with close displace-

Figure 48. Horizontal and vertical flow map with holes.

ment values under the corrected flow map, we assume that pixel in this image should satisfy Laplace's equation:

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0. \tag{7.6}$$

Those pixels disturbed around the ROI can provide us the boundary condition for this partial differential equation. In the simplified discrete version of Laplace's equation, the value of each grid element should be the average of its four connected neighbors:

$$I(x+1, y) + I(x-1, y) + I(x, y+1) + I(x, y-1) = 4 * I(x, y). \tag{7.7}$$

Based on the assumption stated above, we can set up a linear system of equations: $Ax = b$ to describe the pixel-wise relationship in the expected flow map. The left-hand side is one sparse coefficient

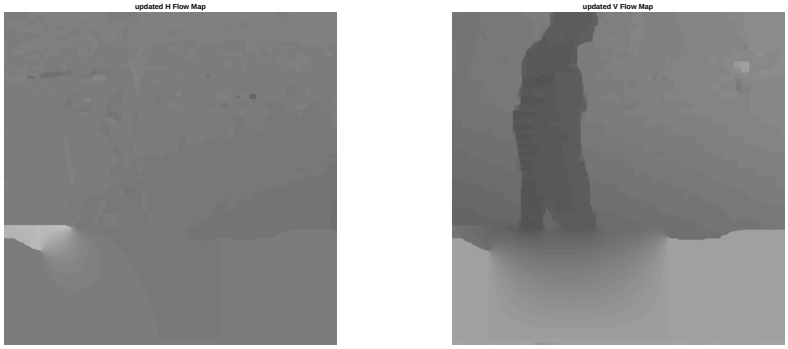


Figure 49. Horizontal and vertical flow map after correction.

matrix that is formulated from the Laplace assumption. The right-hand vector, b, contains either the original pixel values (for pixels outside the ROI) or 0 (for pixels inside the ROI). Thus, the expected pixel value vector x can be easily obtained and reshaped to the size of the original flow map.

In Figure 49, we can observe the updated horizontal and vertical flow map after we fill the empty regions. Compared to the original flow maps, these updated maps achieve much better smoothness inside the marked ROI, which indicates better warping consistency for neighboring pixels.

7.3.4 ROI Reconstruction

Since the updated flow map provide us one smoother warping function to reconstruct the target area, our next step is to produce one new target area without any visible stitching errors. For each pixel $I(x, y)$ in the target ROI, we can find its position displacement along x and y coordinate from the updated flow maps, respectively:

$$Disp_w(x, y) = flow_w(i, j), \tag{7.8}$$

$$Disp_h(x, y) = flow_h(i, j). \tag{7.9}$$

Then we can utilize the RGB intensity value of pixel at $(x + Disp_w(x, y), y + Disp_h(x, y))$ in the reference ROI to replace the pixel at (x, y) in the target ROI coordinate. The warping technique we use here is called reverse mapping. Figure 50 displays the reconstruction process from the original target ROI to the corrected target ROI. Sub-figure(a) is the original target ROI with the discontinuities around human's lower leg. Since most of the pixels in target ROI can correspond to one valid pixel in the reference ROI, sub-figure(b) displays one error-free reconstruction result in most of the image area. However, there are some pixels with an invalid corresponding that fails to obtain updated intensity values and are left as black holes in sub-figure(b). For these outliers in the reverse mapping, we directly keep their original intensity values from the target coordinate. The final reconstruction result is shown at sub-figure(c) of Figure 50.

7.4 Performance Evaluation

7.4.1 Experiment Setup

To validate the feasibility and robustness of our proposed flow map guided panorama correction technique, we test parts of the stitched panoramas and panoramic videos in chapter 6, which include both synthetic and real data-sets. For each original output with visible stitching errors, we manually label the ROI with 500 by 500 rectangles and operate the proposed correction to them. The visual example and numerical analysis of our proposed correction technique are discussed in the following.

(a) Original target ROI

(b) Reconstructed target ROI with Holes

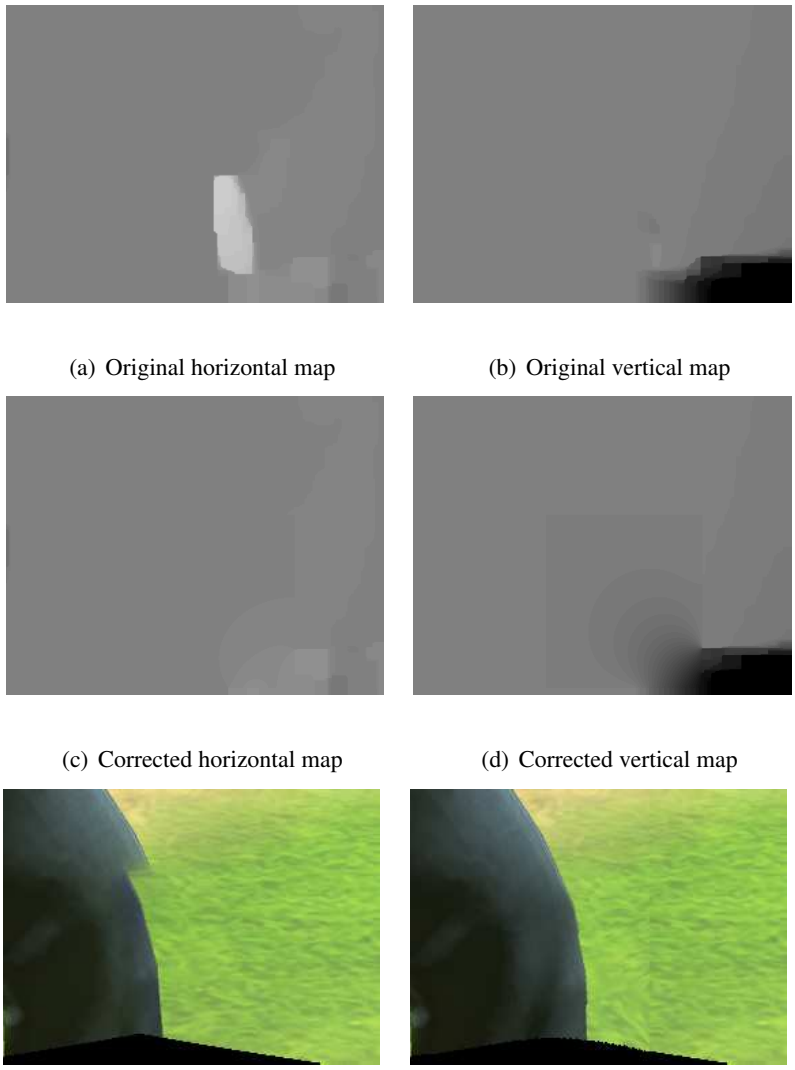
(c) Reconstructed target ROI after hole filling

Figure 50. Reconstructed ROI.

It is noted that our proposed correction can mitigate most of the visible stitching errors in the original output panorama.

7.4.2 Monocular Panorama Correction

In Figure 51, we can see an example of one monocular-view panorama correction.



(a) Original horizontal map

(b) Original vertical map

(c) Corrected horizontal map

(d) Corrected vertical map

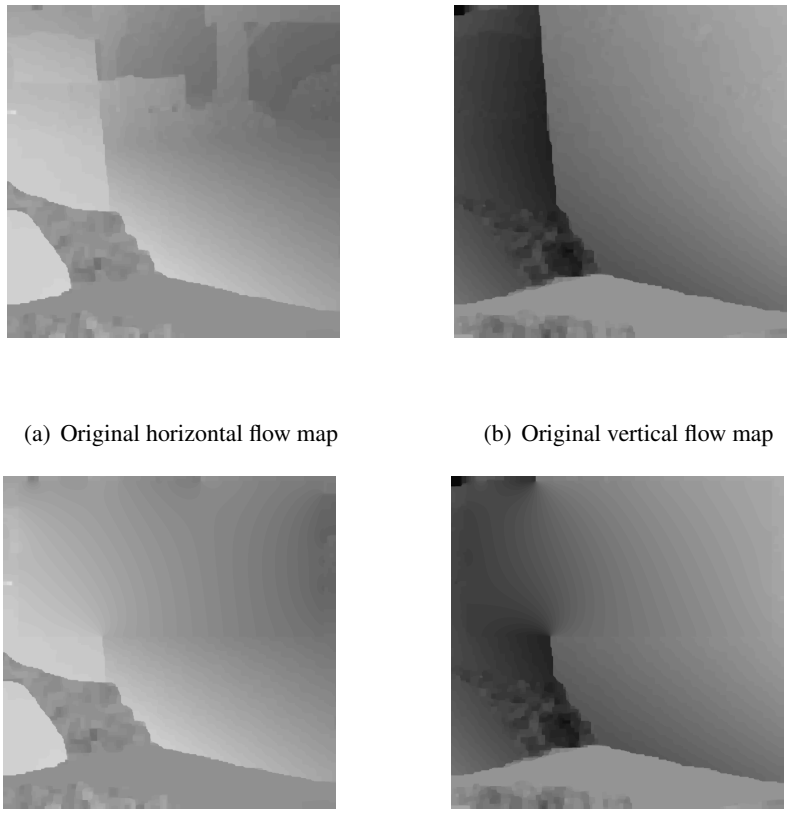
(e) Original ROI

(f) Reconstructed ROI

Figure 51. Monocular correction example.

7.4.3 Application to Stereoscopic Panorama

In this section, we extend the previous monocular correction into the stereoscopic panorama correction. To maintain the stereo consistency between left-view and right-view panorama after the correction, we choose to operate standard operation to one of the binocular views and utilize the updated flow map to guide the ROI reconstruction in another view. For instance, we label the target ROI in the left-view output panorama, find its corresponding ROI in the left-view input image and correct left-view panorama with the updated flow map. For the right-view panorama correction, we also utilize SIFT Flow to find the corresponding target ROI in the right-view output panorama, and reference ROI in the right-view input image. Since the well-rectified input image can ensure the good correspondence between the left-view and the right-view input image, we can follow the routine of left-view panorama correction and reconstruct the ROI in the right-view panorama. Figure 52 and Figure 53 display the flow-map update process in left-view and right-view ROI, respectively. The finally corrected ROIs are depicted in Figure 54.



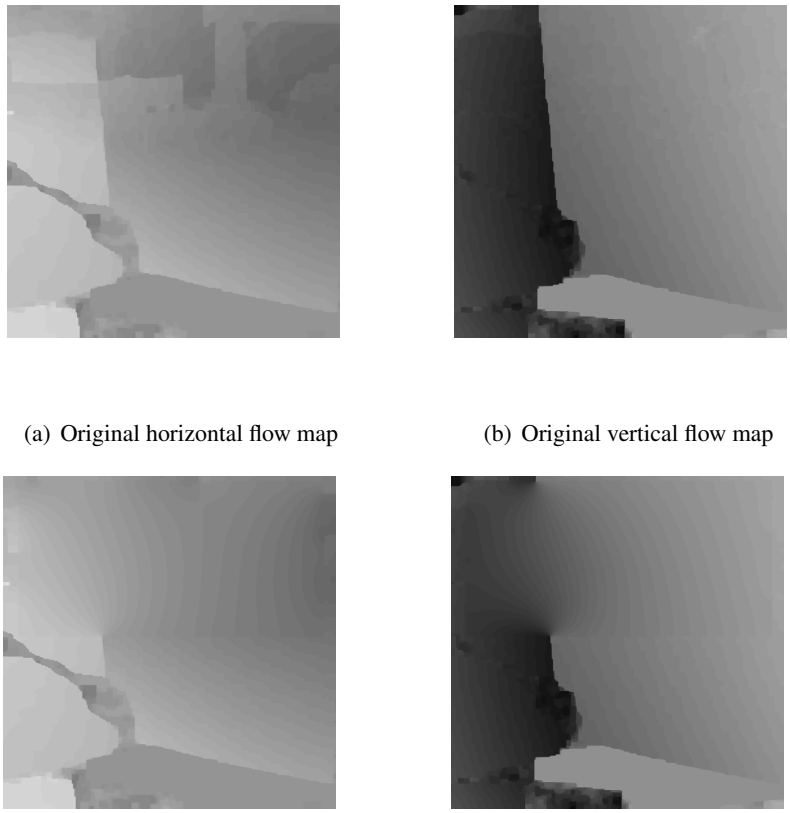
(a) Original horizontal flow map

(b) Original vertical flow map

(c) Corrected horizontal flow map

(d) Corrected vertical flow map

Figure 52. Left-view flow map before and after correction.



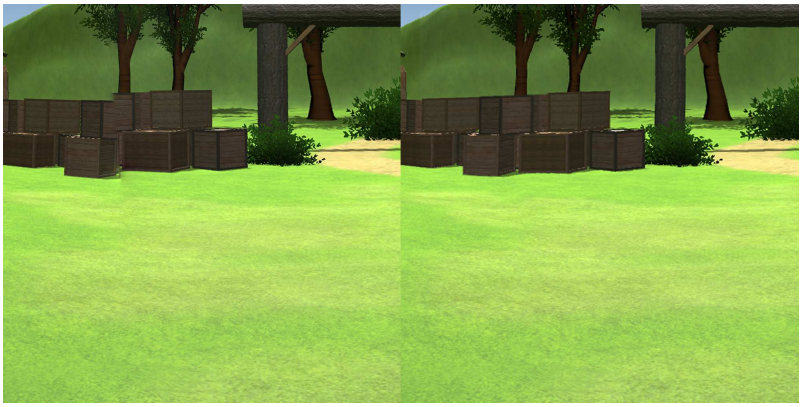
(a) Original horizontal flow map

(b) Original vertical flow map

(c) Corrected horizontal flow map

(d) Corrected vertical flow map

Figure 53. Right-view flow map before and after correction.



(a) Left-view ROI before and after correction

(b) Right-view ROI before and after correction

Figure 54. Left-view and right-view ROI before and after correction.

7.4.4 Application to Panorama Video

In this section, we extend the previous panorama correction into the video case. Unlike the situation of stereoscopic panorama correction, the stitching errors and artifacts always appear at some places in the video sequence because of the fixed blending mask. Thus, there is no need to track the trajectory of initialized ROI in later frames. Once we label the ROI with stitching errors in one frame, we only need to focus the same place in later frames and operate standard monocular panorama correction to the corresponding ROI in the video when it's necessary. In Figure 55, we can see the discontinuities around the leg region in all three frames. Given the same ROI in three consecutive frames, we update the flow maps and interpolate the erroneous areas to obtain new maps for more consistent position displacement. The correction to the left-view and right-view flow maps are show in the Figure 56 and Figure 57 respectively. Finally, we can see the target ROIs and reconstructed ROIs in Figure 58. There are no visible stitching errors detected in the ROI after correction.

(a) Panorama at 1st frame (b) Panorama at 2nd frame (c) Panorama at 3rd frame

Figure 55. Cropped panorama in three consecutive frames.

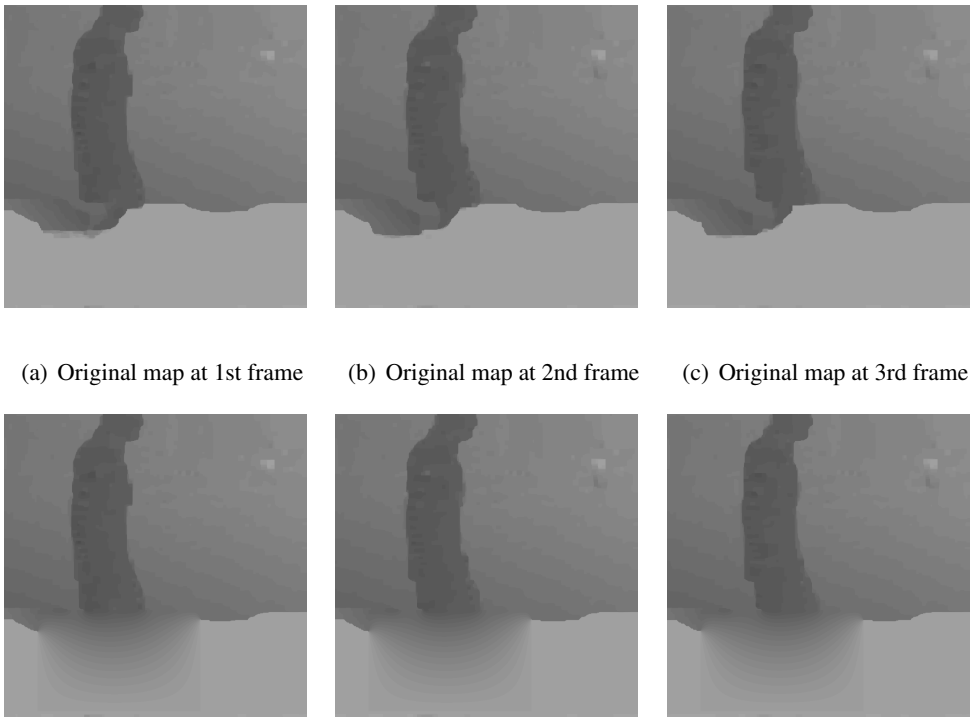
(a) Original map at 1st frame (b) Original map at 2nd frame (c) Original map at 3rd frame

(d) Corrected map at 1st frame (e) Corrected map at 2nd frame (f) Corrected map at 3rd frame

Figure 56. Horizontal flow maps in three consecutive frames.

7.4.5 Quantitative Analysis

To quantify the stitching quality of ROI before and after the proposed correction, we employ the structural similarity index (SSIM) to characterize the similarity between target ROI in the panorama and reference ROI in the input image. The SSIM index quality assessment consists of the luminance term, the contract term, and the structural term. The final similarity score is the multiplicative combination of them.



(a) Original map at 1st frame (b) Original map at 2nd frame (c) Original map at 3rd frame

(d) Corrected map at 1st frame (e) Corrected map at 2nd frame (f) Corrected map at 3rd frame

Figure 57. Vertical flow maps in three consecutive frames.

$$SSIM(x,y) = [l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s[x,y]]^{\gamma}, \tag{7.10}$$

where

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{7.11}$$

$$l(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{7.12}$$

(a) Original ROI at 1st frame (b) Original ROI at 2nd frame (c) Original ROI at 3rd frame

(d) Corrected ROI at 1st frame (e) Corrected ROI at 2nd frame (f) Corrected ROI at 3rd frame

Figure 58. ROI before and after correction in three consecutive frames. The three sub-figures on the first row ares ROI before correction. The three sub-figures on the second row ares ROI after correction

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3}. \tag{7.13}$$

In the above equation, μ_x and μ_y the local means for images x and y. σ_x , σ_y, and σ_{xy} are the corresponding standard deviations and cross-covariance. In our experiment, we usually set coefficients α, β, and γ as 1 and C_3 as $C_2/2$. Therefore, the similarity score can be simplified to:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_x\sigma_y + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{7.14}$$

In Table VII, we can see the stitching quality of target ROI before correction and after correction in different datasets. For those still frames, we manually find all the ROIs and compute their SSIM before and after the correction. The SSIM we stated in Table VII is the average similarity score of these all detected ROIs. For those panoramic video data-sets, we also manually find the ROIs at the first frame and correct the same region in later frames when stitching errors detected. The SSIM we displayed is the average of all target ROIs in all the frames we operate the proposed correction technique.

7.5 Conclusion

In this chapter, we presented one post-stitching flow-map-guided correction technique to mitigate all the manually labeled artifacts in the output panorama. According to the registered ROI in the input image, we establish a pixel-to-pixel correspondence between panorama and image coordinate. The original flow map indicates the position displacement between the target ROI and the reference ROI can be updated with newly fitted movement via interpolation. Based on the more smoothly distributed flow map, we can reconstruct the target ROI with the corresponding pixel from the input image by reverse mapping. The updated flow map then constrains this warping process to be smooth, which inhibits the generation of artifacts in the reconstructed ROI. Finally, panoramas from various synthetic and real datasets are tested with the proposed correction. Both the visual examples and quantitative analysis demonstrate the effectiveness of our proposed post-stitching correction technique. Since the proposed method utilizes the warping information from those corrected stitched regions, it will not work well when there is severe distortions or significant plane misalignment. Another limitation of this correction

technique is its lack of essential consideration of the consistency in the revision to the stereoscopic panorama and panoramic video. For a more robust correction result and broader application scenes, more works need to be investigated:

1. Propose one automatic ROI detection method to replace the manual labeling operation;
2. Define one reasonable geometric error metric to characterize the improvement of correction;
3. Incorporate stereo and temporal constraints to handle the stereoscopic and panoramic video problem better.

The automatic ROI detection can be established on the variance of the pixel position displacements between output panorama and the input image, which may correspond to discontinues and object cropping. To define one reasonable geometric error metric, the neural network and deep learning may help us to train one convincing quality-grader based on a quantity of human-evaluated data-sets.

TABLE VII

ROI SIMILARITY COMPARISON

	Before Correction	After Correction
Atrium	0.8971	0.9025
Basement	0.9144	0.9198
Pavilion	0.9148	0.9206
Classroom	0.9166	0.9216
Courtyard	0.9193	0.9232
Campus	0.9201	0.9235
Rampart	0.9119	0.9230
Synthetic Indoor (Video)	0.9189	0.9216
Synthetic Outdoor (Video)	0.9189	0.9218
Bearstone (Video)	0.9171	0.9204
Courtyard (Video)	0.9323	0.9474

CHAPTER 8

POST-STITCHING FEATURE-BASED DEPTH ADJUSTMENT

The content of this chapter is based on our work that is published in [3] .©2019 IEEE. Reprinted with permission, from [3].

8.1 Background and Related Works

Traditional monocular panorama stitching methods cannot handle stereo consistency well due to its lack of depth information utilization [17, 49, 50]. Therefore, recently proposed stereoscopic panorama generation methods provide various solutions to alleviate the stereo visual discomfort caused by the inaccurate depth information [4, 5, 10, 51]. Based on the rotation of cameras in a circular trajectory or static radial camera array, the omnistereo projection method [11] and its extensions [5,13,14,52] usually implement depth adjustment operation by careful selection of corresponding left-view and right-view strips. However, Richardt et al. proposed to make compensation to the vertical disparity before stitching by projecting undistorted input images onto a cylindrical imaging surface [12]. Zhang and Liu also extended a spatially varying warping method [16] to warp the input images under the guidance of a well-stitched dense disparity map [4]. However, all of the above depth control strategies are highly correlated to their unique hardware setup and stitching algorithm, which indicates the difficulty in generalization and extension. Thus, we intend to propose one general depth correction strategy that can fit different camera arrangements, captured scenarios and panorama generation methods. The whole depth adjust-

ment process can only depend on the input- rectified image pairs and originally stitched stereoscopic panoramas.

In this chapter, we mainly discuss the depth correction and adjustment to the stereoscopic panorama and panoramic video after the monocular correction in the previous chapter. Given the extensive investigations in high-quality stereoscopic panorama generation and widespread usage of VR display equipment, the comfortable immersive visual experience of real-world scenes are always expected by audiences. The criteria for the stitching quality of stereoscopic panoramas not only includes the misalignment, stitching errors, and object distortion but also relies on the accuracy of the depth information. Although many complicated techniques and hardware-orientated solutions are proposed to handle the depth control problem, most of the correction, refinement, and adjustment are operated before or in the panorama generation process. Those later introduced depth disturbance, such as inconsistent blending seams and panorama straightening, are always ignored. Thus, the goal of this investigation is to provide an efficient general post-stitching depth correction strategy that could minimize depth error and stereo inconsistency with sparsely sampled depth information.

8.2 Proposed Depth Adjustment

There are two step in our proposed correction strategy to correct the perceived depth into a comfortable range. Based on the well-matched CIF set from input images, we first operate the global translation to adjust the relative pose between left-view and right-view panoramas. The fitted translation vector, which causes the minimal stereo inconsistency, can ensure the majority of pixels in the stitched panoramas correctly deliver depth information. Then, we utilize the thin-plate-spline warping method to fix all the noticeable depth errors in small regions after the global correction, according to the target disparity

map from the input rectified image. For the stitched stereoscopic video, we will operate the proposed global and local depth adjustment on every single frame.

8.2.1 Global Depth Adjustment

The first correction step can be interpreted as global registration between the left and right panorama. We wish to adjust the output panoramas with a translation vector $< d_v, d_h >$ for better global depth perception. For simplicity, we explain its details in the stitching task for only two pairs of input images, $\{L_1, L_2, R_1, R_2\}$. The basic unit we used for global depth correction is CIF, which refers to the same corner, edge, or region observed and precisely described by all of the adjacent camera views. The technique for the detection and construction of the CIF in [1] is utilized here. Thus, for the two pairs of images, we can produce one corresponding CIF set $S = \{d_{i,1}, d_{i,2}, d_{i,3}, d_{i,4}; i = 1 : N\}$. In the ideal case, those features' final projection position in the output panorama is expected to provide identical depth information as what we perceive from the input images. Then, two corresponding stereo consistency errors can be defined as:

$$E_v(i) = |d^{'}_{i,1}.y - d^{'}_{i,3}.y| + |d^{'}_{i,2}.y - d^{'}_{i,4}.y|, \tag{8.1}$$

$$\begin{aligned} E_h(i) &= |(d^{'}_{i,1}.x - d^{'}_{i,3}.x) - (d_{i,1}.x - d_{i,3}.x)| \\ &+ |(d^{'}_{i,2}.x - d^{'}_{i,4}.x) - (d_{i,2}.x - d_{i,4}.x)|. \end{aligned} \tag{8.2}$$

In the above two equations, $d_{i,1}$, $d_{i,2}$, $d_{i,3}$, and $d_{i,4}$ are the ith four matched features in the CIF set S. The primed symbols $d^{'}_{i,1}$, $d^{'}_{i,2}$, $d^{'}_{i,3}$, and $d^{'}_{i,4}$ are their corresponding features in the generated panorama. Additionally, $d_{i,1}.x$ and $d_{i,1}.y$ represent the feature's center point position.

Thus, the stereo consistency errors after the translation operation with d_v and d_h are:

$$E_v^g(i, d_v) = |d_{i,1}^{'}.y + d_v - d_{i,3}^{'}.y| + |d_{i,2}^{'}.y + d_v - d_{i,4}^{'}.y|, \tag{8.3}$$

$$\begin{aligned} E_h^g(i, d_h) &= |(d_{i,1}^{'}.x + d_h - d_{i,3}^{'}.x) - (d_{i,1}.x - d_{i,3}.x)| \\ &+ |(d_{i,2}^{'}.x + d_h - d_{i,4}^{'}.x) - (d_{i,2}.x - d_{i,4}.x)|. \end{aligned} \tag{8.4}$$

The global depth correction can be formulated as the optimization problem to fit two motion scalars $\hat{d_v}$ and $\hat{d_h}$ that cause the minimal stereo consistency errors for all CIF in set S:

$$\hat{d_v} = arg\min_{d \in \Re} \sum_{i=1}^{N} v_i E_v^g(i, d_v), \tag{8.5}$$

$$\hat{d_h} = arg\min_{d \in \Re} \sum_{i=1}^{N} v_i E_h^g(i, d_h). \tag{8.6}$$

Weight v_i indicate the visual saliency index of the corresponding CIF [23,53], which characterize the visual importance of features. The corporation of saliency weights can force the estimated translation to care more about those features that attract more attention from viewers.

Two disparity maps from stereoscopic panoramas are shown in Figure 59. In the first row, before the global depth correction, we can see that the overall distribution of depth suffers from serious depth compression. All the objects in this disparity map are assigned with small values. The different color around the nearby region indicates the spatial discontinuities of the depth information. However, in the

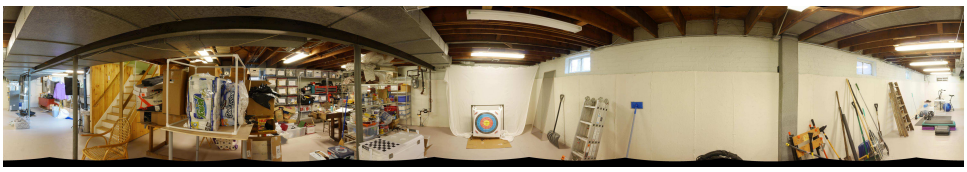
second row, we can see the depth more distributed uniformly and smoothly, which provides an easy understanding scenario to the viewers. Two cropped stereoscopic panoramas in red-cyan anaglyph are shown in Figure 60. In the right panorama, we can see the vertical disparity issue after the global translation is nearly eliminated. The horizontal disparity of bicycle and shovel are also adapted to one reasonable range that delivers correct depth information.

8.2.2 Local Depth Adjustment

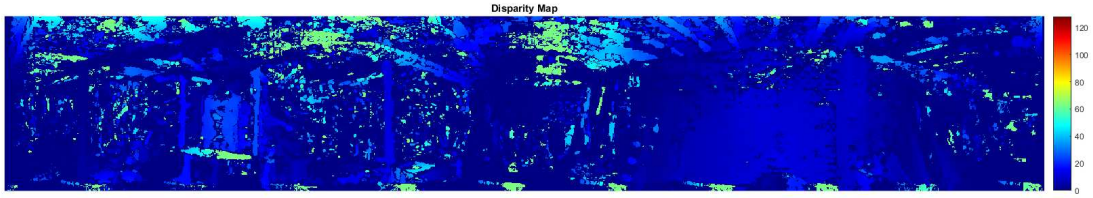
While the global correction can largely relieve the viewing discomfort, we can always find some tiny artifacts in its corrected result. To remove these undesirable issues, one TPS-based morphing method is proposed to fix the region of the interest under the guidance of the target disparity map. Without loss of generality, we consider the right- view panorama with better monocular stitching quality as the reference and perform TPS warping at the left-view panorama in the following discussion.

Control Point Generation

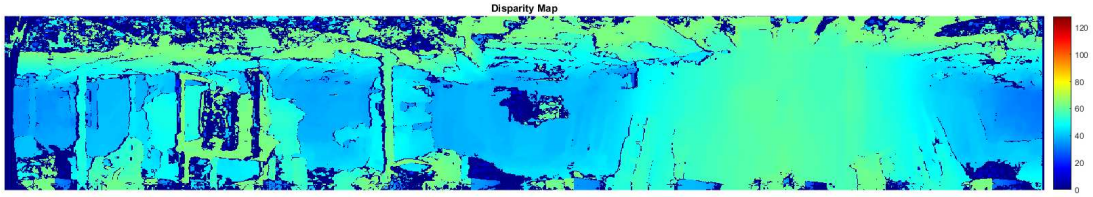
Operation of TPS warping requires two equally sized corresponding point-sets in the areas with depth anomaly. The region of the interest is usually manually labeled for depth correction and denoted as G_L and G_R. Afterward, one group of control points $\{P^L_{i,j}\}$ can be detected and extracted from the left-view ROI, where i is the index for the sampled pixel, and j is the index for ROI. Given the projection matrix H, which describes the geometrical transformation from the camera view to the panorama view, we utilize its inverse function to obtain the position of the corresponding point at the input- rectified image coordinate. According to the target disparity map from input images, the expected disparity of those sampled control points can be computed easily. Thus, the expected position of control point after the local warping can be defined as:



(a) Left-view output panorama

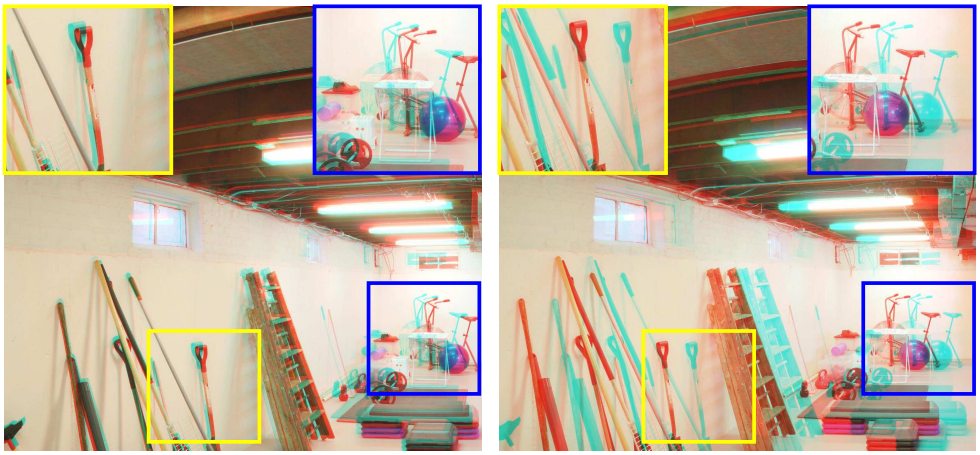


(b) Before global correction



(c) After global correction

Figure 59. Disparity map comparison between before and after global depth correction



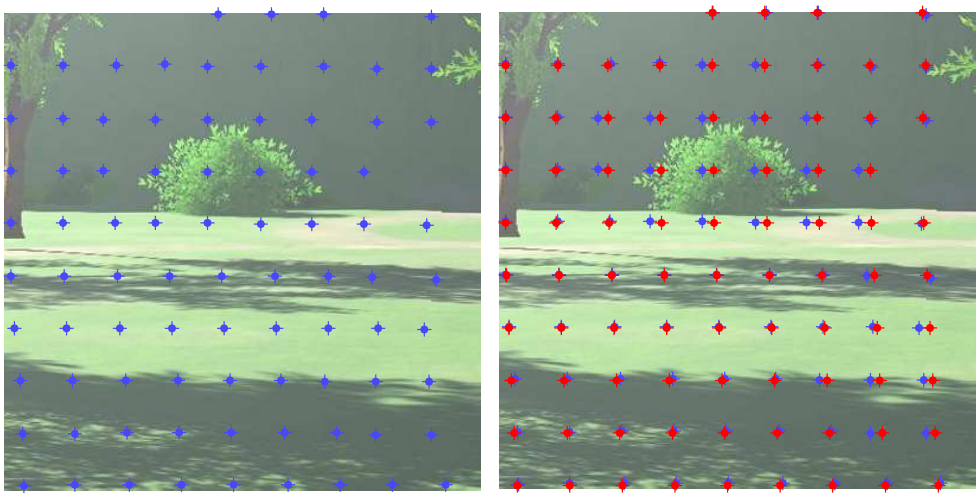
(a) Before global correction (b) After global correction

Figure 60. ROI comparison between before and after global depth correction

$$\hat{P}^L_{i,j}.x = P^R_{i,j}.x + Disp(H^{-1}(P^L_{i,j}.x, P^L_{i,j}.y)) \times R, \tag{8.7}$$

$$\hat{P}^L_{i,j}.y = P^R_{i,j}.y. \tag{8.8}$$

In the above equations, $P^L_{i,j}.x$, $P^L_{i,j}.y$, $P^R_{i,j}.x$, and $P^R_{i,j}.y$ represent the position of the selected control point in the left and right-view panorama. The hatted symbols $\hat{P}^L_{i,j}.x$ and $\hat{P}^L_{i,j}.y$ refer to the expected position of the sampled control points in the left-view panorama. Moreover, H^{-1} is the inverse of the projection matrix, $Disp$ is the disparity map from the rectified image pair and R is the ratio of pixel per degree between the panorama view and camera view.



(a) Original position (b) Expected position

Figure 61. Control point generation result. Blue stars and red stars indicate the original and expected position of sampled control points, respectively. To achieve correct disparity values for each cp between left-view and right-view panorama, the majority of control points are expected to move left-ward in the original left-view ROI.

In the example depicted in Figure 61, the blue stars in sub-figure (a) marks the position of the sampled control points and the red stars in sub-figure (b) indicate their expected position with correct depth. Under this salutation, the right-view panorama is fixed as the reference, and those sampled control points in the left-view ROI are expected to move left-ward for expected disparity values.

Depth-aware TPS Warping

The standard TPS [54] can fit one mapping function, Φ, between the two equally sized corresponding point-sets, $A = \{x_a, y_a\}$ and $B = \{x_b, y_b\}$, with minimal bending energy:

$$\begin{aligned} E_{tps}(\Phi) &= \sum_{i=1}^{M} ||v_i - \Phi(x_{a,i}, y_{a,i})||^2 \\ &+ \iint_R [(\frac{\partial^2 \Phi}{\partial x^2}) + 2(\frac{\partial^2 \Phi}{\partial x \partial y})^2 + (\frac{\partial^2 \Phi}{\partial y^2})]\, dx\, dy \end{aligned} \tag{8.9}$$

In our application, we set v_i equal to the target coordinates (x_b, y_b) in turn to obtain two continuous transformations for the x and y coordinate respectively. Point set A is defined as the original control point $P^L_{i,j}$ and B refers to the corresponding point $\hat{P}^L_{i,j}$ at the expected position.

According to the proof in reference [55], the unique minimizer, Φ, is parameterized as follows:

$$\Phi(x, y) = \gamma_1 + \gamma_2 * x + \gamma_3 * y + \sum_{i=1}^{M} w_i U(|(x_{a,i}, y_{a,i}) - (x, y)|) \tag{8.10}$$

$$U(r) = \begin{cases} r^2 \ln r & r > 0 \\ 0 & r = 0 \end{cases} \tag{8.11}$$

Therefore, we intend to find the coefficients $[w|\gamma_1, \gamma_2, \gamma_3]$ in mapping function Φ. For Φ to be square-integrable at second derivatives, we require that:

$$\sum_{i=1}^{M} w_i = \sum_{i=1}^{M} w_i x_{a,i} = \sum_{i=1}^{M} w_i y_{a,i} = 0. \tag{8.12}$$

Together with the exact interpolation conditions, $\Phi(x_{a,i}, y_{a,i}) = v_i$, this produces a linear system as follows:

$$\begin{bmatrix} \boldsymbol{K} & \boldsymbol{P} \\ \boldsymbol{P}^T & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{0} \end{bmatrix}, \tag{8.13}$$

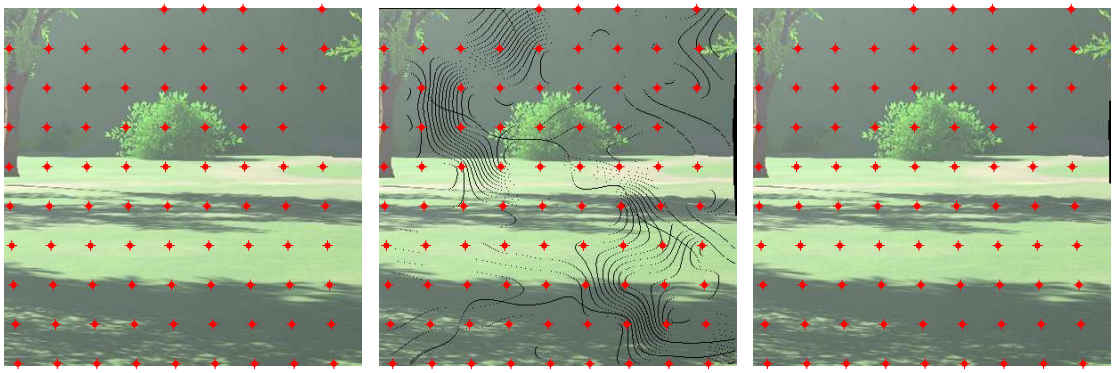
where $\boldsymbol{K}$, $\boldsymbol{P}$, and $\boldsymbol{O}$ are submatrices:

$$\begin{aligned} K_{i,j} &= U(|(x_{a,i}, y_{a,i}) - (x_{a,j}, y_{a,j})|) \\ \boldsymbol{P}_{M\times 3} &= \begin{bmatrix} 1 & x_{a,1} & y_{a,1} \\ 1 & x_{a,2} & y_{a,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{a,M} & y_{a,M} \end{bmatrix} \\ \boldsymbol{O}_{3\times 3} &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \end{aligned} \tag{8.14}$$

and $\boldsymbol{w}, \boldsymbol{\gamma}$, and $\boldsymbol{v}$ are column vectors, which stand for TPS coefficients and target data value respectively. Then, we can obtain the TPS interpolation coefficients as:

$$\begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \boldsymbol{K} & \boldsymbol{P} \\ \boldsymbol{P}^T & \boldsymbol{O} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{0} \end{bmatrix}. \tag{8.15}$$

Once TPS coefficients $[w|\gamma_1, \gamma_2, \gamma_3]$ are computed, we use the TPS equation (10) to find the expected position for those un-sampled points in the ROI. After all pixels in the ROI have been projected into their new position via the TPS coefficients, we will obtain the adjusted ROI with correct depth information. In Figure 62, sub-figure (a) is the unwrapped ROI with control points at their expected position. Then, sub-figure (b) shows the warped ROI after the projection of all pixels under TPS coefficients. Finally, sub-figure (c) is the warped ROI after all black holes have been filled via nearest-neighbor interpolation.



(a) (b) (c)

Figure 62. Thin-plate-spline warping

8.3 Performance Evaluation

8.3.1 Experiment Setup

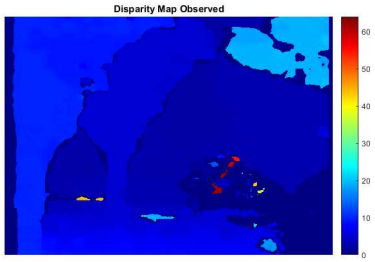
To validate the feasibility and robustness of our proposed flow map guided panorama correction technique, we employ the feature-based stereoscopic panorama stitching framework [1] as the baseline. In the panorama generation process, $vlfeat$ [40] lib is used for SIFT detection and $Enblend$ [8] is the panorama blender. All stitched panoramas are scaled to 12000 by 6000 pixels for $360^\circ \times 180^\circ$. Each local region is manually labeled as a 400 by 400 pixel rectangle, and there are 400 (20 by 20) control points uniformly sampled for each local region warping.

8.3.2 Quantitative Analysis

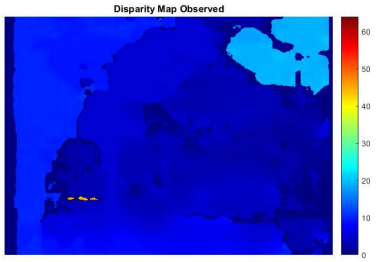
Both the camera-captured and synthetic data are tested to quantify the improvement of our proposed depth correction to the original stitching result. Considering the disparity map from the input rectified image pairs as ground truth and uniformly sampled features as testing control points, the difference between the perceived depth of testing control points from stereoscopic panoramas and expected depth from ground truth is then used as the metric to evaluate the performance of depth adjustment. The pixel-level depth error for global and local correction is stated in Table VIII and Table IX, respectively. The testing data-set includes four frames of real-captured outdoor scenarios, two frames of synthetic indoor scenarios, and five animations. The missing information of several static camera-based datasets-sets in Table IX indicates no local depth anomaly region was found. The depth error recorded in the tables is the average of 30 frames of panoramas in each animation data-set, while sometimes the pixel-wise metric fails to characterize the improvement for human visual perception to depth.

(a) Right-view panorama before correction

(b) Right-view panorama after correction



(c) Before correction



(d) Globally corrected

Figure 63. Example of ROI before and after correction

TABLE VIII

DEPTH ERROR BEFORE AND AFTER GLOBAL CORRECTION

	Horizontal Dist		Vertical Dist	
	Before	After	Before	After
Atrium	1.62px	1.41px	1.15px	1.12px
Basement	2.69px	0.67px	1.89px	0.84px
Campus	2.13px	2.02px	1.53px	1.53px
Rampart	28.87px	1.81px	1.79px	1.19px
Barcelona	1.96px	1.54px	2.30px	2.23px
Classroom	1.30px	1.27px	3.89px	1.58px
Village-a	1.62px	1.41px	1.15px	1.12px
Village-b	1.40px	1.29px	1.21px	1.11px
Village-c	1.35px	1.30px	1.05px	1.02px
Living-room-a	1.09px	0.91px	1.96px	1.94px
Living-room-b	0.68px	0.59px	1.73px	1.65px

TABLE IX

DEPTH ERROR BEFORE AND AFTER LOCAL CORRECTION

	Horizontal Dist		Vertical Dist	
	Before	After	Before	After
Atrium	2.78px	1.93px	1.13px	1.12px
Basement	3.40px	1.10px	1.10px	0.89px
Village-a	1.95px	0.65px	1.08px	1.05px
Village-b	1.70px	0.63px	1.23px	1.04px
Village-c	1.85px	0.57px	1.10px	1.08px
Living-room-a	2.70px	1.58px	2.03px	1.94px
Living-room-b	2.10px	1.47px	1.87px	1.86px

(a) (b) (c) (d) (e)

Figure 64. Disparity map of ROI before and after correction. sub-figure (a) is the expected disparity map. Sub-figure (b) is the disparity map of selected ROI. Sub-figure (c) is the ROI before any depth correction. Sub-figure (d) is the ROI after global correction. Sub-figure (e) is the ROI after local correction.

8.3.3 Visual Comparison

One instance in Figure 63 visually demonstrates improvement to the perceived disparity map. Sub-figure (a) and (b) are the right-view panoramas before the local correction and after the local correction. Sub-figure (c) and (d) are the corresponding disparity map. It is noted that the depth map after local correction makes the pixels in the ROI achieve a color more similar to its neighbors around the ROI, which implies better depth consistency in the spatial coordinate.

Another instance in Figure 64 visually demonstrates improvement to the perceived disparity map. Sub-figure (a) is the full-size target disparity map estimated from input rectified image pairs, and sub-figure (b) is the cropped version for the selected ROI. Sub-figure (c), (d), and (e) show the measured disparity map of ROI before global correction, after global correction and after local correction, respectively. The global correction shifts the overall disparity value of ROI into the range more similar to the

target map. Moreover, the perceived disparity value of the human's right arm, which is marked with the red rectangle, is also corrected.

8.4 Conclusion

In this chapter, we presented a general depth correction strategy in the stereoscopic panoramic video generation system. Given the well-matched commonly -identified features from input image pairs, we can consider one of the stitched stereoscopic panoramas as the reference and globally shift another one for minimal average depth error. After we manually label those regions with local noticeable depth issues, we can fix them with the thin-plate-spline warping under the guidance of a sparse target disparity map. Though the global depth correction can roughly guarantee the overall reasonable depth distribution across the whole image, the local region warping still needs the extra human intervention to label the ROI with noticeable stitching errors. This post-stitching depth adjustment operation is the last step in our proposed stitching system. To extend the proposed depth adjustment technique to more complicated scenes or stitching output from other panorama generation frameworks, we need to propose one automatic ROI detection algorithm to search all regions with abnormal disparity values.

CHAPTER 9

SUMMARY

In this thesis, we presented a general framework to generate stereoscopic panoramic videos based on commonly identified features. To achieve this goal, we improve and extend the standard panorama stitching pipeline, in particular:

1. We proposed the definition of commonly identified features and partially occluded features based on whether they can be fully captured via all input camera views. The proposed feature structure is designed to deal with the stereoscopic panoramic stitching task. Compared to the independent stitching strategy based on the standard feature descriptor, the proposed CIF can align the left-view and right-view input images with better stereo consistency and maintain the high-quality of mononuclear monocular stitching simultaneously.

2. We extended the standard 2D feature matching process into the 3D sense for CIF under the guidance of the depth-aware information. The original 2D feature descriptor can be matched with gradient information, while the proposed CIF is designed to take advantage of depth information in each of the four matched 2D feature descriptors. The extra depth-based qualifying term helps to filter out false corresponded control points and provide more accurate image alignment.

3. We introduced human visual interest to prune the redundant feature descriptors and adjust the distribution of the selected control points for image alignment. The control points after saliency-based feature selection are distributed across the overlapping region more uniformly and reason-

ably. Appropriate distribution of control points according to the proposed energy map can adapt the image alignment to place more emphasis on regions that attract more attention from viewers.

4. We proposed the corresponding saliency-aware and depth-aware CIF tracking strategy in the video sequence. The stitching strategy based on pure independent detection for every single frame can maintain good stitching- quality but will cause serious video instability between consecutive frames. The stitching strategy based on pure tracking will lose the trajectory of control points due to the occlusion and drift problem after several frames. Compared to those two strategies, the proposed saliency-aware stitching strategy can keep the temporally consistent image alignment. For more details, the proposed strategy will keep original control points when no moving object enters the overlapping region and only conducts updates when noticeable content change appears in the overlapping area. This integrated strategy can maintain image alignment information for monocular stitching correctness and also concerns about the temporal consistency between consecutive frames.

5. We proposed one modified version of the RANSAC-based homography estimation algorithm for a more consistent output panorama. Instead of feature selection from the standard control point set in left-view and right-view independently, the proposed stereo-constrained algorithm will select corresponding control points from the CIF set. The similarity penalty term also constrains the left-view and right-view input images to be aligned in the final output canvas under similar projection matrices.

6. We proposed one post-stitching correction technique to mitigate the visible artifacts in the output panorama under the guidance of densely corresponded position displacement maps. The pixel-

based corresponding map between the target ROI from the panorama and reference ROI from the input image can indicate the warping relationship. Thus, with the updated position displacement of target ROI according to the corrected stitched neighbors, we can mitigate those discontinues and stitching errors that were not fully removed by the refinement and optimization techniques in the standard panorama generation framework. After this flow-map-guided panorama correction, the monocular sense stitching quality can be improved.

7. We proposed one post-stitching correction technique to globally adjust the left-view and right-view panorama for more accurate depth delivery under the guidance of sparsely distributed commonly identified features. Given one of the output stereoscopic panorama as the reference, we intend to globally translate another view panorama horizontally and globally so that the majority of the CIF can be aligned reasonably in the output canvas. For those patches and areas that still suffer from depth issues, we utilize one feature-based image morphing method, thin-plate-spline, to warp all the selected control points into their position with expected disparity values. After this feature-based panorama correction, the perceived depth information from the output panorama can be more similar to the depth information estimated input image pairs, which is considered as ground truth in our correction process.

8. We tested our proposed stereoscopic panoramic video generation system with several different image acquisition equipment: SENSEICam Simulator, StarCam Design, and Chameleon Design. The difference in camera array structures always needs carefully adapted calibration or even an extra raw image preprocessing operation. The experiments and simulation based on these image

acquisition equipment demonstrate the compatibility of our proposed stitching system to various camera position setups for image capture.

9. We tested our proposed stereoscopic panoramic video generation system under various scenarios: still and dynamic data-sets, synthetic and camera-based data-sets, and indoor and outdoor data-sets. The experiments and simulation based on these image data-sets illustrate the robustness of our proposed stitching system in various scenarios.

Though the proposed stitching system works well under various camera array designs and scenarios, it still has some limitations in several areas:

1. Our current stitching system can work well with the fixed camera array and limited moving objects in the overlapping field. Under the situation of moving camera array and complicated scenarios, the current stitching system will not promise high-quality output due to the depth change of the background and the degraded feature tracking strategy.
2. The proposed stitching framework includes several pre-precessing and post-correction operations and is implemented without any optimization so that the real-time stereoscopic panoramic video processing is unavailable now. To speed up the panorama generation process, some steps in the stitching framework should be carefully modified to benefit from parallel computation techniques.
3. Both the monocular-view post-stitching artifact correction and local depth adjustment need manual ROI labeling before the operation, which indicates these corrections may not be stable and sufficient for the perfect stitching output. Automatic erroneous ROI detection and selection is necessary for the efficient correction and adjustment to the output video.

APPENDIX

COPYRIGHT PERMISSIONS

This Appendix includes the copyright permissions granted from the IEEE to use published work in thesis.

Copyright Permission from IEEE: The following statement has been copied from the CopyRight Clearance Center (RightsLink).

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you must follow the requirements listed below:

Textual Material

In the case of textual material (e.g., using short quotes or referring to the work within these papers), users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line ©20XX IEEE.

In the case of illustrations or tabular material, we require that the copyright line ©[Year of original publication] IEEE appear prominently with each reprinted figure and/or table.

If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior authors approval.

Full-Text Article

If you are using the entire IEEE copyright owned article, the following IEEE copyright/ credit notice should be placed prominently in the references: ©[year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]

Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of University of Illinois at Chicago's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to $http://www.ieee.org/publications\_standards/publications/rights/rights\_link.html$ to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

CITED LITERATURE

1. Wang, H., Sandin, D. J., and Schonfeld, D.: A common feature-based disparity control strategy in stereoscopic panorama generation. In Visual Communications and Image Processing (VCIP), 2017 IEEE, pages 1–4. IEEE, 2017.

2. Wang, H., Sandin, D. J., and Schonfeld, D.: Saliency-based feature selection strategy in stereoscopic panoramic video generation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1837–1841. IEEE, 2018.

3. Wang, H., Sandin, D. J., and Schonfeld, D.: Post-stitching depth adjustment for stereoscopic panorama. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2282–2286. IEEE, 2019.

4. Zhang, F. and Liu, F.: Casual stereoscopic panorama stitching. In CVPR, pages 2002–2010, 2015.

5. Anderson, R., Gallup, D., Barron, J. T., Kontkanen, J., Snavely, N., Hernández, C., Agarwal, S., and Seitz, S. M.: Jump: virtual reality video. ACM Transactions on Graphics (TOG), 35(6):198, 2016.

6. Furukawa, Y. and Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE transactions on pattern analysis and machine intelligence, 32(8):1362–1376, 2010.

7. Dersch, H.: Panotools contributors (2010).

8. Mihal, A., dAngelo, P., et al.: Enblend, 2012.

9. Szeliski, R.: Image alignment and stitching: A tutorial. Foundations and Trends® in Computer Graphics and Vision, 2(1):1–104, 2006.

10. Huang, H.-C. and Hung, Y.-P.: Panoramic stereo imaging system with automatic disparity warping and seaming. Graphical Models and Image Processing, 60(3):196–208, 1998.

11. Peleg, S., Ben-Ezra, M., and Pritch, Y.: Omnistereo: Panoramic stereo imaging. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3):279–290, 2001.

12. Richardt, C., Pritch, Y., Zimmer, H., and Sorkine-Hornung, A.: Megastereo: Constructing high-resolution stereo panoramas. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 1256–1263. IEEE, 2013.

13. Couture, V., Langer, M. S., and Roy, S.: Panoramic stereo video textures. In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 1251–1258. IEEE, 2011.

14. Couture, V. C., Langer, M. S., and Roy, S.: Omnistereo video textures without ghosting. In 3D Vision-3DV 2013, 2013 International Conference on, pages 64–70. IEEE, 2013.

15. Chapdelaine-Couture, V. and Roy, S.: The omnipolar camera: A new approach to stereo immersive capture. In Computational Photography (ICCP), 2013 IEEE International Conference on, pages 1–9. IEEE, 2013.

16. Liu, F., Gleicher, M., Jin, H., and Agarwala, A.: Content-preserving warps for 3d video stabilization. In ACM Transactions on Graphics (TOG), volume 28, page 44. ACM, 2009.

17. Brown, M. and Lowe, D. G.: Automatic panoramic image stitching using invariant features. International journal of computer vision, 74(1):59–73, 2007.

18. Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding, 110(3):346–359, 2008.

19. Lowe, D. G. et al.: Object recognition from local scale-invariant features. In iccv, volume 99, pages 1150–1157, 1999.

20. Grundmann, M., Kwatra, V., Castro, D., and Essa, I.: Calibration-free rolling shutter removal. In Computational Photography (ICCP), 2012 IEEE International Conference on, pages 1–8. IEEE, 2012.

21. Bay, H., Tuytelaars, T., and Van Gool, L.: Surf: Speeded up robust features. In European conference on computer vision, pages 404–417. Springer, 2006.

22. Yen, T.-C., Tsai, C.-M., and Lin, C.-W.: Maintaining temporal coherence in video retargeting using mosaic-guided scaling. IEEE Transactions on Image Processing, 20(8):2339–2351, 2011.

23. Harel, J., Koch, C., and Perona, P.: Graph-based visual saliency. In Advances in neural information processing systems, pages 545–552, 2007.

24. Itti, L. and Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision research, 40(10-12):1489–1506, 2000.

25. Zheng, M., Chen, X., and Guo, L.: Stitching video from webcams. In International Symposium on Visual Computing, pages 420–429. Springer, 2008.

26. He, B., Zhao, G., and Liu, Q.: Panoramic video stitching in multi-camera surveillance system. In Image and Vision Computing New Zealand (IVCNZ), 2010 25th International Conference of, pages 1–6. IEEE, 2010.

27. El-Saban, M. A., Refaat, M., Kaheel, A., and Abdul-Hamid, A.: Stitching videos streamed by mobile phones in real-time. In Proceedings of the 17th ACM international conference on Multimedia, pages 1009–1010. ACM, 2009.

28. Shimizu, T., Yoneyama, A., and Takishima, Y.: A fast video stitching method for motion-compensated frames in compressed video streams. In Consumer Electronics, 2006. ICCE'06. 2006 Digest of Technical Papers. International Conference on, pages 173–174. IEEE, 2006.

29. Xu, W. and Mulligan, J.: Panoramic video stitching from commodity hdtv cameras. Multimedia systems, 19(5):407–426, 2013.

30. Jiang, W. and Gu, J.: Video stitching with spatial-temporal content-preserving warping. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on, pages 42–48. IEEE, 2015.

31. El-Saban, M., Izz, M., Kaheel, A., and Refaat, M.: Improved optimal seam selection blending for fast video stitching of videos captured from freely moving devices. In Image Processing (ICIP), 2011 18th IEEE International Conference on, pages 1481–1484. IEEE, 2011.

32. El-Saban, M., Izz, M., and Kaheel, A.: Fast stitching of videos captured from freely moving devices by exploiting temporal redundancy. In Image Processing (ICIP), 2010 17th IEEE International Conference on, pages 1193–1196. IEEE, 2010.

33. Su, T., Nie, Y., Zhang, Z., Sun, H., and Li, G.: Video stitching for handheld inputs via combined video stabilization. In SIGGRAPH ASIA 2016 Technical Briefs, page 25. ACM, 2016.

34. Lin, K., Liu, S., Cheong, L.-F., and Zeng, B.: Seamless video stitching from hand-held camera inputs. In Computer Graphics Forum, volume 35, pages 479–487. Wiley Online Library, 2016.

35. Guo, H., Liu, S., He, T., Zhu, S., Zeng, B., and Gabbouj, M.: Joint video stitching and stabilization from moving cameras. IEEE Transactions on Image Processing, 25(11):5491–5503, 2016.

36. Nie, Y., Su, T., Zhang, Z., Sun, H., and Li, G.: Dynamic video stitching via shakiness removing. IEEE Transactions on Image Processing, 27(1):164–178, 2018.

37. Fischler, M. A. and Bolles, R. C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In Readings in computer vision, pages 726–740. Elsevier, 1987.

38. Hartley, R. and Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press, 2003.

39. Meyer, D., Wang, H., Sandin, D., McFarland, C., Lo, E., Dawe, G., Dai, J., Nguyen, T., Baker, H., Brown, M., et al.: Starcam-a 16k stereo panoramic video camera with a novel parallel interleaved arrangement of sensors. Electronic Imaging, 2019(3):646–1, 2019.

40. Vedaldi, A. and Fulkerson, B.: Vlfeat: An open and portable library of computer vision algorithms. In Proceedings of the 18th ACM international conference on Multimedia, pages 1469–1472. ACM, 2010.

41. Vincent, M. L., DeFanti, T., Schulze, J., Kuester, F., and Levy, T.: Stereo panorama photography in archaeology: Bringing the past into the present through cavecams and immersive virtual environments. In 2013 Digital Heritage International Congress (DigitalHeritage), volume 1, pages 455–455. IEEE, 2013.

42. Liu, C., Yuen, J., and Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE transactions on pattern analysis and machine intelligence, 33(5):978–994, 2010.

43. Liu, S., Yuan, L., Tan, P., and Sun, J.: Bundled camera paths for video stabilization. ACM Transactions on Graphics (TOG), 32(4):78, 2013.

44. Lin, C.-C., Pankanti, S. U., Natesan Ramamurthy, K., and Aravkin, A. Y.: Adaptive as-natural-as-possible image stitching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1155–1163, 2015.

45. Chen, Y.-S. and Chuang, Y.-Y.: Natural image stitching with the global similarity prior. In European Conference on Computer Vision, pages 186–201. Springer, 2016.

46. Zhang, F. and Liu, F.: Parallax-tolerant image stitching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3262–3269, 2014.

47. Cabral, B. K., Hsu, J., and Coward, A. H.: Panoramic virtual reality camera, October 9 2018. US Patent App. 29/569,876.

48. Brown, M., Lowe, D. G., et al.: Recognising panoramas. In ICCV, volume 3, page 1218, 2003.

49. Szeliski, R. et al.: Image alignment and stitching: A tutorial. Foundations and Trends® in Computer Graphics and Vision, 2(1):1–104, 2007.

50. Gao, J., Li, Y., Chin, T.-J., and Brown, M. S.: Seam-driven image stitching. In Eurographics (Short Papers), pages 45–48, 2013.

51. Perazzi, F., Sorkine-Hornung, A., Zimmer, H., Kaufmann, P., Wang, O., Watson, S., and Gross, M.: Panoramic video from unstructured camera arrays. In Computer Graphics Forum, volume 34, pages 57–68. Wiley Online Library, 2015.

52. Wang, C. and Sawchuk, A. A.: Region-based stereo panorama disparity adjusting. In Multimedia Signal Processing, 2006 IEEE 8th Workshop on, pages 186–191. IEEE, 2006.

53. Hou, X., Harel, J., and Koch, C.: Image signature: Highlighting sparse salient regions. IEEE transactions on pattern analysis and machine intelligence, 34(1):194–201, 2012.

54. Bookstein, F. L.: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on pattern analysis and machine intelligence, 11(6):567–585, 1989.

55. Wahba, G. et al.: Spline models for observational data. society for industrial and applied mathematics. 1990.

VITA

NAME Haoyu Wang

EDUCATION

Ph.D. Electrical and Computer Engineering 2011 – 2019
University of Illinois at Chicago, Chicago, United States

B.Eng. Electrical and Information Engineering 2006 – 2010
Huazhong University of Science and Technology, Wuhan, China

PUBLICATIONS

Wang, H., Bouaynaya, N., Shterenberg, R., & Schonfeld, D. (2013, May). Sparse biologically-constrained optimal perturbation of gene regulatory networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1167-1171). IEEE.

Wang, H., Sandin, D. J., & Schonfeld, D. (2017, December). A common feature-based disparity control strategy in stereoscopic panorama generation. In 2017 IEEE Visual Communications and Image Processing (VCIP) (pp. 1-4). IEEE..

Wang, H., Sandin, D. J., & Schonfeld, D. (2018, April). Saliency-based feature selection strategy in stereoscopic panoramic video generation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1837-1841). IEEE.

Sandin, D., Wang, H., Atra, A., Ainsworth, R. A., Brown, M., & DeFanti, T. A. (2019). The Quality of Stereo Disparity in the Polar Regions of a Stereo Panorama. Electronic Imaging, 2019(3), 642-1.

Meyer, D. E., Wang, H., Sandin, D., McFarland, C., Lo, E., Dawe, G., ... & DeFanti, T. (2019). StarCAM-A 16K stereo panoramic video camera with a novel parallel interleaved arrangement of sensors. Electronic Imaging, 2019(3), 646-1.

Meyer, D., Lo, E., McFarland, C., Strawson, J., Drohobytsky, D., Dai, J., ... & Wang, H. (2019, March). Omniscopic Vision for Robotic Control. In 2019 IEEE Aerospace Conference (pp. 1-13). IEEE.

Wang, H., Sandin, D. J., & Schonfeld, D. (2019, May). Post-stitching Depth Adjustment for Stereoscopic Panorama. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2282-2286). IEEE.

EXPERIENCE

Research Assistant in Electronic Visualization Laboratory 2015 - 2019
Department of Computer Science, University of Illinois at Chicago

Research Assistant in Multimedia Communications Laboratory 2011 - 2015
Department of Electrical and Computer Eng., University of Illinois at Chicago