A Study on Monocular Stereo Vision Using High-Speed Catadioptric System

（高速カタディオプトリックシステムを用いた単眼ステレオ視に関する研究）

胡　少鵬

Graduate School of Engineering
Hiroshima University
July, 2018

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 Stereo Vision

Stereo vision is a range-sensing technique for distant real-world scenes using multiple images observed at different viewpoints with triangulation, and many stereo matching algorithms have been reported for stereo disparity map estimation [1–8]; they are classified into (1) global algorithms to perform global optimization for the whole image to estimate the disparity of every pixel with numerical methods [9–14], and (2) local algorithms with window-based matching that only requires local image features in a finite-size window when computing disparity at a given point with the winner-take-all strategy [15–21]. Compared with accurate but time-consuming global algorithms, local algorithms are much less time-consuming in estimating disparity maps, and therefore many real-time stereo systems capable of executing local algorithms have been reported, such as Graphic Processing Unit (GPU)-based stereo matching [22–26] and Field Programmable Gate Array (FPGA)-based embedded systems [27–30].

Typical stereo vision system is made up of two cameras with synchronization to capture stereo images just like human eyes. Depth information can be extracted by examining the relative positions of scene in the two image panels by comparing information from two vantage points, The difference between two images for the same target is called as disparity map, which encodes the difference in horizontal coordinates of corresponding image points. The baseline of these systems are always fixed and short, because it is

difficult to conect two or more cameras for synchronization. Therefore, the field of view is limited and it can not be used as the active vision system to track or capture moving object.

For a wider field of view in stereo measurement without decreasing resolution, many active stereo systems that mount cameras on pan-tilt mechanisms have been reported [31–34]; they are classified into (1) multiple cameras on a single pan-tilt mechanism; and (2) multiple pan-tilt cameras, on which each camera has its pan-tilt mechanism. In the former approach, the relative geometrical relationship between cameras are fixed on the common pan-tilt mechanism in a way that the camera parameters can be easily calibrated for stereo measurement; its measurable range in depth is limited because the vergence angle between cameras is fixed. The latter approach can expand the measurable range in the depth direction, as well as those in the pan and tilt directions, because the vergence angle between cameras can be freely controlled; the camera parameters should be calibrated for accurate stereo measurement frame by frame according to the time-varying vergence angle in stereo tracking.

With the recent spread of distributed camera networks for wide-area video surveillance, many studies that concern on gaze control [35–37], camera calibration [38–43], and image rectification in stereo matching [44, 45], for the latter approach, have been reported for stereo tracking using multiple PTZ (pan-tilt-zoom) cameras located at different sites. However, These systems using two or multiple pan-tilt cameras are difficult to switch their viewpoints quickly and are high cost using multiple cameras or mechanisms. The pros and cons of the stereo vision techniques and the active stereo systems are summarized in Table 1.1.

1.1.2 Monocular Stereo Vision System

Monocular stereo vision means that stereo vision images are obtained using only one camera. Monocular stereo system has the advantages of only one camera used as multiple virtual cameras. Many monocular stereo methods using a single camera have been proposed to reduce the complexity in calibration of camera parameters and synchroniza-

Table 1.1: Pros and cons of stereo vision techniques and active stereo systems.

	Stereo Vision Techniques		Active Stereo Systems	
classification	local methods [15–21]	global methods [9–14]	single pan-tilt mechanism [31, 34]	multiple pan-tilt mechanisms [32, 33]
calibration	direct calibration (eg. Zhang's method [46]) Pros: high calibration precision Cons: suitable for fixed stereo system		self-calibration [41, 43, 44] Pros: automatic parameter acquisition Cons: complex theory and control LUT-based calibration [33, 38] Pros: on-line parameter acquisition Cons: complex preprocessing for LUT feature-based calibration [39, 40, 42] Pros: parameters from image features Cons: time-consuming and imprecise	
advantages	efficient for stereo matching and less time-consuming	accurate matching particularly for ambiguous regions	easy stereo calibration and gaze control	flexible views and extensive depth range
disadvantages	sensitive to locally ambiguous regions	very time-consuming	fixed baseline and limited depth range	real-time stereo calibration and complex gaze control

tion of camera shutter timings when using multiple cameras, These systems have been proposed mainly using catadioptric system or other mechanism. A simpler approach to realize stereo vision by only one camera is to capture two or more images from different veiws at different time by moving the single camera [47, 49]. Another common way is to use catadioptric system such as planar mirrors [53, 54], convex mirrors [67, 68], bi-prism mirrors [61, 62], rotation mirrors and so on. The former is called motion stereo and the later is called catadioptric stereo. Other methods including image layering stereo by using lens aperture, coded aperture or micro lens array are also proposed.

Motion stereo can freely set the baseline width and vergence angle between virtual cameras at different timings for accurate stereo measurement, whereas it is limited to measuring stationary scenes due to the synchronization error caused by the delay time among multiple images. Image-layering stereo requires a decoding process to extract multi-view data from a single image; it is limited in accuracy due to the very narrow baseline width of stereo measurement on their designed apertures. Corresponding to the number of viewpoints, catadioptric stereo has to divide the cameras field of view into

smaller fields for multi-view, whereas it can provide a relatively long baseline width and large vergence angle between mirrored virtual cameras for accurate stereo measurement. However, these stereo systems have not been used as an active stereo to switch quickly and expand the field of view for wide-area surveillance.

1.1.3 High-speed Catadioptric Vision

For both typical stereo vision system and monocular stereo vision systems, the vision device is one of the most importance parts for capturing stereo pairs. many kinds of vision systems have been applied to various fields, such as multimedia, industrial inspection, three-dimensional reconstruction, traffic system, and so on. Most of conventional vision systems with standard video signals are designed on the basis of the characteristics of the human eye, which implies that the processing speed of these systems is limited to the recognition speed of human eye. Therefore, various high-speed vision systems that can operate at sound-level frame rate have been developed for various hyper-human applications.

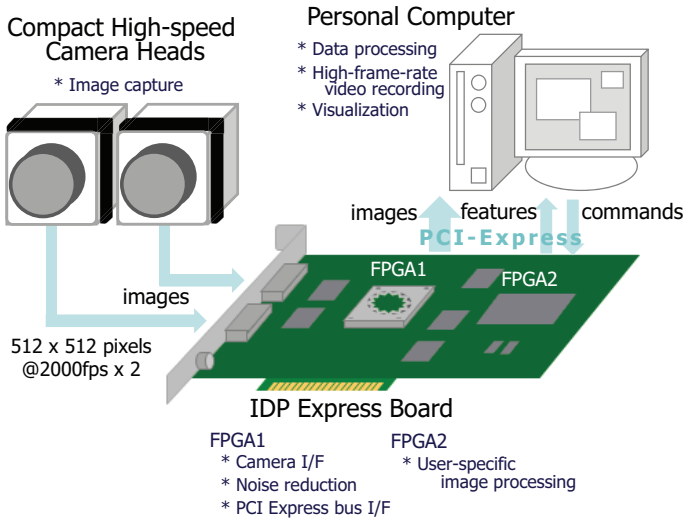
The key issue of high speed vision is the transmission speed from photo-detectors (PD) to processing elements (PE). To accelerate the transmission speed, vision chips or FPGAs have been developed and execute real-time processes at a rate of 1000 fps or more by integrating sensors and processors compactly. Bernard et al. proposed an on-chip array of bare Boolean processors with half toning facilities and developed a 65×76 Boolean retina on a 50 mm^2 CMOS 2 μm circuit for the imager of an artificial retina [69]. Eklund et al. verified the near-sensor image processing (NSIP) concept, which describes a method to implement a two-dimensional (2-D) image sensor array with processing capacity in every pixel, and have fabricated and measured a 32×32 pixels NSIP [70]. Ishikawa et al. have developed a COMS vision chip for 1ms image processing and proposed the S^3PE(simple and smart sensory processing elements) vision chip architecture with each PE connected to a PD without scanning circuits [71, 72]. Komuro et al. proposed a dynamically reconfigurable single-instruction multiple-data (SIMD) processor for a vision chip and developed a prototype vision chip based on their proposed architecture, which

has 64×64 pixels in a 5.4 mm 5.4 mm area fabricated using the 0.35 μm TLM CMOS process [73]. Ishii et al. proposed a new vision chip architecture specialized for target tracking and recognition, and developed a prototype vision chip using 0.35 μm CMOS DLP/TLM(3LM) process [74].

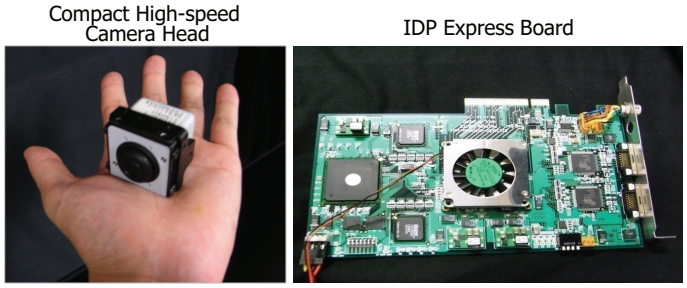
Many real-time HFR-vision systems that can process images at hundreds of frames per second or more have been made by implementing image processing algorithms using parallel processing circuits on a field-programmable gate array (FPGA) board that is directly connected to an HFR camera head. Hirai et al. developed an flexibility FPGA-based vision system using the logic circuit to implement the image algorithm [75]. Watanabe et al. developed a high-speed vision system for real-time shape measurement of a moving/deforming object at a rate of 955 fps (256×256 resolution) [76]. Ishii et al. developed a high-resolution high-speed vision platform, H^3(Hiroshima Hyper Human) Vision, which can simultaneously process a 1024×1024 pixels image at 1000 fps and a 256×256 pixels image at 10000 fps by implementing image processing algorithms as hardware logic on a dedicated FPGA board [77]. In the latest two years, Ishii et al. developed a high-speed vision system called IDP Express, as shown in Figure 1.1., which can execute real-time image processing at a rate from 2000 fps (512×512 resolution) to 10000 fps (512×96 resolution), and high frame rate video recording simultaneously [78].

Some real-time monocular stereo vision systems can be obtained by combining high-speed vision and catadioptric system. For the fixed catadioptric system, the image resolution is reduced as half of original resolution because the wo images are acquired for stereo. For rotating mirror based monocular stereo system, the image resolution is integrated. However, these rotating mirrors move slowly because it is difficult to drive these heavy mirrors. The mirror galvanometer, however, can drive and switch mirrors quickly.

In general, high-speed mirror galvanometers are employed in laser light shows to move the laser beams and produce colorful geometric patterns in fog around the audience. Such high speed mirror galvanometers have proved to be indispensable in industry for laser marking systems for everything from laser etching hand tools, containers, and parts to batch-coding semiconductor wafers in semiconductor device fabrication. nowadays,



(a) configuration of IDP Express vision system



(b) photo of IDP Express vision system

Figure 1.1: IDP Express high-speed imaging system

many systems based on high-speed mirror galvanometers are proposed for object tracking by combining high-speed vision.

1.2 Outline of Thesis

This thesis is organized as 7 Chapters including this introduction.

Chapter 2 summarizes related works on monocular stereo vision using catadioptric system, 3-D motion tracking methods, and wide baseline stereo vision system.

In Chapter 3, I proposed the concept of monocular stereo active vision system based on the high speed vision and ultrafast pan-tilt mirror system for 3-D shape measurement and real-time marker based 3-D motion capturing. Besides, a concept of wide baseline

monocular stereo system is also proposed.

In Chapter 4, monocular stereo measurement system based on high-speed catadioptric device is proposed to track the target and obtain the stereo pairs at the same time.

In Chapter 5, marker based real-time 3-D motion tracking method is introduced by using the monocular catadioptric stereo system.

In Chapter 6, a novel monocular wide baseline stereo measurement system is proposed to realize wide baseline stereo even in limited indoor environment.

Chapter 7, the final chapter, summarizes the contributions of this study and also discusses the future work.

Chapter 2

Related Works

2.1 Monocular Stereo Measurement

Monocular stereo vision system can be classified into (1) motion stereo that calculates range information from multiple images captured at different timings [47–49]; (2) image layering stereo that incorporates multi-view information in a single image via a single-lens aperture [50] and coded aperture [51,52]; and (3) catadioptric stereo for which a single image involves mirror-reflected multi-view data; the camera's field of view is divided either by a single planar mirror [53, 54], two or three planar mirrors [55–57], four planar mirrors [58, 60], bi-prism mirrors [61, 62], or convex mirrors [63–68].

Considering camera calibration and stereo rectification [79, 80], several real-time catadioptric stereo systems have also been developed [81–86]. However, most catadioptric stereo systems have not been used for an active stereo to expand the field of view and baseline distance. This is because catadioptric stereo systems involving large mirrors are too heavy to quickly change their orientations, and it is difficult to control the pan and tilt angles of mirrored virtual cameras independently. Monocular stereo systems that can quickly switch their viewpoints with dynamic changing apertures, such as a programmable iris with a liquid crystal device [87] and multiple pinhole apertures with a rotating slit [88], also have been reported as expansions of an image-layering stereo with designed apertures, whereas they have not considered an active stereo for a wide field of view due to a very narrow baseline stereo measurement.

2.2 Three-Dimensional Motion Tracking

Motion tracking [89, 90] is an important technique in computer vision for capturing and analyzing the movements of objects, and motion capture systems have been used widely in many application areas such as entertainment [91–93], surveillance [94, 95], human interfaces [96, 97], and robotics [98, 99]. Vision-based motion capture systems are mainly categorized as marker-based systems that use passive retroreflective markers [100–102] or LED markers [103, 104], and markerless systems where motion estimation employs object kinematic models [105, 106]. Many real-time marker-based motion capture systems that operate at hundreds or thousands of frames per second (fps) have been already developed in order to rapidly capture the motions of objects in real time [107, 108]. During marker-based motion capture, image processing is employed to extract the regions with markers and calculate their positions in images, which is much less time-consuming than marker-less motion capture.

Most motion capture systems employ multiple cameras at fixed locations to obtain the three-dimensional (3-D) positions of the markers via triangulation using multiple images with different views. However, the control and management of these systems becomes more complex and costly as the number of cameras increases in order to obtain a wider view field without decreasing the resolution especially when multiple moving objects should be independently tracked and captured for accurate stereo measurement. Besides, for multiple target objects it is hard to track and capture their 3D motions independently. 3-D motion capture could be executed to obtain a wider view field without increasing the number of cameras if we can mechanically track moving objects for observation in the view fields of cameras in order to capture different views for triangulation. For monocular stereo measurement, Hu et al. [109] had developed a catadioptric stereo tracking system that can function as a single pair of virtual left and right pan-tilt tracking cameras by switching hundreds of different views in a second, whereas it had been limited in offline monocular stereo measurement for a single moving object.

Many high-speed vision systems have been developed for capturing and processing images in real time at hundreds or more fps in order to track fast-moving objects

[77, 78, 110]. High frame-rate image processing accelerated by field-programmable gate arrays (FPGAs) and graphic processing units has been reported for applications in optical flow [111], multi-object tracking [112], and face tracking [113]. The effectiveness of high-speed vision has been demonstrated in tracking applications such as robot-hand grasping [114], drone tracking [115, 116], cell analysis in microchannels [117, 118], and vibration analysis [119]. However, in contrast to the accelerated sensor and computing based on these tracking systems, the actuator cannot be accelerated sufficiently so the tracking control requires dozens of fps for convergence. Recently, a mirror-drive active vision system [120] was developed that uses galvanomirrors with accelerated pan-tilt actuators to achieve ultrafast gaze control for tracking fast-moving objects, and it can function as a virtual multi-pan and tilt camera [121] to observe different views in one second. The installation and management costs would be reduced greatly in many 3-D tracking applications with a wider view of field if a single ultrafast tracking system can simultaneously track the same object in multiple images with different views for stereo measurement instead of multi-camera stereo.

2.3 Wide Baseline Stereo Vision System

Stereo vision techniques have been widely used in many robotic vision applications to extract depth information. Classical stereo vision system which is made up of two parallel placed cameras with short baseline has been studied widely [122]. Unlike short baseline stereo techniques, wide baseline stereo can yield more accurate depth estimates and tolerate a large change in viewpoint between the images [123, 124]. Most studies about wide baseline stereo have mainly focused on epipolar geometry [125], stereo matching algorithms [126] and synchronization between cameras [127]. However, these studies are almost based on two or more real-camera systems, which have the problems of connecting and adjusting two cameras and are impossible to make the wide baseline stereo system in limited space.

Monocular stereo system in contrast has the advantages of easier setting and controlling the baseline, because that of the virtual cameras. Many monocular stereo systems

have been presented such as motion stereo, image layering stereo, catadioptric stereo and so on. However, the baselines of these systems are almost short distance and difficult to adjust. Hu et al. [109] had summarized these different kinds of monocular stereo measurement systems and proposed an adjustable baseline monocular stereo system for both stereo measurement and multiple targets stereo tracking [128]. However, drawbacks such as limited baseline and additional illumination still remain. We designed the new catadioptric system to make the stereo baseline wider and make it possible to work without additional illumination.

Chapter 3

Concept

3.1 Active Monocular Stereo Measurement

This section describes our concept of catadioptric stereo tracking on an ultrafast mirror-drive active vision system that can perform as two virtual pan-tilt cameras for left and right-side views by frame-by-frame switching the direction of the mirrors on the active vision system. Figure 3.1 shows the concept of catadioptric stereo tracking. Our catadioptric stereo tracking system consists of a mirror-based ultrafast active vision system and a catadioptric mirror system. The former consists of a high-speed vision system that can capture and process images in real time at a high frame rate, and a pan-tilt mirror system for ultrafast gaze control. It can be unified as an integrated pan-tilt camera and its complexity in system management is similar to those of standard PTZ cameras, which are commonly used in video surveillance applications. Figure 3.1 shows a catadioptric mirror system consisting of multiple planar mirrors on the left and right sides, and a pan-tilt mirror system installed in front of the lens of the high-speed vision system to switch between left- and right-side views by alternating the direction of its mirrors. The images on the side of the left-side mirror and the left half of the angle mirror are captured as the left-view images, and those on the side of the right-side mirror and the right half of the angle mirror are captured as the right-view images.

Originating from multithreaded processing in which threads conducting tasks are simultaneously running on a computer using the time-sharing approach, our catadioptric stereo tracking method extends the concept of multithread gaze control to the pan-tilt

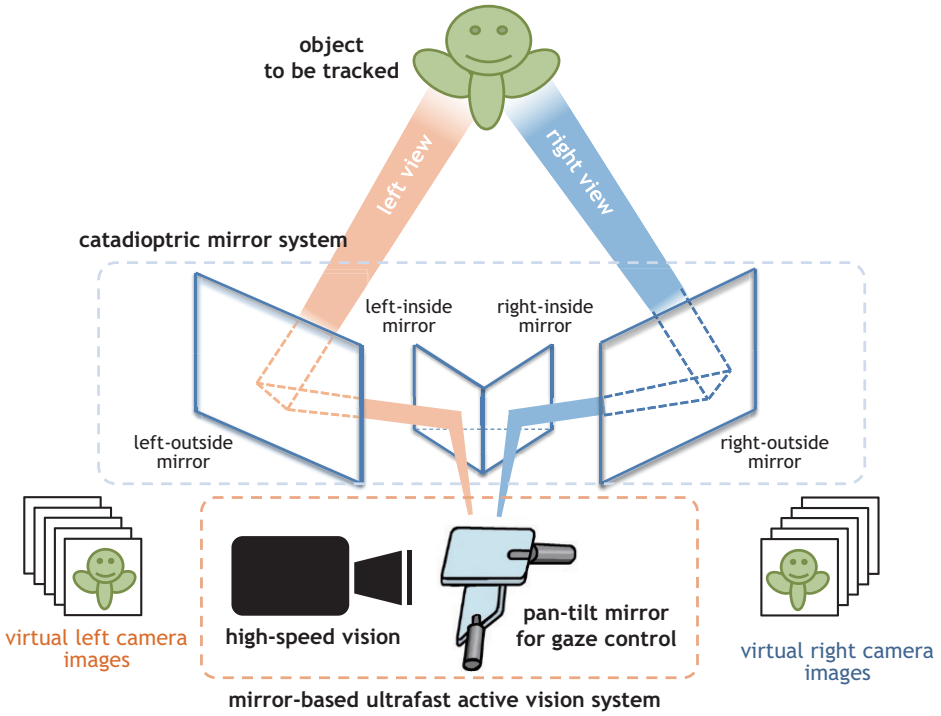


Figure 3.1: Catadioptric stereo tracking with multithread gaze control.

camera by parallelizing a series of operation with video-shooting, processing, and gaze control into time-division thread processes with a fine temporal granularity to realize multiple virtual pan-tilt cameras on a single active vision system. The following conditions are required so that a single active vision system with multithread gaze control has a potency equivalent to that of left and right pan-tilt cameras for accurate and high-speed stereo tracking with sufficient large parallax.

(1) Acceleration of video-shooting and processing

When left and right virtual pan-tilt cameras are shooting at the rate of dozens or hundreds of frames per second for tracking fast-moving objects in 3D scenes, the frame capturing and processing rate of an actual single vision system must be accelerated at a rate of several hundreds or thousands of frames per second to perform the video-shooting, processing, and tracking for left and right-view images of the virtual pan-tilt cameras.

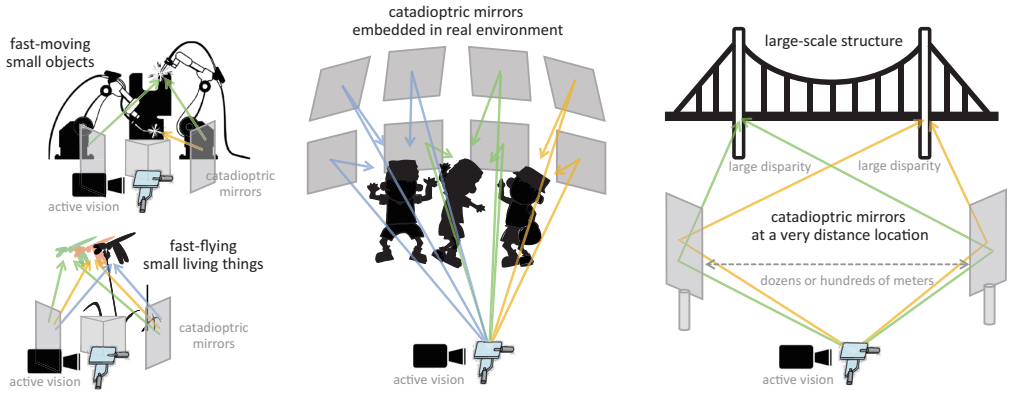
(2) Acceleration of gaze control

To control the gaze of every frame independently, high-speed gaze control must en-

sure that a given frame does not affect the next frame. Corresponding to the acceleration of video-shooting and processing at a rate of several hundreds of frames per second, the temporal granularity of the time-division thread gaze control processes must be minimized at the millisecond level, and a high-speed actuator that has a frequency characteristic of a few kHz is required for acceleration of gaze control.

Compared with catadioptric systems with a fixed camera, catadioptric stereo tracking has the advantage of being able to mechanically track a target object as active stereo while zooming in the fields of views of virtual left and right pan-tilt cameras; multithread gaze control enables zoom-in tracking when the target is moving in the depth direction by controlling the vergence angle between two virtual pan-tilt cameras as well as when the target moves in the left-right or up-down direction. In catadioptric stereo tracking, correspondences among left and right-view images can be easily established because their camera internal parameters, such as focal length, gain, and exposure time, are the same in virtual left and right pan-tilt cameras, whose lens and image sensors are perfectly matched due to differences in their cameras' internal parameters.

The catadioptric mirror system in catadioptric stereo tracking can be designed flexibly in accordance with the requirements of its practical applications. Moreover, catadioptric stereo tracking has the following advantages over active stereo systems with multiple PTZ cameras: (1) space-saving installation that enables stereo measurement in a small space, where multiple PTZ cameras cannot be installed; (2) easy expandability for multi-view stereo measurement with a large number of mirrors; and (3) stereo measurement with arbitrary disparity without any electrical connection that enables precise long-distance 3D sensing. Figure 3.2 illustrates the configuration examples of the catadioptric mirror systems, referring to the practical applications of catadioptric stereo tracking: (a) precise 3D digital archiving/video logging for fast-moving small objects and creatures; (b) 3D human tracking without a dead angle, which functions as a large number of virtual pan-tilt cameras by utilizing multiple mirrors embedded in the real environment; and (c) remote surveillance for a large-scale structure with left and right mirrors at a distant location that requires a large disparity of dozens or hundreds of meters for precise 3D sensing. In this study, the catadioptric mirror system used in the experiments detailed in



(a) precise 3-D digital archiving / 3-D video logging

(b) 3-D human tracking without dead angle

(c) remote monitoring for a large-scale structure

Figure 3.2: Configuration examples of catadioptric mirror systems, referring to practical applications of catadioptric stereo tracking: (a) precise 3-D digital archiving or 3-D video logging; (b) 3-D human tracking without dead angle; (c) remote monitoring for a large-scale structure.

Section 4.4 was set up for a short-distance measurement to verify the performance of our catadioptric stereo tracking system in a desktop environment, corresponding to precise 3D digital archiving for fast-moving small objects.

Catadioptric stereo tracking has disadvantages: (1) inefficient use of incident light, owing to the small-size pan-tilt mirror, which is designed for ultrafast switching of viewpoints; and (2) synchronization errors in stereo measurement of moving targets, due to the delay time between virtual left and right-view images captured at different timings. These synchronization errors in catadioptric stereo tracking can be reduced by accelerating alternative switching of left and right views with multithreaded gaze control with a temporal granularity at the millisecond level. The pros and cons of the catadioptric systems with fixed mirrors, fixed camera systems, and our catadioptric stereo tracking system are summarized in Table 3.1.

3.2 Real-time Monocular Three-Dimensional Multiple Targets Motion Tracking

In this study, we extended a real-time monocular 3-D tracking system for multiple moving objects based on a catadioptric stereo tracking system with multithread active

Table 3.1: Pros and cons of catadioptric stereo systems, fixed camera systems, and catadioptric stereo tracking system.

	Catadioptric Systems	**Fixed Camera Systems**	**Catadioptric Stereo Tracking System**
classification	planar mirror [53–58, 60, 81] bi-prism mirror [61, 62] convex mirror [63–68, 82–86]	lens aperture based [50, 87, 88], coded aperture based [51, 52]	our proposed method
advantages	multi-view/wide field of view (convex)/no synchronization error/single camera	compact structure/rapid viewpoint-switching/easy calibration/single camera	active stereo/full image resolution/multi-view tracking/single camera
disadvantages	image distortion (convex)/half image resolution (planar or bi-prism)/inactive stereo	narrow baseline/limited field of view/synchronization errors/inactive stereo	insufficient incident light/synchronization errors/complex stereo calibration.

vision, which we proposed as an offline monocular stereo measurement method for a single moving object in a previous study [109]. Figure 3.3 illustrates the concept of multiple virtual stereo tracking cameras employed by this system.

We use an ultrafast active vision system comprising a high-speed vision system for image acquisition and processing at a high frame rate, a pan-tilt galvanomirror system for ultrafast gaze control, and a catadioptric mirror system with a pair of planar mirrors on the left and right sides. This system can serve as virtual left and right pan-tilt tracking cameras to capture the same scene from different views for triangulation by frame-by-frame viewpoint switching in the pan and tilt directions. By accelerating gaze control as well as video-shooting and processing, multithread gaze control parallelizes a series of operation comprising video shooting, processing, and gaze control into time-division processes with a fine temporal granularity, thereby providing multiple pairs of left and right pan-tilt tracking cameras with different views via a catadioptric mirror system on a single active vision system for monocular stereo measurement for multiple moving objects.

Compared with catadioptric systems with a fixed camera, catadioptric stereo tracking has the advantage of obtaining accurate stereo measurements by tracking multiple target objects as active stereo while zooming in on the fields of views observed by multiple pairs of virtual left and right pan-tilt cameras. However, catadioptric stereo tracking has the disadvantage of synchronization errors when obtaining stereo measurements of

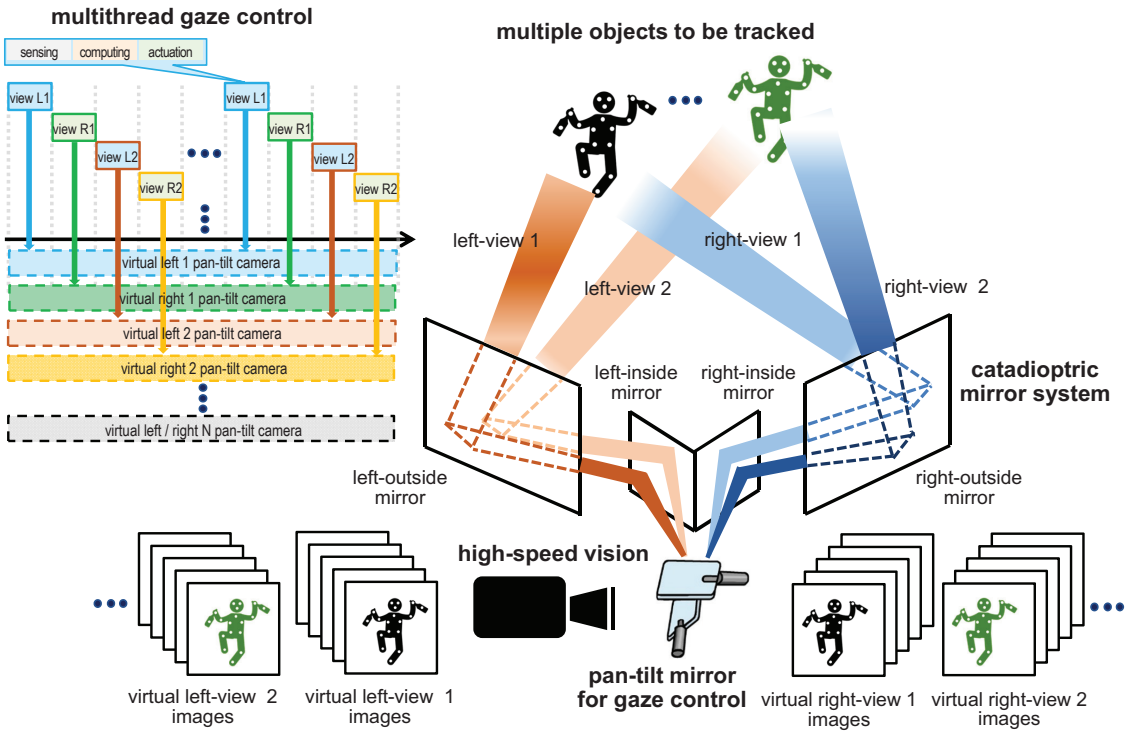


Figure 3.3: Catadioptric 3-D motion tracking with multithread gaze control for multiple virtual stereo tracking cameras.

moving target objects due to the time difference when capturing the virtual left- and right-view images.

3.3 Monocular Wide Baseline Stereo Vision System

The concept of proposed wide baseline monocular catadioptric stereo system is based on an ultrafast pan-and-tilt mirror device that can switch hundreds of different views in one second to make two virtual pan-tilt cameras for stereo pairs capturing. Figure 3.4 illustrates the concept of proposed monocular wide baseline stereo vision system including a high-speed vision system, pan-tilt mirror system for ultrafast gaze control, and a catadioptric mirror system consisting two plane mirrors. The pan-tilt mirror device is installed in front of the lens of the high-speed vision system to switch left and right-side view directions to the two side-mirrors alternately. Two virtual cameras with wider baseline can be obtained and images of each virtual camera are captured frame by frame

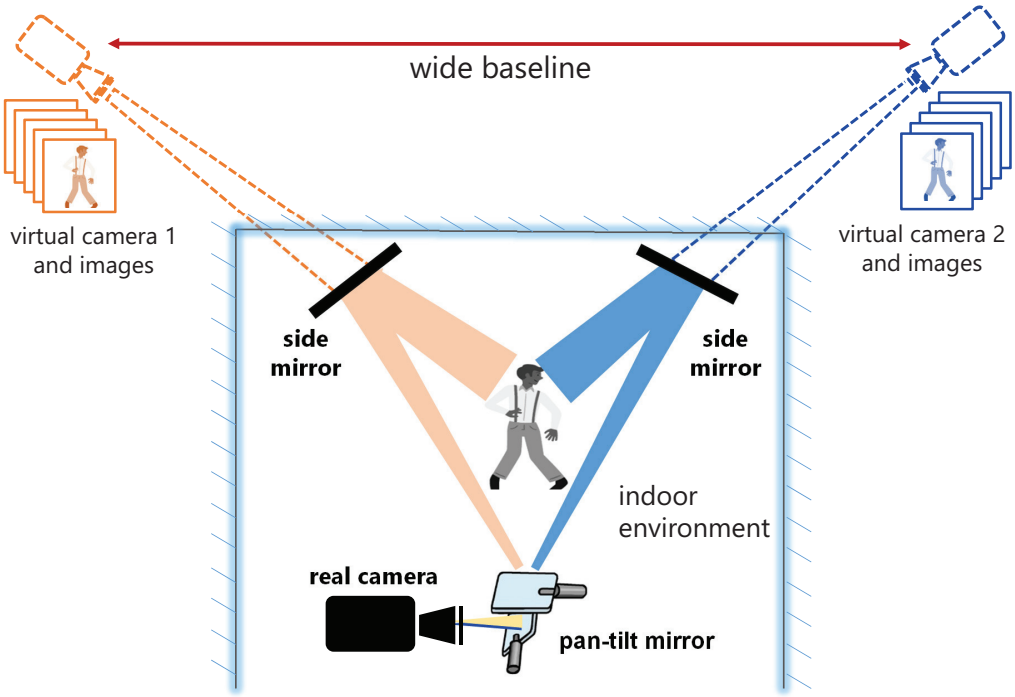


Figure 3.4: Concept of proposed monocular wide baseline stereo measurement system.

alternately according to the multithread time division control of pan and tilt mirrors.

Chapter 4

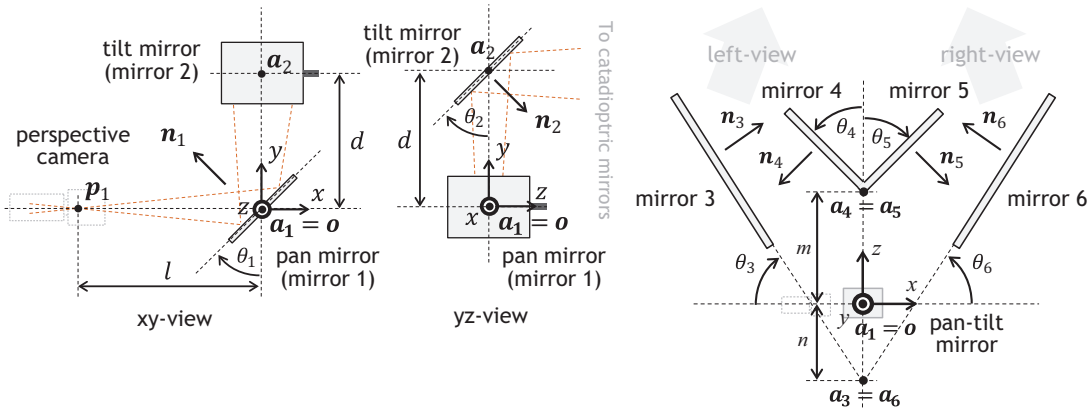
Monocular Stereo Measurement Using High-Speed Catadioptric Tracking

4.1 Introduction

In this chapter, we implement a monocular stereo tracking system that expands on a concept of catadioptric stereo with a relatively long-width baseline to an active stereo that can control the pan and tilt directions of mirrored virtual cameras for the wider field of view, and develop a mirror-based ultrafast active vision system with a catadioptric mirror system that enables a frame-by-frame viewpoint switching of pan and tilt controls of mirrored virtual tracking cameras at hundreds of frames per second.

4.2 Geometry of Catadioptric Stereo Tracking

This section describes the geometry of a catadioptric stereo tracking system that uses a pan-tilt mirror system with a single perspective camera and a catadioptric mirror system with four planar mirrors as illustrated in Figure 3.1, and derives the locations and orientations of virtual left and right pan-tilt cameras for triangulation in active stereo for time-varying 3D scenes.



(**a**) pan-tilt mirror (**b**) catadioptric mirror

Figure 4.1: Geometries of the pan-tilt mirror system and the catadioptric mirror system: (a) pan-tilt mirror; (b) catadioptric mirror.

4.2.1 Geometrical Definitions

4.2.1.1 Pan-Tilt Mirror System

The pan-tilt mirror system assumed in this study has two movable mirrors in the pan and tilt directions: pan mirror and tilt mirror. Figure 4.1a shows the xy-view and yz-view of the geometrical configuration of the pan-tilt mirror system with a perspective camera; the xyz-coordinate system is set so that the x-axis corresponds to the optical axis of the camera, the y-axis corresponds to the line between the center points of the pan mirror and tilt mirror. The depth direction in stereo measurement corresponds to the z-direction. The center of the pan mirror (mirror 1) is set to $\boldsymbol{a}_1 = (0,0,0)^T$, which is the origin of the xyz-coordinate system. The pan mirror can rotate around the z-axis, and its normal vector is given as $\boldsymbol{n}_1 = (-\cos\theta_1, \sin\theta_1, 0)^T$. The center of the tilt mirror (mirror 2) is located at $\boldsymbol{a}_2 = (0,d,0)^T$, where its distance from that of the pan mirror is represented by d. The tilt mirror can rotate around a straight line parallel to the x-axis at a distance d, and its normal vector is given as $\boldsymbol{n}_2 = (0,-\sin\theta_2, \cos\theta_2)^T$. θ_1 and θ_2 indicate the pan and tilt angles of the pan-tilt mirror system, respectively. The optical center of the perspective camera is set to $\boldsymbol{p}_1 = (-l,0,0)^T$, where its distance from the center of the pan mirror is represented by l.

4.2.1.2 Catadioptric Mirror System

The catadioptric mirror system with four planar mirrors is installed in front of the pan-tilt mirror so that all the planar mirrors are parallel to the y-axis. Figure 4.1b shows the xz-view of its geometry. The locations of mirrors 3 and 4 for the left-side view and mirrors 5 and 6 for the right-side view are given. The normal vectors of the mirror planes $i(= 3, 4, 5, 6)$ are given as $\boldsymbol{n}_3 = (\sin\theta_3, 0, \cos\theta_3)^T$, $\boldsymbol{n}_4 = (-\cos\theta_4, 0, -\sin\theta_4)^T$, $\boldsymbol{n}_5 = (\cos\theta_5, 0, -\sin\theta_5)^T$, and $\boldsymbol{n}_6 = (-\sin\theta_6, 0, \cos\theta_6)^T$, respectively. As illustrated in Figure 4.1b, θ_3 and θ_6 are the angles formed by the xy-plane and the planes of mirrors 3 and 6, respectively; θ_4 and θ_5 are those formed by the yz-plane and the planes of mirrors 4 and 5, respectively. A pair of mirrors 3 and 6 at the outside are located symmetrically with the yz-plane as well as a pair of mirrors 4 and 5 at the inside; $\theta_3 = \theta_6$ and $\theta_4 = \theta_5$. The planes of mirrors 4 and 5, and those of mirrors 3 and 6 are crossed on the yz-plane, respectively; their crossed lines pass through the points $\boldsymbol{a}_4 = (0, d, m)(= \boldsymbol{a}_5)$ and $\boldsymbol{a}_3 = (0, d, n)(= \boldsymbol{a}_6)$ in front of or behind the center of the tilt mirror, respectively.

4.2.2 Camera Parameters of Virtual Pan-Tilt Cameras

4.2.2.1 Mirror Reflection

The camera parameters of a virtual pan-tilt camera, whose optical path is reflected on a pan-tilt mirror system and a catadioptric mirror system multiple times, can be described by considering the relationship between the real camera and its virtual camera, reflected by a planar mirror as illustrated in Figure 4.2. The optical center of the real camera is given as $\boldsymbol{p}_i$, and it is assumed that the mirror plane, whose normal vector is $\boldsymbol{n}_i$, involves the point $\boldsymbol{a}_i$. The optical center of the mirrored virtual camera $\boldsymbol{p}_{i+1}$, which is the reflection of the real camera view on the mirror plane, can be expressed with a 4×4 homogeneous transformation matrix $\boldsymbol{P}_i$ as follows:

$$\begin{pmatrix} \boldsymbol{p}_{i+1} \\ 1 \end{pmatrix} = \boldsymbol{P}_i \begin{pmatrix} \boldsymbol{p}_i \\ 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} - 2\boldsymbol{n}_i\boldsymbol{n}_i^T & 2(\boldsymbol{n}_i\boldsymbol{n}_i^T)\boldsymbol{a}_i \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{p}_i \\ 1 \end{pmatrix}, \tag{4.1}$$

where $\boldsymbol{I}$ is the 3×3 identity matrix.

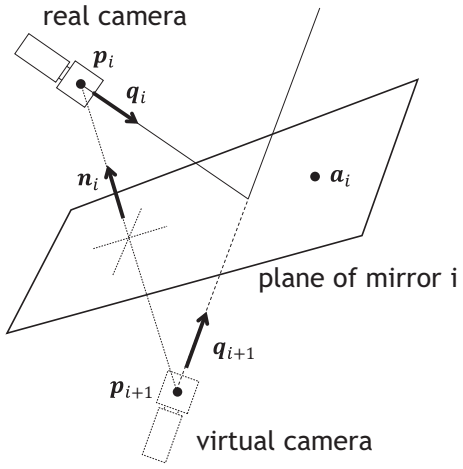


Figure 4.2: Relationship between real camera and its virtual camera reflected by a planar mirror.

4.2.2.2 Pan-Tilt Mirror System

Using the geometric parameters of the pan-tilt mirror system as defined in Section 4.2.1.1, the optical center of the virtual camera $\boldsymbol{p}_{pt}$, which is reflected by its pan and tilt mirrors, can be expressed with reflection transformation as follows:

$$\begin{pmatrix} \boldsymbol{p}_{pt} \\ 1 \end{pmatrix} = \boldsymbol{P}_2\boldsymbol{P}_1 \begin{pmatrix} \boldsymbol{p}_1 \\ 1 \end{pmatrix} = \boldsymbol{P}_2\boldsymbol{P}_1 \begin{pmatrix} -l \\ 0 \\ 0 \\ 1 \end{pmatrix}, \tag{4.2}$$

where

$$\boldsymbol{P}_1 = \begin{pmatrix} -\cos 2\theta_1 & \sin 2\theta_1 & 0 & 0 \\ \sin 2\theta_1 & \cos 2\theta_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{P}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos 2\theta_2 & \sin 2\theta_2 & d(1-\cos 2\theta_2) \\ 0 & \sin 2\theta_2 & -\cos 2\theta_2 & -d\sin 2\theta_2 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.3}$$

Considering $\boldsymbol{q}_1 = (1,0,0)^T$, the optical center of the virtual camera $\boldsymbol{p}_{pt}$ and the direction of its optical axis $\boldsymbol{q}_{pt}$ can be derived from Equations (4.2) and (4.3), as the

following functions of the pan and tilt angles θ_1 and θ_2,

$$\boldsymbol{p}_{pt}(\theta_1,\theta_2)=\begin{pmatrix} l\cos 2\theta_1 \\ -(l\sin 2\theta_1+d)\cos 2\theta_2+d \\ -(l\sin 2\theta_1+d)\sin 2\theta_2 \end{pmatrix},\quad \boldsymbol{q}_{pt}(\theta_1,\theta_2)=\begin{pmatrix} -\cos 2\theta_1 \\ \sin 2\theta_1\cos 2\theta_2 \\ \sin 2\theta_1\sin 2\theta_2 \end{pmatrix}. \quad (4.4)$$

4.2.2.3 Catadioptric Mirror System

In the catadioptric mirror system, the pan angle of the pan-tilt mirror system determines whether the camera gazes the left view via mirrors 3 and 4 or the right view via mirror 5 and 6.

When the virtual pan-tilt camera gazes the left view, the optical center of the virtual pan-tilt camera after the mirror reflections of the catadioptric mirror system, $\boldsymbol{p}_L$, can be expressed by using its geometric parameters described in Section 4.2.1.2 as follows:

$$\begin{pmatrix} \boldsymbol{p}_L \\ 1 \end{pmatrix}=\boldsymbol{P}_3\boldsymbol{P}_4\begin{pmatrix} \boldsymbol{p}_{pt} \\ 1 \end{pmatrix}=\boldsymbol{P}_3\boldsymbol{P}_4\boldsymbol{P}_2\boldsymbol{P}_1\begin{pmatrix} \boldsymbol{p}_1 \\ 1 \end{pmatrix}, \quad (4.5)$$

where

$$\boldsymbol{P}_3=\begin{pmatrix} \cos 2\theta_3 & 0 & -\sin 2\theta_3 & n\sin 2\theta_3 \\ 0 & 1 & 0 & 0 \\ -\sin 2\theta_3 & 0 & -\cos 2\theta_3 & n(1+\cos 2\theta_3) \\ 0 & 0 & 0 & 1 \end{pmatrix},\ \boldsymbol{P}_4=\begin{pmatrix} -\cos 2\theta_4 & 0 & -\sin 2\theta_4 & m\sin 2\theta_4 \\ 0 & 1 & 0 & 0 \\ -\sin 2\theta_4 & 0 & \cos 2\theta_4 & m(1-\cos 2\theta_4) \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (4.6)$$

Thus, the optical center of the virtual left pan-tilt camera $\boldsymbol{p}_L$ and the direction of its optical axis $\boldsymbol{q}_L$ can be derived from Equations (4.4) and (4.6) as follows:

$$\boldsymbol{p}_L=\begin{pmatrix} -C_{34}l\cos 2\theta_1+S_{34}(l\sin 2\theta_1+d)\sin 2\theta_2+E \\ -(l\sin 2\theta_1+d)\cos 2\theta_2+d \\ S_{34}l\cos 2\theta_1+C_{34}(l\sin 2\theta_1+d)\sin 2\theta_2+F \end{pmatrix},\quad \boldsymbol{q}_L=\begin{pmatrix} C_{34}\cos 2\theta_1-S_{34}\sin 2\theta_1\sin 2\theta_2 \\ \sin 2\theta_1\cos 2\theta_2 \\ -S_{34}\cos 2\theta_1-S_{34}\sin 2\theta_1\sin 2\theta_2 \end{pmatrix} \quad (4.7)$$

where C_{34}, S_{34}, E, and F are constants, which are determined by the parameters of the

catadioptric mirror system as follows:

$$C_{34} = \cos 2(\theta_3 + \theta_4), \quad S_{34} = \sin 2(\theta_3 + \theta_4), \tag{4.8}$$

$$E = m(-\sin 2\theta_3 + S_{34}) + n \sin 2\theta_3, \quad F = -m(\cos 2\theta_3 - C_{34}) + n(1 + \cos 2\theta_3) \tag{4.9}$$

In a similar manner as the left view via mirrors 3 and 4, the optical center of the virtual pan-tilt camera when the virtual pan-tilt camera gazes the right view via mirrors 5 and 6, $\boldsymbol{p}_R$, can be expressed by as follows:

$$\begin{pmatrix} \boldsymbol{p}_R \\ 1 \end{pmatrix} = \boldsymbol{P}_6 \boldsymbol{P}_5 \begin{pmatrix} \boldsymbol{p}_{pt} \\ 1 \end{pmatrix} = \boldsymbol{P}_6 \boldsymbol{P}_5 \boldsymbol{P}_2 \boldsymbol{P}_1 \begin{pmatrix} \boldsymbol{p}_1 \\ 1 \end{pmatrix}, \tag{4.10}$$

where

$$\boldsymbol{P}_5 = \begin{pmatrix} -\cos 2\theta_5 & 0 & \sin 2\theta_5 & -m \sin 2\theta_5 \\ 0 & 1 & 0 & 0 \\ \sin 2\theta_5 & 0 & \cos 2\theta_5 & m(1-\cos 2\theta_5) \\ 0 & 0 & 0 & 1 \end{pmatrix}, \boldsymbol{P}_6 = \begin{pmatrix} \cos 2\theta_6 & 0 & \sin 2\theta_6 & -n \sin 2\theta_6 \\ 0 & 1 & 0 & 0 \\ \sin 2\theta_6 & 0 & -\cos 2\theta_6 & n(1+\cos 2\theta_6) \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.11}$$

Considering that the mirrors are symmetrically located with the yz-plane ($\theta_6 = \theta_3$, $\theta_5 = \theta_4$), the optical center of the virtual right pan-tilt camera $\boldsymbol{p}_R$ and the direction of its optical axis $\boldsymbol{q}_R$ can be derived as follows:

$$\boldsymbol{p}_R = \begin{pmatrix} -C_{34} l \cos 2\theta_1 - S_{34}(l \sin 2\theta_1 + d) \sin 2\theta_2 - E \\ -(l \sin 2\theta_1 + d) \cos 2\theta_2 + d \\ -S_{34} l \cos 2\theta_1 + C_{34}(l \sin 2\theta_1 + d) \sin 2\theta_2 + F \end{pmatrix}, \boldsymbol{q}_R = \begin{pmatrix} C_{34} \cos 2\theta_1 + S_{34} \sin 2\theta_1 \sin 2\theta_2 \\ \sin 2\theta_1 \cos 2\theta_2 \\ S_{34} \cos 2\theta_1 - C_{34} \sin 2\theta_1 \sin 2\theta_2 \end{pmatrix} \tag{4.12}$$

In our catadioptric stereo tracking, the optical centers and the directions of the optical axes of the virtual left and right pan-tilt cameras are controlled so that the apparent target positions on their image sensor planes, $\boldsymbol{u}_L = (u_L, v_L)$ and $\boldsymbol{u}_R = (u_R, v_R)$, are tracked in the view fields of the virtual left and right pan-tilt cameras, respectively.

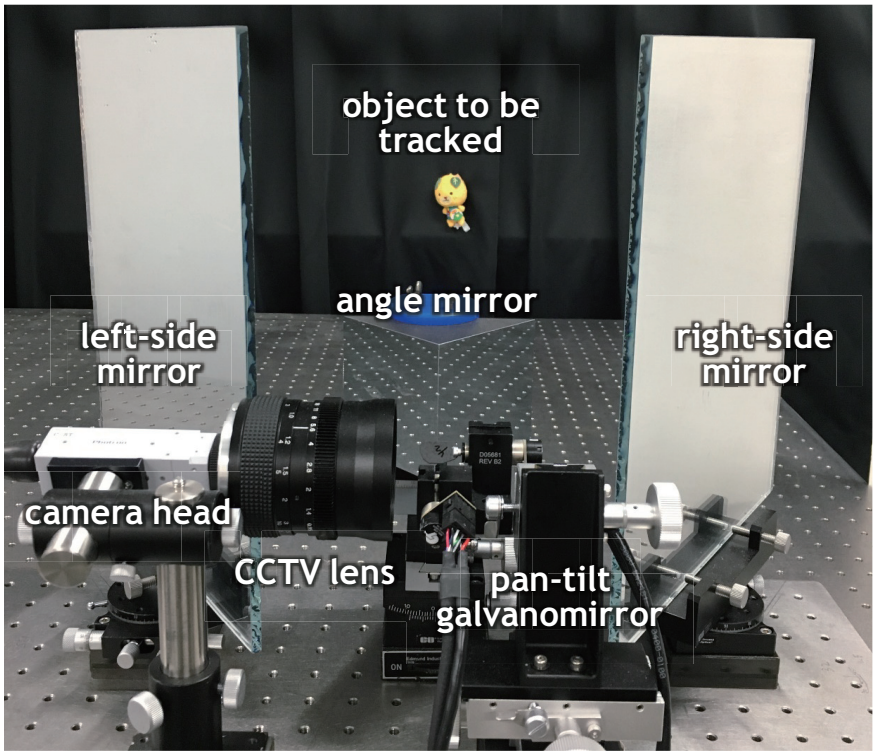


Figure 4.3: Overview of catadioptric stereo tracking system.

4.3 Monocular Catadioptric Stereo Tracking System

4.3.1 System Configuration

We developed a catadioptric stereo tracking system, designed for multithreaded gaze control for switching left and right viewpoints frame-by-frame to capture a pair of stereo images for fast-moving objects in 3D scenes. The system consists of a high-speed vision platform (IDP Express) [78], a pan-tilt galvano-mirror (6210H, Cambridge Technology Inc., Bedford, MA, USA), a right-angle mirror and two flat mirrors on the left and right sides, and a personal computer (PC) (Windows 7 Enterprise 64-bit OS (Microsoft, Redmond, WA, USA); ASUS P6T7 WS SuperComputer motherboard (ASUS, Taiwan, China); Intel Core (TM) i7 3.20-GHz CPU, 6 GB memory (Intel, Santa Clara, CA, USA)). A D/A board (PEX-340416, Interface Inc., Hiroshima, Japan) is used to send control signals to the galvano-mirror and an A/D board (PEX-321216, Interface Inc., Hiroshima, Japan) is used to collect the sensor signals of the pan and tilt angles of the galvano-mirror. Figure 4.3 provides an overview of our developed catadioptric stereo tracking system.

The high-speed vision platform IDP Express (R2000, Photron, Tokyo, Japan) consists of a compact camera head and an FPGA image processing board (IDP Express board). The camera head has a Complementary Metal Oxide Semiconductor (CMOS) image sensor (C-MOS, Photron, Tokyo, Japan) of 512×512 pixels, with a sensor size and pixel size of 5.12×5.12 mm and 10×10 m, respectively. The camera head can capture 8-bit RGB (Red, Green, Blue) images of 512×512 pixels at 2000 fps with a Bayer filter on its image sensor. A $f = 50$ mm CCTV (Closed Circuit Television) lens is attached to the camera head. The IDP Express board was designed for high-speed video processing and recording, and image processing algorithms can be hardware-implemented on the FPGA (Xilinx XC3S5000, Xilinx Inc., San Jose, CA, USA). The 8-bit color 512×512 images and processed results are simultaneously transferred at 2000 fps from the IDP Express board via the Peripheral Component Interconnect (PCI)-e 2.0×16 bus to the allocated memory in the PC.

The galvano-mirror can control two-degrees-of-freedom (DOF) gazes using pan and tilt mirrors, whose sizes are 10.2 mm^2 and 17.5 mm^2, respectively. By applying a voltage signal via the D/A board, the angles of the pan and tilt mirrors are movable in the range of -10 to 10 degrees, and they can be controlled within 1 ms in the range of 10 degrees. The pan mirror of the galvano-mirror was installed 25 mm in front of the CCTV lens, and the tilt mirror was installed 10 mm in front of the pan mirror. A right-angle mirror, whose lengths of the hypotenuse and legs are 106.1 mm and 75.0 mm, respectively, and two 250×310 mm flat mirrors were symmetrically located in front of the galvano-mirror. The catadioptric mirror system was set for short-distance stereo measurement to verify the performance of our catadioptric stereo tracking system in a desktop environment. This enabled us to easily manage the lighting condition to cope with insufficient incident light, owing to the small pan-tilt mirror, and to quantitatively moving at apparently high speeds in images using a linear slider. The configuration of a catadioptric mirror system consisting of these mirrors is illustrated in Figure 4.4. The right-angle mirror was installed in front of the tilt mirror of the galvano-mirror, so that the optical axis of the CCTV lens, which was reflected on the galvano-mirror, comes to the middle point of the right-angled side of the right-angle mirror when the pan and

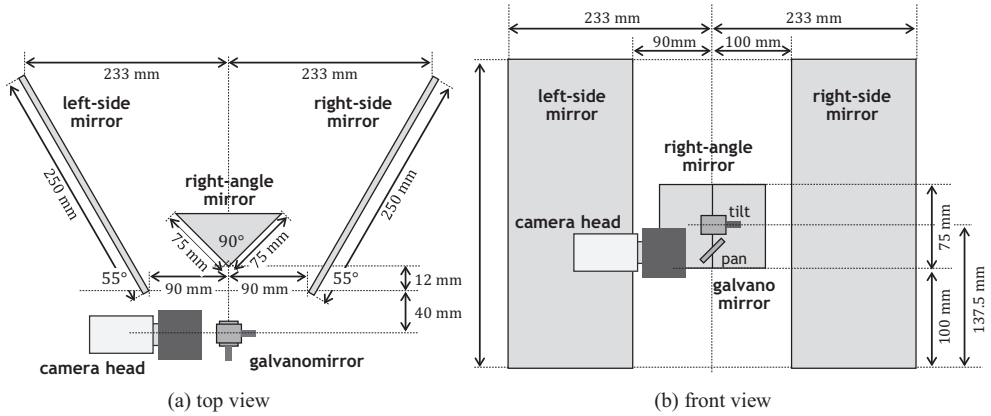


Figure 4.4: Arrangement of mirror system: (a) top view; (b) front view.

tilt angles of the galvano-mirror were zero, corresponding to the center angles in their movable ranges. On the left and right sides, two flat mirrors were vertically installed with an angle of 55 degrees.

The right angle mirror and the two side mirrors can cover the view angle of the camera view completely when the pan and tilt angles of the galvano-mirror were in the range of -10 to 10 degrees, whereas the camera view involved the left-side and right-side views, which were split by the right-angle mirror, when the pan angle of the galvano-mirror was around zero; the camera view only involved the left-view image via the left-side mirror or the right-view image via the right-side mirror, when the pan angle was in the range of -10 to -2 degrees or in the range of 2 to 10 degrees, respectively. It is assumed that the reference positions of the virtual left and right pan-tilt cameras were set when the pan and tilt angles of the galvano-mirror were -5 and 0 degrees, and 5 and 0 degrees, respectively. Figure 4.5 illustrates the locations of the virtual left and right pan-tilt cameras. The optical centers and the normal direction vectors of the optical axes of the virtual left and right cameras at their reference positions were $(-152 \text{ mm}, 10 \text{ mm}, -105 \text{ mm})$ and $(0.174, 0.000, 0.985)$, and $(152 \text{ mm}, 10 \text{ mm}, -105 \text{ mm})$ and $(-0.174, 0.000, 0.985)$, respectively; the *xyz*-coordinate system was set so that its origin was set to the center of the pan mirror of the galvano-mirror as illustrated in Figure 4.5. The virtual left and right pan-tilt cameras can change their virtual pan and tilt angles around their reference positions in the range of -10 to 10 degrees and in the range of -20 to 20 degrees, respec-

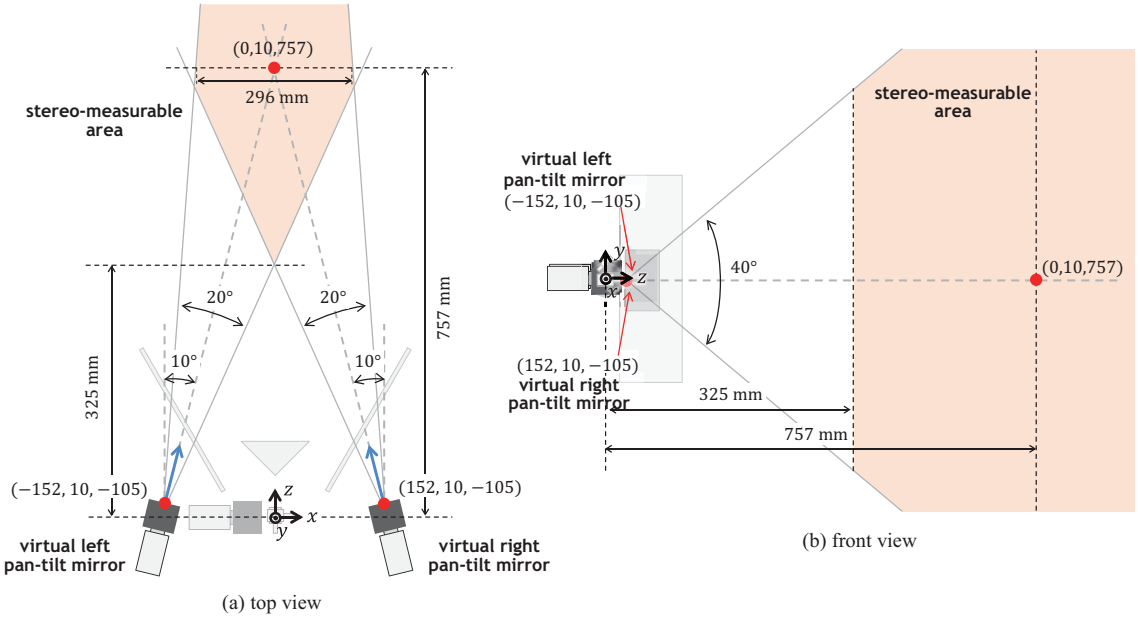


Figure 4.5: Virtual pan-tilt cameras and stereo-measurable area: (a) top view; (b) front view.

tively, which corresponded to twice of the movable ranges of the pan and tilt angles of the galvano-mirror for each virtual pan-tilt camera, whereas the view angle of the camera view was 8.28 degrees in both the pan and tilt directions, respectively, which were determined by the focal distance $f = 50$ mm of the CCTV lens and the 5.12×5.12 mm size of the image sensor.

The stereo-measurable area where a pair of left-view and right-view images can be captured is also illustrated in Figure 4.5. The stereo-measurable area is 296×657 mm on a vertical plane 757 mm in front of the tilt mirror of the galvano-mirror, on which the optical axes of two virtual pan-tilt cameras at their reference positions were crossed, whereas the left-view and right-view images of 512×512 pixels corresponded to 94×94 mm on the vertical plane. When the virtual left and right pan-tilt cameras were at their reference positions, the error in stereo measurement of a nonmoving object 757 mm in front of the tilt mirror of the galvano-mirror, was ± 0.10 mm in the x-direction, ± 0.05 mm in the y-direction, and ± 0.2 mm in the z-direction, respectively.

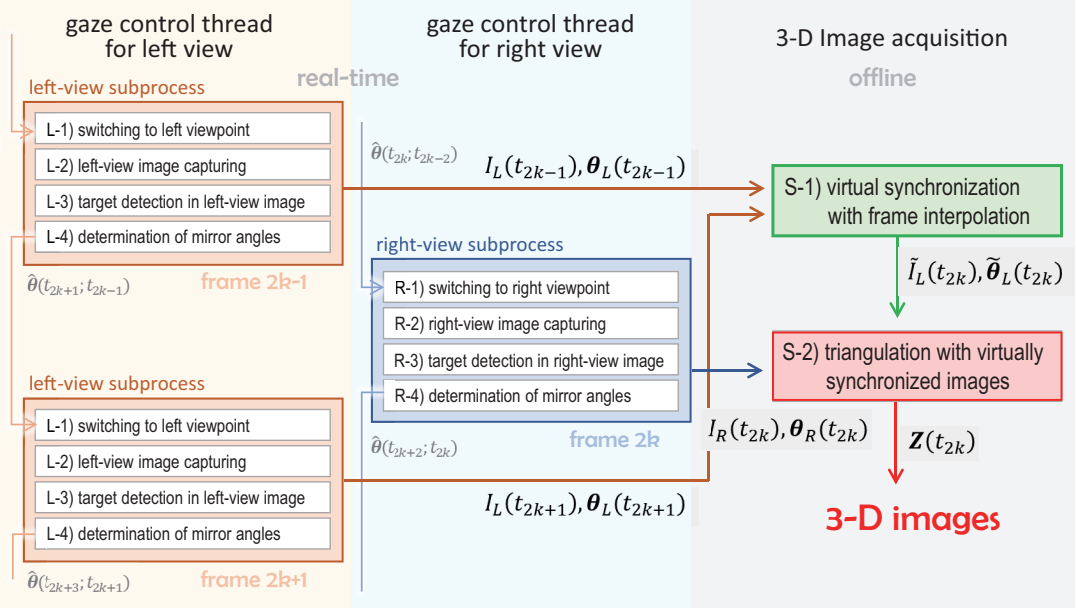


Figure 4.6: Flowchart of the implemented algorithm.

4.3.2 Implemented Algorithm

Assuming that the target scene to be tracked is textured with a specific color, we implement an algorithm to calculate the 3D images using the virtual left and right-view images captured through the ultrafast pan-tilt mirror system: (1) a stereo tracking process with multithread gaze control; and (2) a 3D image estimation process with virtually synchronized images. In this study, the stereo tracking process (1) is executed in real time for mechanical tracking control to keep the target object in view of the virtual left and right pan-tilt cameras with visual feedback, while the 3D image estimation process (2) is executed offline using the left and right-view images, and the pan and tilt angles of the virtual pan-tilt mirror systems, which are being logged during the stereo tracking. Figure 4.6 shows the flowchart of the algorithm.

4.3.2.1 Stereo Tracking Process with Multithread Gaze Control

multiple virtual pan-tilt cameras on a single active vision system as illustrated in Figure 4.7. In the stereo tracking process, the left and right-view subprocesses for multithread gaze control are alternatively switched at a small interval of Δt. The left-view subprocess works for $t_{2k-1} - \tau_m \leq t < t_{2k} - \tau_m$, and that of the right view works for

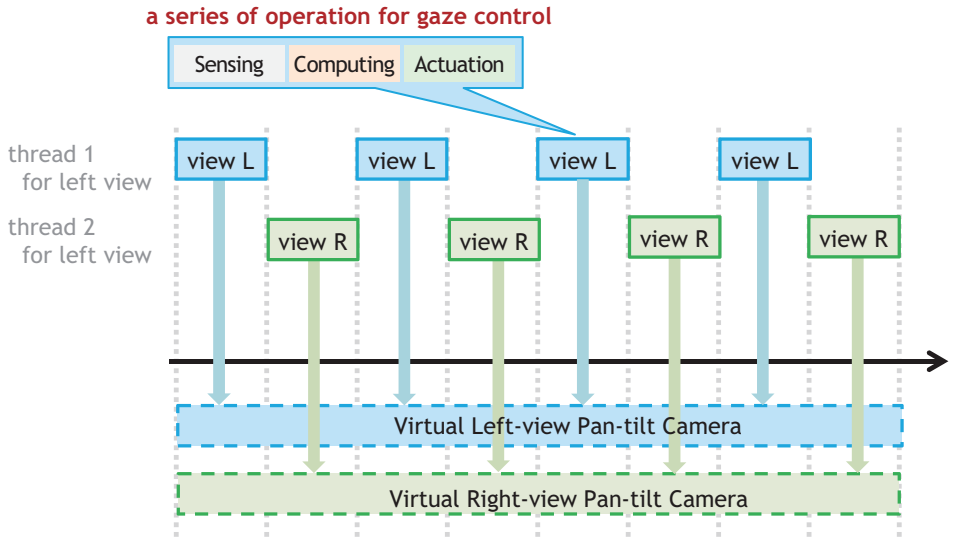


Figure 4.7: Time-division thread processes for virtual pan-tilt cameras in multi-thread gaze control.

$t_{2k} - \tau_m \leq t < t_{2k+1} - \tau_m$ as the time-division thread executes with a temporal granularity of Δt. $t_k = t_0 + k\Delta t$ (k: integer) indicates the image-capturing time of the high-speed vision system, and τ_m is the settling time in controlling the mirror angles of the pan-tilt mirror system.

[Left-view subprocess]

(L-1) Switching to the left viewpoint

For time $t_{2k-1} - \tau_m$ to t_{2k-1}, the pan and tilt angles of the pan-tilt mirror system are controlled to within their desired values $\hat{\boldsymbol{\theta}}(t_{2k-1}; t_{2k-3}) = (\hat{\theta}_1(t_{2k-1}; t_{2k-3}), \hat{\theta}_2(t_{2k-1}; t_{2k-3}))$ at time t_{2k-1}, which is estimated at time t_{2k-3} when capturing the left-view image in the previous frame.

(L-2) Left-view image capturing

The left-view image $I(t_{2k-1})$ is captured at time t_{2k-1}; $I(t)$ indicates the input image of the high-speed vision system at time t.

(L-3) Target detection in left-view image

The target object with a specific color is localized by detecting its center position $\boldsymbol{u}(t) = (u(t), v(t))$ in the image $I(t)$ at time t. Assuming that the color of the target object

to be tracked is different from its background color in this study, $\boldsymbol{u}(t_{2k-1})$ is calculated as a moment centroid of a binary image $C(t_{2k-1}) = C(u, v, t_{2k-1})$ for the target object as follows:

$$\boldsymbol{u}(t_{2k-1}) = (M_u/M_0, M_v/M_0), \tag{4.13}$$

$$M_0 = \sum_{u,v} C(u, v, t_{2k-1}), \quad M_u = \sum_{u,v} uC(u, v, t_{2k-1}), \quad M_v = \sum_{u,v} vC(u, v, t_{2k-1}), \tag{4.14}$$

where the binary image $C(t)$ is obtained at time t by setting a threshold for the HSV (Hue, Saturation, Value) images as follows:

$$C(t) = \begin{cases} 1, & (H_l \leq H < H_h, S > S_l, V > V_l), \\ 0, & \text{(otherwise)}, \end{cases} \tag{4.15}$$

where H, S, and V are the hue, saturation, and value images of $I(t)$, respectively. H_l, H_h, S_l, and V_l are parameters for HSV color thresholding.

(L-4) Determination of mirror angles at the next left-view frame

Assuming that the u- and v-directions in the image correspond to the pan and tilt directions of the pan-tilt mirror system, respectively, the pan and tilt angles at time t_{2k+1} when capturing the left-view image at the next frame, are determined so as to reduce the error between the position of the target object and its desired position $\boldsymbol{u}_L^d$ in the left-view image with proportional control as follows:

$$\hat{\boldsymbol{\theta}}(t_{2k+1}; t_{2k-1}) = -K(\boldsymbol{u}(t_{2k-1}) - \boldsymbol{u}_L^d) + \boldsymbol{\theta}(t_{2k-1}), \tag{4.16}$$

where $\boldsymbol{\theta}(t) = (\theta_1(t), \theta_2(t))$ is collectively the measured values of the pan and tilt angles at time t, and K is the gain parameter for tracking control.

[Right-view subprocess]

(R-1) Switching to right viewpoint

For time $t_{2k} - \tau_m$ to t_{2k}, the pan and tilt angles are controlled to $\hat{\boldsymbol{\theta}}(t_{2k}; t_{2k-2})$, which is estimated at time t_{2k-2} when capturing the right-view image in the next frame.

(R-2) Right-view image capturing

The right-view image $I(t_{2k})$ is captured at time t_{2k}.

(R-3) Target detection in right-view image

$\boldsymbol{u}(t_{2k}) = (u(t_{2k}), v(t_{2k}))$ is obtained as the center position of the target object in the right-view image at time t_{2k}, by calculating a moment centroid of $C(t_{2k})$, which is a sub-image $I(t_{2k})$ of the right-view image, constrained by a color threshold at time t_{2k}, in a similar manner as that described in L-3.

(R-4) Determination of mirror angles in the next right-view frame

Similarly, with the process described in L-4, the pan and tilt angles at time t_{2k+2} when capturing the right-view image in the next frame are determined as follows:

$$\hat{\boldsymbol{\theta}}(t_{2k+2}; t_{2k}) = -K(\boldsymbol{u}(t_{2k}) - \boldsymbol{u}_R^d) + \boldsymbol{\theta}(t_{2k}), \tag{4.17}$$

where $\boldsymbol{u}_R^d$ is the desired position of the target object in the right-view image.

The input images and the mirror angles captured in the stereo tracking process are stored as the left-view images $I_L(t_{2k-1}) = I(t_{2k-1})$ and the pan and tilt angles $\boldsymbol{\theta}_L(t_{2k-1}) = \boldsymbol{\theta}(t_{2k-1})$ at time t_{2k-1}, for the virtual left pan-tilt camera at the odd-numbered frame, and the right-view images $I_R(t_{2k}) = I(t_{2k})$ and the pan and tilt angles $\boldsymbol{\theta}_R(t_{2k}) = \boldsymbol{\theta}(t_{2k})$ at time t_{2k} for the virtual right pan-tilt camera at the even-numbered frame.

4.3.2.2 3D Image Estimation with Virtually Synchronized Images

Left and right-view images in catadioptric stereo tracking are captured at different timings, and the synchronization errors in stereo measurement increase as the target object's movement increases. To reduce such errors, this study introduces a frame interpolation technique for virtual synchronization between virtual left and right pan-tilt cameras, and 3D images are estimated with stereo processing for the virtually synchronized left and right-view images. Frame interpolation is a well-known video processing technique in which intermediate frames are generated between existing frames by means of interpolation using space-time tracking [129–132], view morphing [133–135], and opti-

cal flow [136, 137]; it has been used for many applications, such as frame rate conversion, temporal upsampling for fluid slow motion video, and image morphing.

(S-1) Virtual Synchronization with Frame Interpolation

Considering the right-view image $I_R(t_{2k})$ captured at time t_{2k} as the standard image for virtual synchronization, the left-view image virtually synchronized at time t_{2k}, $\tilde{I}_L(t_{2k})$, is estimated with frame interpolation using the two temporally neighboring left-view images $I_L(t_{2k-1})$ at time t_{2k-1} and $I_L(t_{2k+1})$ at time t_{2k+1} as follows:

$$\tilde{I}_L(t_{2k}) = f_{FI}(I_L(t_{2k-1}), I_L(t_{2k+1})), \tag{4.18}$$

where $f_{FI}(I_1, I_2)$ indicates the frame interpolation function using two images I_1 and I_2. We used Meyer's phase-based method [138] as the frame interpolation technique in this study.

In a similar manner, the pan and tilt angles of the left pan-tilt camera are virtually synchronized with those of the right pan-tilt camera at time t_{2k}, $\tilde{\boldsymbol{\theta}}_L(t_{2k})$, are also estimated using the temporally neighboring mirror angles $\boldsymbol{\theta}_L(t_{2k-1})$ at time t_{2k-1} and $\boldsymbol{\theta}_L(t_{2k+1})$ at time t_{2k+1} as follows:

$$\tilde{\boldsymbol{\theta}}_L(t_{2k}) = \frac{1}{2}(\boldsymbol{\theta}_L(t_{2k-1}) + \boldsymbol{\theta}_L(t_{2k+1})), \tag{4.19}$$

where it is assumed that the mirror angles of the virtual left pan-tilt camera vary linearly for the interval $2\Delta t$ during time t_{2k-1} and t_{2k+1}.

(S-2) Triangulation Using Virtually Synchronized Images

The virtually synchronized left and right-view images at time t_{2k}, $\tilde{I}_L(t_{2k})$ and $I_R(t_{2k})$, are used to compute the 3D image of the tracked object in a similar way as those in the standard stereo methodologies for multiple synchronized cameras. Assuming that the camera parameters of the virtual pan-tilt camera at arbitrary pan and tilt angles $\boldsymbol{\theta}$ are initially given as the 3×4 camera calibration matrix $\boldsymbol{P}(\boldsymbol{\theta})$, the 3D image $\boldsymbol{Z}(t_{2k})$ can be

estimated at time t_{2k} as a disparity map as follows:

$$\boldsymbol{Z}(t_{2k}) = f_{dm}(\tilde{I}_L(t_{2k}), I_R(t_{2k}); \boldsymbol{P}(\tilde{\boldsymbol{\theta}}_L(t_{2k})), \boldsymbol{P}(\boldsymbol{\theta}_R(t_{2k})), \tag{4.20}$$

where $f_{dm}(I_L, I_R; \boldsymbol{P}_L, \boldsymbol{P}_R)$ indicates the function of stereo matching using a pair of left and right-view images, I_L and I_R, when the 3×4 camera calibration matrices of the left- and right cameras are given as $\boldsymbol{P}_L$ and $\boldsymbol{P}_R$, respectively. We used the rSGM method [139] as the stereo matching algorithm in this study.

4.3.3 Specifications

In the stereo tracking process, the viewpoint switching steps (L-1, R-1) require $\tau_m = 1$ ms for the settling time in mirror control, and the image capturing steps (L-2, R-2) require 0.266 ms. The execution time of the target detection steps (L-3,4, R-3,4) is within 0.001 ms, which is accelerated by hardware-implementing the target detection circuit by setting a threshold for the HSV color in Equations (4.13)–(4.15) on the user-specific FPGA of the IDP Express board. Here, the mirrors of the pan-tilt system should be in a state of rest for motion blur reduction in the captured images; the camera exposure in the image capturing steps cannot be executed during the viewpoint switching steps, whereas the target detection steps can be executed in parallel with the next viewpoint switching steps. Thus, the switching time of the left and right-view images is set to $\Delta t = 2$ ms, so that the settling time in mirror control is $\tau_m = 1$ ms and the exposure time is 1 ms. We have confirmed that the stereo tracking process could capture and process a pair of the left- and right-view 8-bit color 512×512 images in real time at 250 fps using a single camera operating at 500 fps.

In this study, the 3D image estimation is executed offline because the computation it requires is too heavy to process 512×512 images in real time at 250 fps on our catadioptric tracking system; the execution time for virtual synchronization with frame interpolation is 54 s, and that for 3D image estimation is approximately 906.3 ms. The 3D image estimation with the rSGM method is too time-consuming to conduct real-time applications such as Simultaneous Localization and Mapping (SLAM) problems and large scale

mapping, whereas the current setup can be used for precise 3D digital archiving/video logging to estimate the 3D images of small objects and creatures fast-moving in the wide area. Our catadioptric stereo tracking system functions as an active stereo system, and its complexity in calibration is similar to those of standard active stereo systems, where there is a trade-off between the complexity and accuracy of the calibration. In this study, focusing on calibration accuracy, the initial look-up-table camera-calibration matrices $\boldsymbol{P}_{lut}(\boldsymbol{\theta}_{ij})$ $(i = 1, \cdots, 52, j = 1, \cdots, 31)$ for 3D image estimation are determined at 52×31 different mirror angles by applying Zhang's calibration method [46] to the captured images of a calibration checkered board at each mirror angle; the pan angle in the range of -10 to -5 degrees and 5 to 10 degrees at intervals of 0.2 degrees, and the tilt angle in the range of -3 to 3 degrees at intervals of 0.2 degrees. In this study, the camera calibration matrix $\boldsymbol{P}(\boldsymbol{\theta})$ at the mirror angle $\boldsymbol{\theta}$ is linearly interpolated with the look-up-table matrices $\boldsymbol{P}_{lut}$ at the four nearest neighbor mirror angles around $\boldsymbol{\theta}$; it can be measured accurately by the angular sensor of the galvano-mirror system at all times, including when the mirror angle is not controlled perfectly to its desired value. Here, it is noted that the offline 3D image estimation that involves heavy computation and the complexity in the camera calibration are the common issues in standard active stereo systems using multiple PTZ cameras, as well as in our catadioptric stereo tracking system.

4.4 Experiments

4.4.1 3D Shapes of Stationary Objects

First, we measured the 3D shapes for the stationary objects at different depths. Figure 4.8 shows the target objects to be measured; a cyan-colored bear doll of $55 \times 65 \times 45$ mm size sitting on a box of $30 \times 15 \times 55$ mm size, two 100 mm-height textured cones, and two differently textured background planes with a depth gap of 10 mm. Except for the bear doll to be tracked, all the objects are black-and-white textured. They were fixed as a rigid-body scene and can move in the x- or z-directions by a linear slider. In the experiment, the cyan-colored regions in the virtual left and right-view images were mechanically tracked at $\boldsymbol{u}_L^d = (310, 255)$ and $\boldsymbol{u}_R^d = (200, 255)$, respectively; the parameters

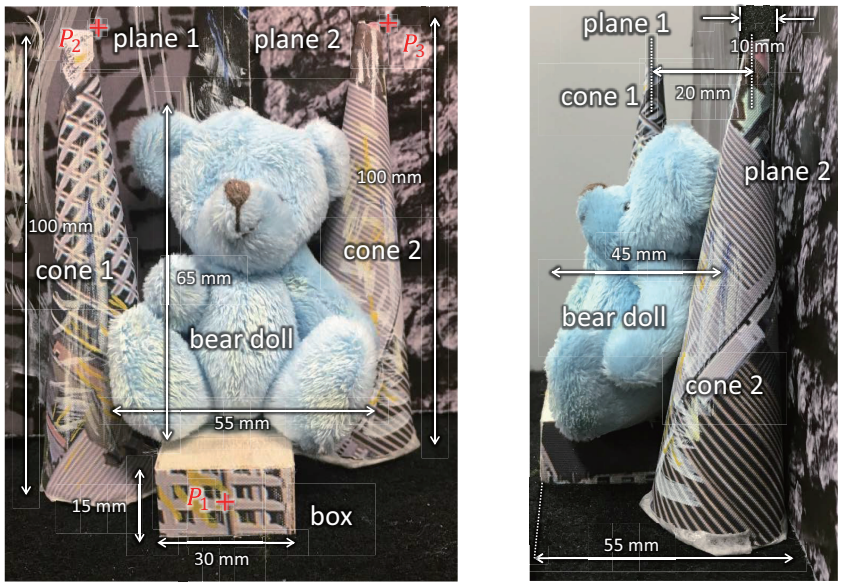
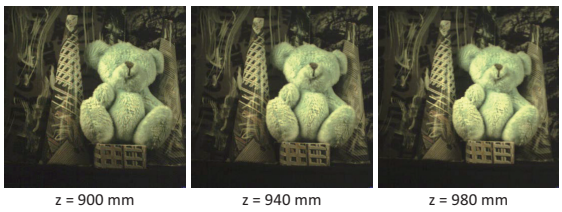


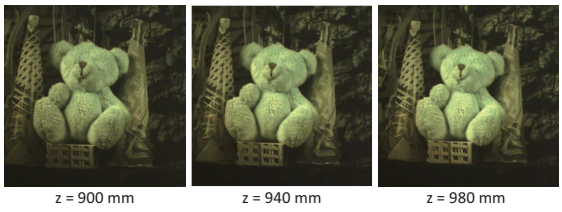
Figure 4.8: 3D scene to be observed.

of the cyan-colored region extraction for 8-bit HSV images were set to $H_l = 85$, $H_h = 110$, $S_l = 20$, and $V_l = 80$. The gain parameter for mirror control was set to $K = 0.01$.

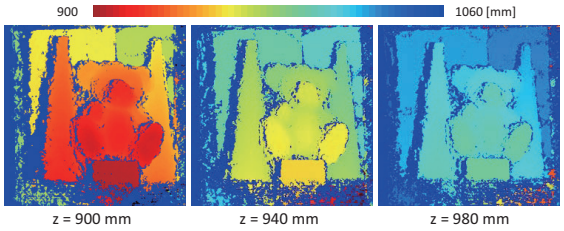
Figure 4.9 shows (a) the left-view images; (b) the right-view images; and (c) the measured 3D images when the distance between the point P_1 on the box under the doll and the system varied along a straight line of $x = 0.9$ mm and $y = 5.3$ mm at $z = 900.0$, 940.0, and 980.0 mm. (d) the xyz coordinate values of the points P_1, P_2, and P_3, (e) the xy-centroids, and (f) the pan and tilt angles of the virtual left and right pan-tilt cameras when the target scene was located at different depths from $z = 900$ to 1000 mm at intervals of 20 mm. The point P_1 is located at the center of the front surface of the box, and the points P_2 and P_3 are located at the left and right-side background planes, respectively; their actual xyz-coordinate values were P_1 (0.9 mm, 5.3 mm, 900.0 mm), P_2 (−24.0 mm, 97.8 mm, 945.0 mm), and P_3 (31.0 mm, 97.8 mm, 955.0 mm). In Figure 4.9c, the 3D shapes of the bear doll, the box, the two cones, and the background planes with a 10 mm depth gap are accurately measured, and they were translated in the z-direction, corresponding to the distance from the system. The xyz-coordinate values almost match the actual coordinate values, and the measurement errors were always within 1.0 mm. The pan angles in both the virtual left and right pan-tilt cameras slightly decreased as the dis-



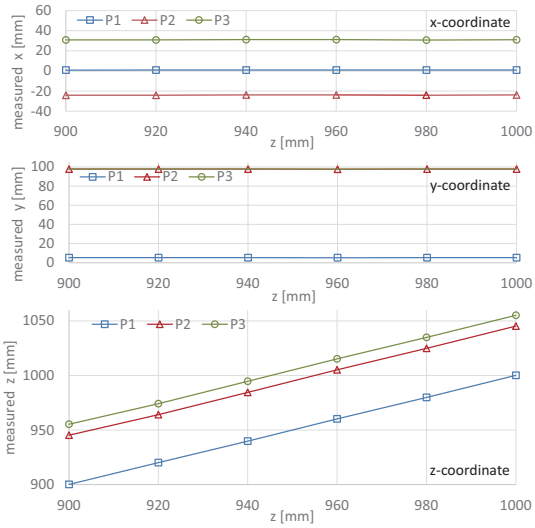
(**a**) left-view images



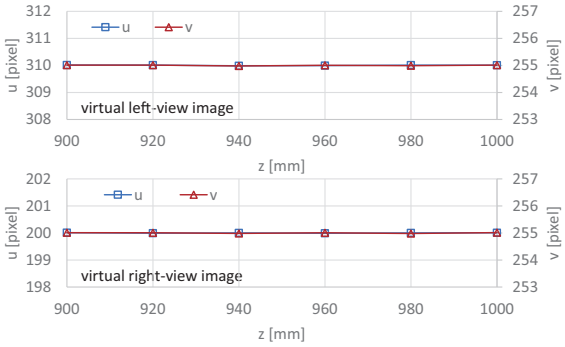
(**b**) right-view images



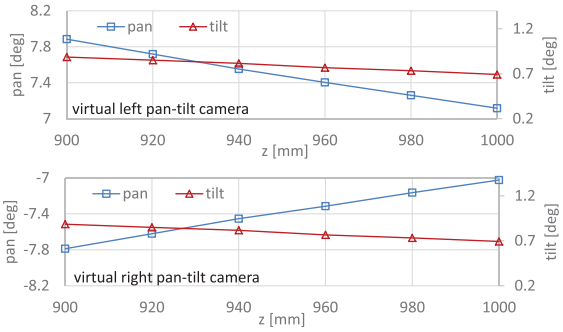
(**c**) measured 3D images



(**d**) measured 3D positions



(**e**) image centroids



(**f**) pan and tilt angles

Figure 4.9: Measured 3D images and positions, image centroids, and pan and tilt angles of virtual left and right pan-tilt cameras for stationary 3D scenes at different depths: (a) left-view images; (b) right-view images; (c) measured 3D images; (d) measured 3D positions; (e) image centroids; (f) pan and tilt angles.

tance between the target object and the system became larger, whereas the *xy* centroids in the left and right-view images were always held to within $(310, 255)$ and $(200, 255)$, respectively; the tracking errors were always within 0.1 pixel. Thus, our catadioptric stereo tracking system can correctly measure the 3D shapes of the stationary target objects.

Next, we measured the 3D images of the target object when the distance between the point P_1 and the system varied along a straight line of $y = 5.3$ mm and $z = 940.0$ mm at $x = -60.0, -30.0, 0.0, 30.0$, and 60.0 mm. Figure 4.10 shows the (a) left-view images; (b) right-view images; and (c) measured 3D images when the target object was mechanically tracked in both the left- and right-view images with multithread gaze control. For the same scenes at different locations, Figure 4.11 shows the experimental results when the mirror angles of the virtual left and right pan-tilt cameras were fixed without tracking; the target object located at $x = 0.0$ mm was observed at $(310, 255)$ and $(200, 255)$ in the left- and right-view images, respectively. In Figure 4.11, the target object located at $x = -60.0$ and 60.0 mm was almost out of the measurable range in the stereo measurement without tracking, whereas the 3D images of the target object were observable continuously in the stereo measurement with tracking, as illustrated in Figure 4.10. Thus, our catadioptric stereo tracking system can expand the measurable area without decreasing resolution by mechanically tracking the target object in both the left- and right-view images, even for the short-distance experiments detailed in this subsection.

4.4.2 3D Shape of Moving Objects

Next, the 3D images of a moving scene at different velocities were measured; the same scene used in the previous subsection was conveyed in the x- and z-directions by a linear slider. The virtual left and right-view images were tracked by setting the same parameters used in the previous subsection.

Figure 4.12a–c shows the left and right-view images, and the measured 3D images when the point P_1 on the box was around $(x, y, z) = (0.5$ mm, 5.3 mm, 930.0 mm); the target scene moved at 500 mm/s in the x-direction. The 3D images $\boldsymbol{Z}(t_{2k})$ measured by using the virtually synchronized left and right-view images ((c) the "FI" method) were illustrated as well as $\boldsymbol{Z}_-(t_{2k})$ and $\boldsymbol{Z}_+(t_{2k})$, which were measured by using the left-view image and the right-view one with a 2 ms delay ((a) the "LR" method), and the right-view

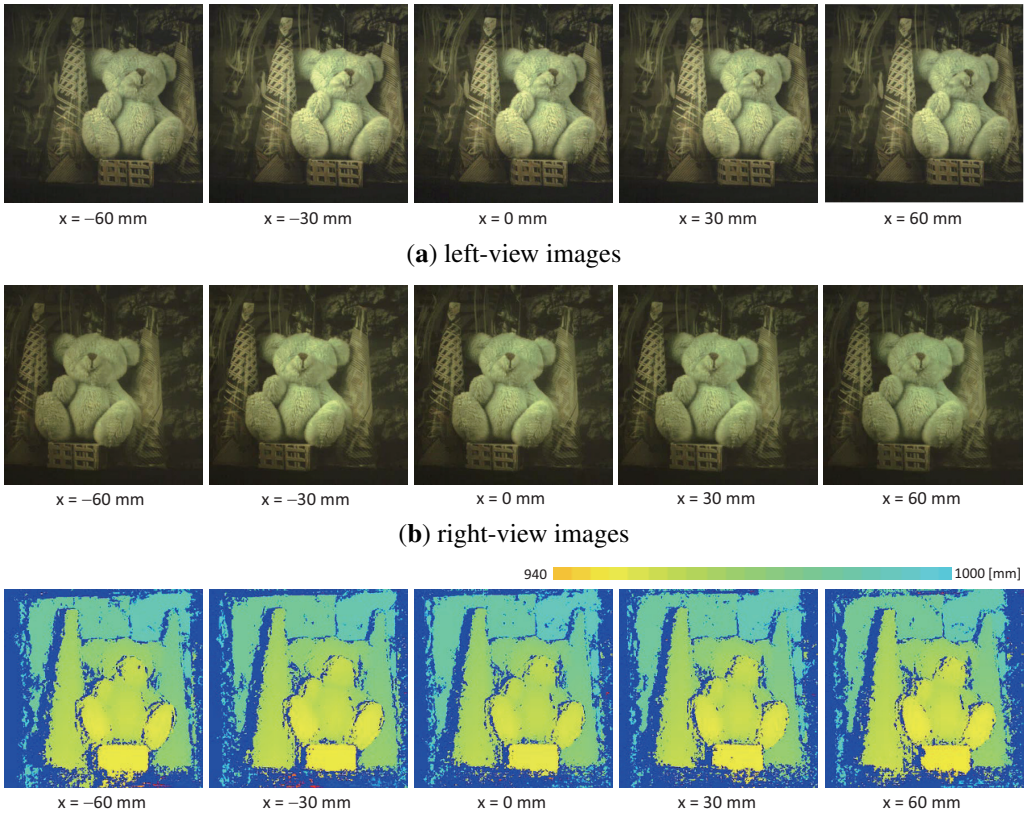


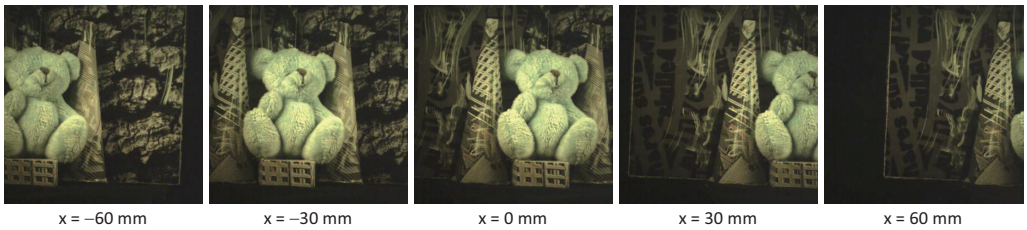
Figure 4.10: Measured 3D images for stationary 3D scenes at different x-coordinates when the target object was mechanically tracked in both the left- and right-view images: (a) left-view images; (b) right-view images; (c) measured 3D images.

image and the left-view one with a 2 ms delay ((b) the 'RL" method), respectively:

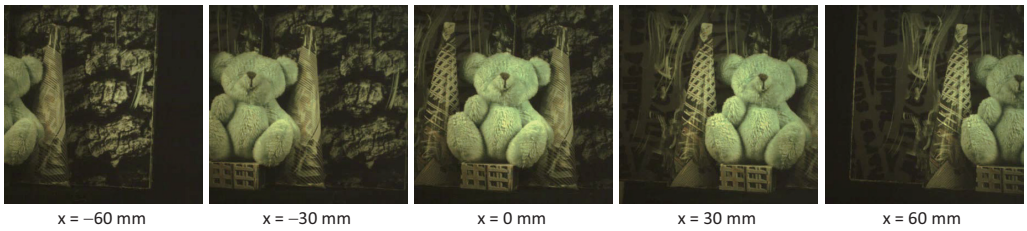
$$\boldsymbol{Z}_{-}(t_{2k}) = f_{dm}(I_L(t_{2k-1}), I_R(t_{2k}); \boldsymbol{P}(\boldsymbol{\theta}_L(t_{2k-1})), \boldsymbol{P}(\boldsymbol{\theta}_R(t_{2k})), \tag{4.21}$$

$$\boldsymbol{Z}_{+}(t_{2k}) = f_{dm}(I_L(t_{2k+1}), I_R(t_{2k}); \boldsymbol{P}(\boldsymbol{\theta}_L(t_{2k+1})), \boldsymbol{P}(\boldsymbol{\theta}_R(t_{2k})). \tag{4.22}$$

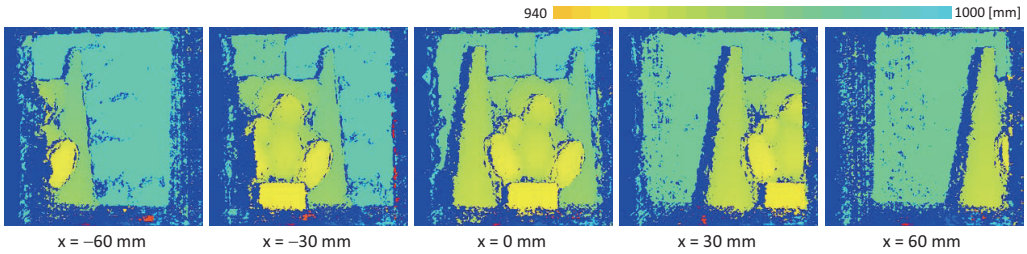
Figure 4.12d–f shows the measured 3D positions at the point P_1, the deviation errors from the actual 3D positions at the points P_1, P_2, and P_3, the image centroids, and the pan and tilt angles of the virtual left and right pan-tilt cameras when the target scene moved at different speeds from −500, −300, −100, 0, 100, 300, and 500 mm/s in the x-direction; the actual positions were P_1(0.5 mm, 5.3 mm, 930.0 mm), P_2(−24.5 mm, 97.8 mm, 975.0 mm), and P_3(30.5 mm, 97.8 mm, 985.0 mm). The 3D positions and errors measured by the "FI" method were compared with those measured by the "LR"



(**a**) left-view images



(**b**) right-view images



(**c**) measured 3D images

Figure 4.11: Measured 3D images for stationary 3D scenes at different x-coordinates when the mirror angles of the virtual left and right pan-tilt cameras were fixed without tracking: (a) left-view images; (b) right-view images; (c) measured 3D images.

and "RL" methods. The pan and tilt angles and image centroids of the virtual right pan-tilt camera were common in all of the measurements, whereas those of the virtual left one differed according to whether virtual synchronization was active. Similarly, Figure 4.13 shows (a)–(c) the left and right-view images, and the measured 3D images when the target scene moved at 500 mm/s in the z-direction; (d) the measured 3D positions at the point P_1; (e) the deviation errors from the actual 3D positions at the points P_1, P_2, and P_3; (f) the image centroids; and (g) the pan and tilt angles of the virtual left and right pan-tilt cameras when the target scene moved at different speeds in the z-direction.

The 3D positions measured by the "FI" method were almost constant when the target scene moved at different speeds in the x- and z-directions, whereas the deviations of the y- and z-coordinate values measured by "LR" and "RL" methods from those measured

when the target scene had no motion increased with the amplitude of the target's speed. The deviation errors at the point P_1 when the target scene moved at 500 mm/s in the x-direction were 2.87. 2.46, and 0.20 mm, respectively, and those when the target scene moved at 500 mm/s in the z-direction, were 1.03, 1.08, and 0.11 mm, respectively; the deviation errors in the "FI" measurement were approximately 1/10 of those in the "LR" and "RL" measurements. A similar tendency was observed in the deviation errors at the points P_2 and P_3. In our system, the target object was always tracked to the desired positions in the left and right-view images by controlling the pan and tilt angles of the virtual left and right pan-tilt cameras.

In Figures 4.12f and 4.13f, the image centroids in the left and right-view images slightly deviated from their desired positions in proportion to the target's speed, which was dependent on the operational limit of the pan-tilt mirror system. This tendency was similar in the "FI", "LR", and "RL" measurements; the deviations of the image centroids in the left and right-view images when the target scene moved at 500 mm/s in the x- and z-directions were 1.2 and 1.3 pixels and 1.0 and 1.0 pixels, respectively. In Figures 4.12 and 4.13, the left-view images in the "FI", "LR", and "RL" measurements were similar, and the differences of the apparent positions of P_1, P_2, and P_3 in all the measurements were within one pixel when the target scene moved at 500 mm/s in the x- and z-directions. This is because the target scene moved together with its background objects as a rigid body, and the left-view images were not so largely varied by tracking the color-patterned object to its desired position in the images.

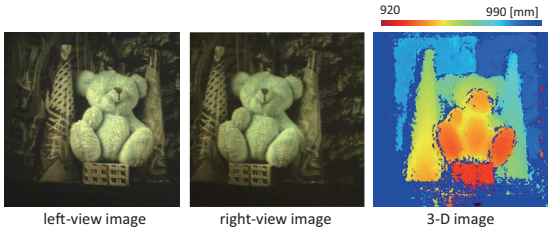
In Figures 4.12g and 4.13g, the pan angle in the left pan-tilt camera in the "FI" measurement when the target scene moved at different speeds in the x- and z-directions was 7.478 degrees, the same as that when the target scene had no motion, whereas those in the "LR" and "RL" measurements deviated in proportion to the target's speed; those in the "LR" and "RL" measurements when 500 mm/s in the x- and z-directions were 7.449 and 7.508 degrees, and 7.492 and 7.461 degrees, respectively. The tilt angle of the left pan-tilt camera and the pan and tilt angles of the right pan-tilt camera were almost similar at different speeds in the x- and z-directions. The measurement errors in the "LR" and "RL" measurements, in which the virtual left pan-tilt camera was not virtually

synchronized with the right one, were mainly caused by these deviations in the pan angle of the left pan-tilt camera, whereas the left-view images were almost similar in the "LR", "RL", and "FI" measurements. This indicates that the 3D images measured by the "FI" method were accurately estimated as the same information as when the target scene moved at different speeds in the x- and z-directions because the synchronization errors in stereo computation were remarkably reduced by synchronizing virtual pan-tilt cameras with frame interpolation. In contrast, the 2-ms interval between the left- and right-view images was not sufficiently large, and the deviation errors were not serious even when the object speed in the experiment was 500 mm/s. Virtual synchronization between left- and right-view images is more effective when a large galvano-mirror system is used for viewpoint-switching with sufficient incident light. This is because the synchronization errors increase when the switching time between the left- and right-view images increases according to the mirror size.

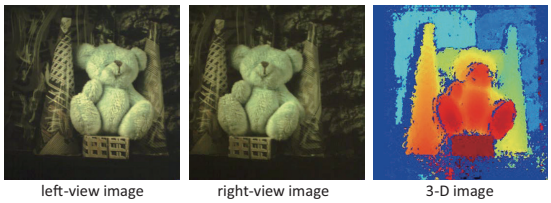
4.4.3 Dancing Doll in 3D Space

Finally, we measured the 3D shapes of a dancing horse doll of size 63 × 25 × 100 mm as illustrated in Figure 4.14. The doll was dancing 40 mm in front of a background plane with black and white patterns. The surface of doll was textured with red color. The virtual left and right-view images were tracked by setting the same parameters used in the previous subsection, excluding the parameters for red-colored region extraction, $H_l = 0$, $H_h = 65$, $S_l = 61$, and $V_l = 115$. The doll and the background plane was moved together at 100 mm/s in the z-direction from $z = 900$ to 1000 mm by the linear slider while the doll was dancing with shaking its body and legs at a frequency of approximately 2.5 Hz.

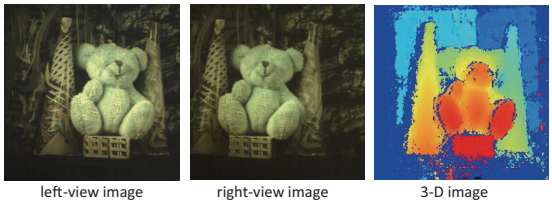
Figure 4.15 shows (a) a sequence of the experimental overviews, monitored using a standard video camera at a fixed position; (b) a sequence of the left-view images; and (c) a sequence of the estimated 3D images with virtual synchronization ("FI" measurement), respectively, which are taken at intervals of 0.2 s. Figure 4.16 shows (a)–(c) the left and right-view images and the 3D images in the "LR", "RL", and "FI" measurements at



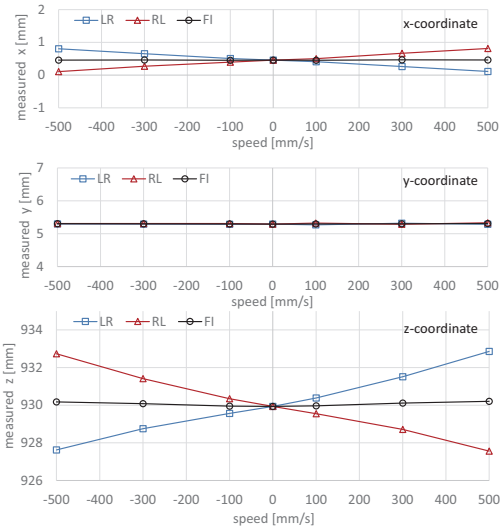
(**a**) "LR" method



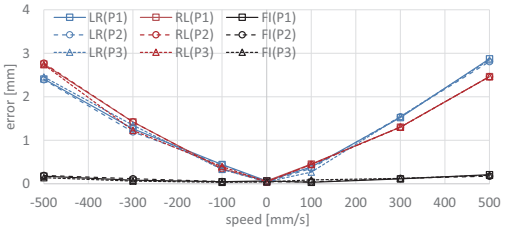
(**b**) "RL" method



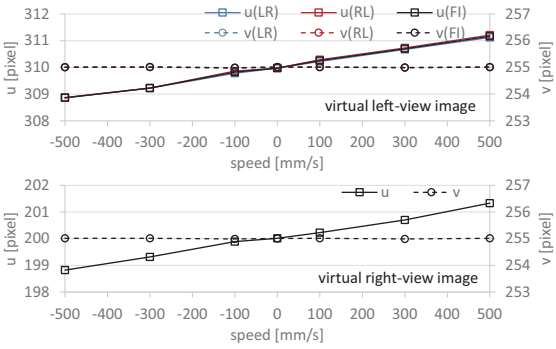
(**c**) "FI" method



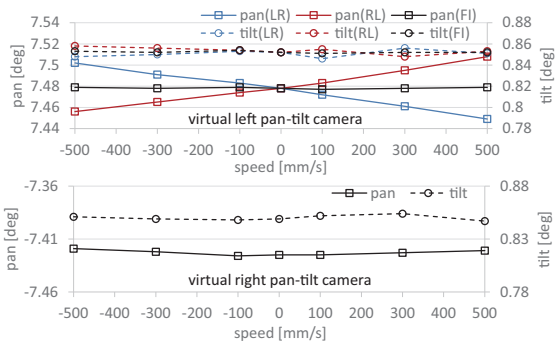
(**d**) measured 3D positions at P_1



(**e**) deviation errors at P_1, P_2, and P_3



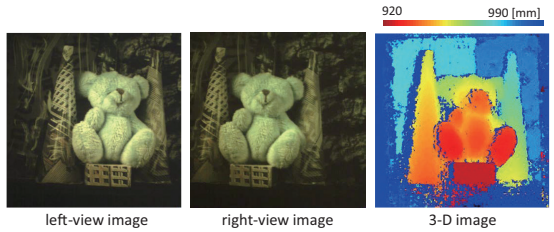
(**f**) image centroids



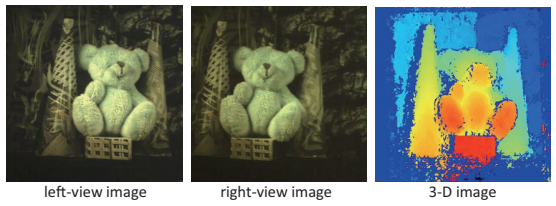
(**g**) pan and tilt angles

Figure 4.12: Left- and right-view images, measured 3D images and positions, deviation errors, image centroids, and pan and tilt angles of virtual left and right pan-tilt cameras when moving at 500 mm/s in the x-direction: (a) LR method; (b) RL method; (c) FI method; (d) measured 3D positions at P_1; (e) deviation errors at P_1 , P_2 , and P_3; (f) image centroids; (g) pan and tilt angles.

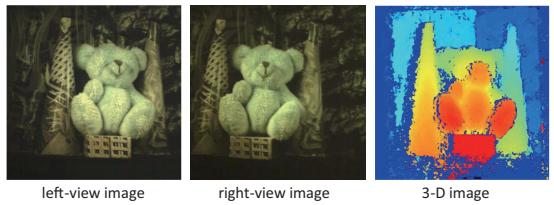
$t = 0.712$ s when the body of the doll quickly moved from up to down with the right-to-left motion of its front legs; (d) the 3D positions of the points P_1 and P_2, (e) the image centroids, and (f) the pan and tilt angles of the left and right pan-tilt cameras in the "FI"



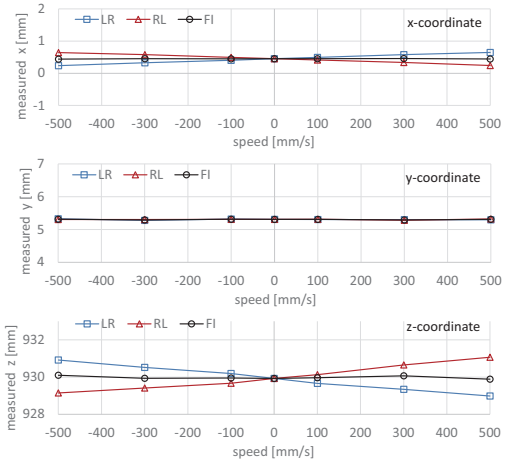
(**a**) "LR" method



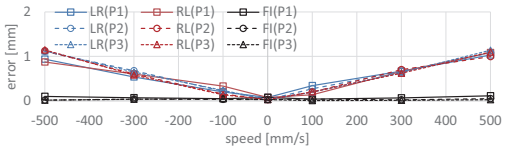
(**b**) "RL" method



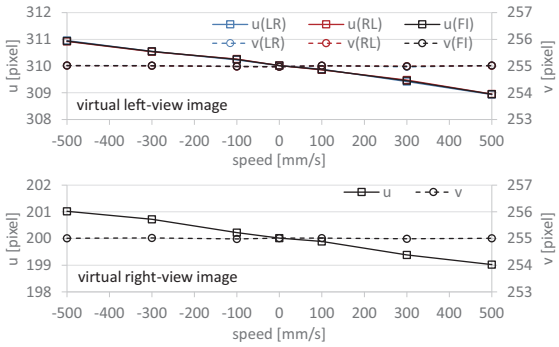
(**c**) "FI" method



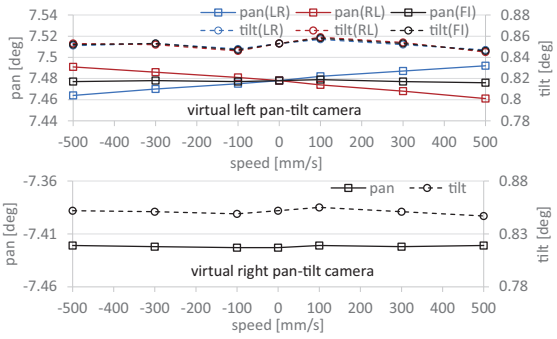
(**d**) measured 3D positions at P_1



(**e**) deviation errors at P_1, P_2, and P_3



(**f**) image centroids



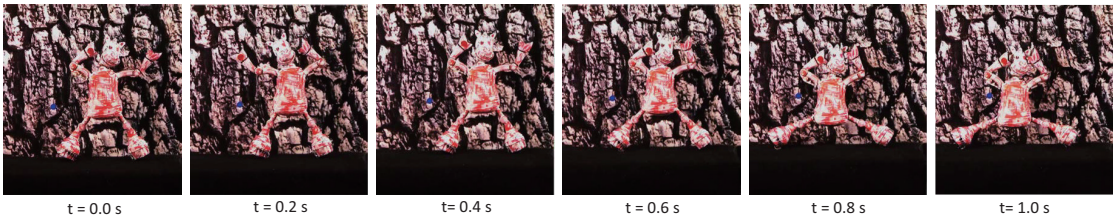
(**g**) pan and tilt angles

Figure 4.13: Left- and right-view images, measured 3D images and positions, deviation errors, image centroids, and pan and tilt angles of virtual left and right pan-tilt cameras when moving at 500 mm/s in the z-direction: (a) LR method; (b) RL method; (c) FI method; (d) measured 3D positions at P_1; (e) deviation errors at P_1 , P_2 , and P_3; (f) image centroids; (g) pan and tilt angles.

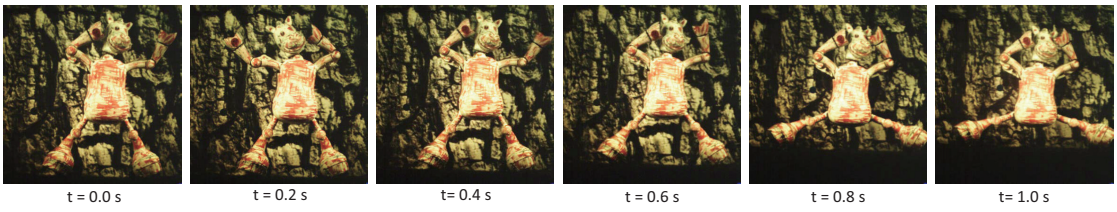
measurement for 1 s. The points P_1 and P_2 were located on the leg of the dancing doll and its background plane, respectively, as illustrated in Figure 4.14.

Figure 4.15 shows that the 3D shapes of the doll's legs, which were cylinder-shaped with 5 mm diameter, were accurately measured when the legs moved quickly at different speeds from those around its other parts, whereas the body of the doll was so controlled

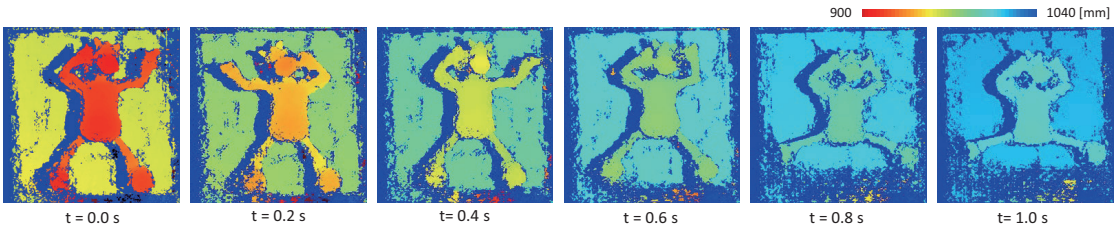
Figure 4.14: Dancing horse doll to be observed.



(**a**) experimental overviews



(**b**) left-view images



(**c**) measured 3D images

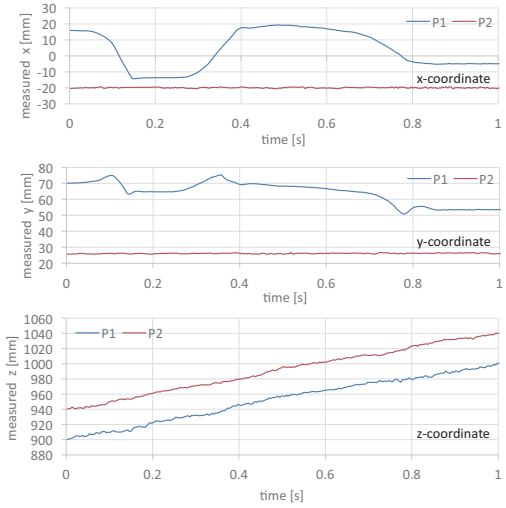
Figure 4.15: [-15](a) Experimental overviews; (b) captured left-view images; and (c) measured 3D images of a dancing doll.

to its desired value $(310, 255)$ in the left-view images that the z-coordinate values of the whole scene increased according to the linear slider's motion from $z = 900$ to 1000 mm

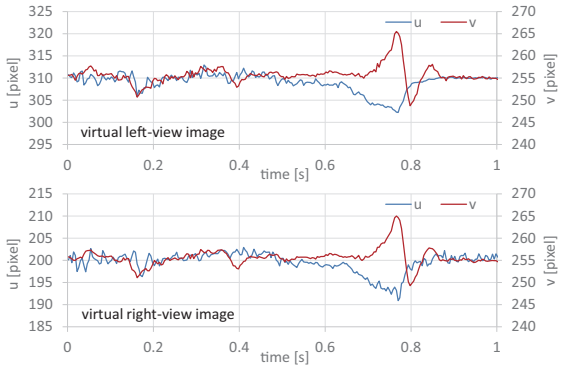
(**a**) "LR" method

(**b**) "RL" method

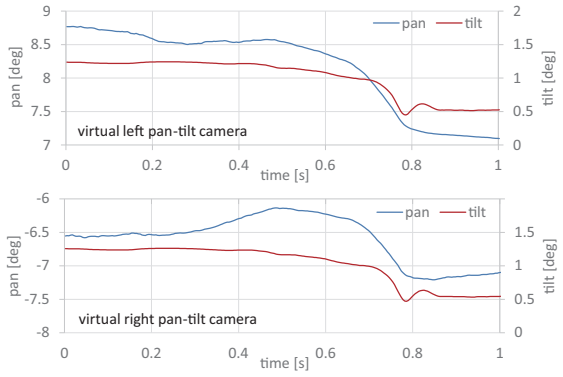
(**c**) "FI" method



(**d**) measured 3D positions



(**e**) image centroids



(**f**) pan and tilt angles

Figure 4.16: Measured 3D images and positions, image centroids, and pan and tilt angles of virtual left and right pan-tilt cameras when observing a dancing doll: (a) LR method; (b) RL method; (c) FI method; (d) measured 3D positions; (e) image centroids; (f) pan and tilt angles.

in the z-direction. In Figure 4.16d, the x- and y-coordinate values measured at the point P_2 were always around $x = -20$ mm and $y = 26$ mm, respectively, and its z-coordinate values varied from $z = 940$ to 1040 mm in a second when the linear slider moved at 100 mm/s in the z-direction; the difference between the z-coordinate values measured at

the points P_1 and P_2 was always around 40 mm, corresponding to the 40-mm distance between the doll and the background plane. The x-coordinate values measured at the point P_1, which was located on the shaking leg of the doll, varied periodically in the range of $x = -14$ to 19 mm, and its y-coordinate values decreased from $y = 70$ to 54 mm according to the up-to-down motion of its body. In Figure 4.16e,f, the pan and tilt angles of the left and right pan-tilt cameras were so controlled that the image centroids in the left- and right-view images were held to around their desired values $(310, 255)$ and $(200, 255)$, respectively, when the dancing doll was moved in the z-direction by the linear slider. For $t = 0.6$–0.8 s when the body of the doll quickly moved from up and down, the image centroids in the left and right-view images slightly deviated from their desired values; these deviations corresponded to the tracking errors when the pan-tilt mirror system could not perfectly track the quick motion of the doll's body in the y-direction. In Figure 4.16a–c, the 3D shapes in "LR", "RL", and "FI" measurements were similarly obtained, whereas the numbers of unmeasurable pixels in the 3D images in the "LR" and "RL" measurements were larger than that in the "FI" measurement. This is because the deviation errors were within 1 mm in the "LR" and "RL" measurements without virtual synchronization when the slider speed was 100 mm/s in the range of $z = 900$ to 1000 mm as illustrated in Figure 4.13e, whereas stereo correspondence was somewhat uncertain or inaccurate in the "LR" and "RL" measurements, due to the vertical displacement between the unsynchronized left and right-view images when capturing the doll moving in the y-direction.

These experimental results indicate that our catadioptric stereo tracking system can accurately measure the 3D shapes of time-varying-shape objects that have local parts at different speeds, while the target object is always tracked at the desired positions in the left and right-view images.

4.5 Conclusions

In this study, we implemented a catadioptric stereo tracking system for monocular stereo measurement by switching 500 different-view images in a second with an ultrafast mirror-drive pan-tilt camera. It can function as two virtual left and right pan-tilt

cameras for stereo measurement that can capture a stereo pair of 8-bit color 512×512 images at 250 fps. Several 3D measurement results were evaluated using the high-frame-rate videos, which were being stored in stereo tracking with multithread gaze control; this evaluation verified the effectiveness of monocular stereo measurement using our catadioptric stereo tracking system with ultrafast viewpoint switching. Synchronization errors in monocular stereo measurement can be reduced by virtually synchronizing the right-view image with the frame-interpolated left-view image.

Chapter 5

Real-Time Monocular Three-Dimensional Multiple Targets Motion Tracking

5.1 Introduction

In this study, we developed a monocular stereo tracking system for use as a marker-based three-dimensional (3-D) motion capture system to localize dozens of markers on multiple moving objects in real time by switching 500 different views in 1 s. The ultrafast mirror-drive active vision used in our catadioptric stereo tracking system can accelerate a series of operation for multithread gaze control with video shooting, computation, and actuation within 2 ms. By switching between 500 different views in one second with real-time video processing for marker extraction at 500 fps, our system can function as J virtual left and right pan-tilt tracking cameras that operate at 250/J fps to simultaneously capture and process J pairs of 512×512 stereo images with different views via the catadioptric mirror system. We conducted several real-time 3-D motion experiments to capture multiple fast-moving objects with markers, where the results demonstrated the effectiveness of our monocular 3-D motion tracking system. We extended the catadioptric stereo tracking system based on an ultrafast pan-tilt mirror system to real-time monocular 3-D motion capture for multiple moving objectsby employing multithread active vision that can function as multiple virtual stereo tracking cameras. The system including high-speed vision and catadioptric device is the same as previous Chapter.

5.2 Monocular 3-D Motion Tracking Algorithm

To estimate the 3-D positions of markers attached to multiple moving objects, we implemented a monocular 3-D motion tracking algorithm with multithread gaze control using the ultrafast pan-tilt mirror system; it can function as J pairs of virtual stereo tracking cameras. For the j-th virtual stereo tracking camera ($j = 0, \cdots, J-1$), the left- and right-view processes for multithread gaze control were alternated at an interval of Δt in order to capture and extract the markers in the left- and right-view images with connected component labels at different timings. The 3-D positions of the markers were estimated by triangulation using the stereo pairs of their corresponding points. The J pairs of virtual left and right pan-tilt cameras independently alternated their tracked areas at an interval of $2\Delta t$ for different moving target objects.

In the algorithm, the virtual left and right pan-tilt cameras are virtually synchronized with frame interpolation in order to reduce stereo measurement errors when the observed objects make large movements. The left-view image and the mirror angles of the virtual left pan-tilt camera are frame-interpolated in the left-view process so their capture timing corresponds to that of the virtual right pan-tilt camera, and the 3-D positions of the markers are estimated in the right-view process. For the j-th virtual stereo tracking camera, the left-view process works for $t_{2Jk+2j-1} - \tau_{mt} \leq t < t_{2Jk+2j} - \tau_{mt}$ and the right-view one works for $t_{2Jk+2j} - \tau_{mt} \leq t < t_{2Jk+2j+1} - \tau_{mt}$. $I(\boldsymbol{u}, t_k)$ indicates the image intensity at pixel $\boldsymbol{u} = (u, v)$ in the input image captured at time t_k and $t_k = t_0 + k\Delta t$ (k: integer) indicates the capture time for the vision system. τ_{mt} is the time required to settle the mirror angles in the pan-tilt mirror system. For the j-th virtual stereo tracking camera, the 3-D measuremental results are computed at an interval of $2J\Delta t$ by executing the following left- and right-view processes.

5.2.1 Left-view Process

L1) Viewpoint switching

At time $t_{2Jk+2j-1}$, the pan and tilt angles of the pan-tilt mirror system complete their movement to

$$\hat{\boldsymbol{\alpha}}(t_{2Jk+2j-1}; t_{2J(k-1)+2j-1}) = (\hat{\alpha}(t_{2Jk+2j-1}; t_{2J(k-1)+2j-1}), \hat{\beta}(t_{2Jk+2j-1}; t_{2J(k-1)+2j-1})), \tag{5.1}$$

where $\hat{\boldsymbol{\alpha}}(t_{2Jk+2j-1}; t_{2J(k-1)+2j-1})$ is the desired value of the pan and tilt angles at time $t_{2Jk+2j-1}$, which is estimated at time $t_{2J(k-1)+2j-1}$ when capturing the left-view image for the j-th virtual stereo tracking camera in the previous frame.

L2) Image acquisition and binarization

The left-view image $I(\boldsymbol{u}, t_{2Jk+2j-1})$ of $M \times N$ pixels is captured at time $t_{2Jk+2j-1}$ and it is converted into a binary image $B(\boldsymbol{u}, t_{2Jk+2j-1})$ with a threshold B_θ as follows.

$$B(\boldsymbol{u}, t_{2Jk+2j-1}) = \begin{cases} 1 & (I(\boldsymbol{u}, t_{2Jk+2j-1}) \geq B_\theta) \\ 0 & (\text{otherwise}) \end{cases}. \tag{5.2}$$

L3) Calculation of cell-based moment features

To reduce the number of pixels scanned for labeling, $B(\boldsymbol{u}, t_{2Jk+2j-1})$ is divided into $M'N'$ cells Γ_{ab} ($a = 0, \ldots, M'-1;\ b = 0, \ldots, N'-1$) with $m \times n$ pixels, where $M = mM'$ and $N = nN'$, and the 0th- and 1st-order moment features are calculated for each cell as follows:

$$M_{pq}(\Gamma_{ab}, t_{2Jk+2j-1}) = \sum_{u=am}^{a(m+1)-1} \sum_{v=bn}^{b(n+1)-1} u^p v^q \cdot B(\boldsymbol{u}, t_{2Jk+2j-1}), \tag{5.3}$$

where $(p, q) = (0, 0)$, $(1, 0)$, and $(0, 1)$. Γ_{ab} is expressed as follows:

$$\Gamma_{ab} = \{(u, v) \mid (am+\xi, bn+\eta),\ 0 \leq \xi < m, 0 \leq \eta < n\}, \tag{5.4}$$

where u is $am+\xi$, and v is $bn+\eta$.

L4) Marker extraction with cell-based labeling

To obtain the moment features of multiple markers in an image, the connected components labeling process is accelerated by implementing the cell-based labeling al-

gorithm [112], which can reduce the computational complexity in the order of $O(M'N')$, where this is $1/mn$ of the pixel-level complexity of the order $O(MN)$. In addition to scanning a flag map $F(\Gamma_{ab}, t_{2Jk+2j-1})$ of $M'N'$ cells, the connected regions $O_i(t_{2Jk+2j-1})$ $(i = 0, \ldots, I_j - 1)$ are labeled in the left-view binary image $B(\boldsymbol{u}, t_{2Jk+2j-1})$ at time t_{2k-1}, and the label-domain moment features $M_{pq}(O_i(t_{2Jk+2j-1}))$ are accumulated as follows:

$$M_{pq}(O_i(t_{2Jk+2j-1})) = \sum_{\boldsymbol{u} \in O_i(t_{2k-1})} u^p v^q \cdot B(\boldsymbol{u}, t_{2Jk+2j-1}), \tag{5.5}$$

where the flag map $F(\Gamma_{ab}, t_{2Jk+2j-1})$ is defined by thresholding $M_{00}(\Gamma_{ab}, t_{2Jk+2j-1})$ with F_θ as follows:

$$F(\Gamma_{ab}, t_{2Jk+2j-1}) = \begin{cases} 1 & (M_{00}(\Gamma_{ab}, t_{2Jk+2j-1}) \geq F_\theta) \\ 0 & (\text{otherwise}) \end{cases}. \tag{5.6}$$

The center positions $\boldsymbol{u}_L^{ji}(t_{2Jk+2j-1}) = \boldsymbol{u}(O_i(t_{2Jk+2j-1}))$ of the labeled regions $O_i(t_{2Jk+2j-1})$ at time $t_{2Jk+2j-1}$ $(i = 0, \cdots, I_j - 1)$ are obtained as those of the markers in the left-view image for the j-th virtual stereo tracking camera using their 0th- and 1st-order moment features, as follows:

$$\boldsymbol{u}_L^{ji}(t_{2Jk+2j-1}) = \left(\frac{M_{10}(O_i(t_{2Jk+2j-1}))}{M_{00}(O_i(t_{2Jk+2j-1}))}, \frac{M_{01}(O_i(t_{2Jk+2j-1}))}{M_{00}(O_i(t_{2Jk+2j-1}))} \right) \tag{5.7}$$

The detailed processes in the cell-based labeling algorithm were described previously by [140].

L5) Determination of mirror angles for the next frame

Assuming that the u and v directions in the image correspond to the pan and tilt directions of the pan-tilt mirror system, respectively, then the pan and tilt angles at time $t_{2J(k+1)+2j-1}$ are determined when capturing the left-view image of the j-th virtual stereo tracking camera for the next frame in order to reduce the error between the position of the target object and its desired position $\boldsymbol{u}_L^d$ in the left-view image with proportional control

as follows:

$$\hat{\boldsymbol{\alpha}}(t_{2J(k+1)+2j-1}; t_{2Jk+2j-1}) = \\ -K(\bar{\boldsymbol{u}}_L^j(t_{2Jk+2j-1}) - \boldsymbol{u}_L^d) + \boldsymbol{\alpha}(t_{2Jk+2j-1}), \tag{5.8}$$

where K is the gain parameter for tracking control, and $\boldsymbol{\alpha}(t_{2Jk+2j-1}) = (\alpha(t_{2Jk+2j-1}), \beta(t_{2Jk+2j-1}))$ comprise the measured values of the pan and tilt angles at time, $t_{2Jk+2j-1}$. $\bar{\boldsymbol{u}}_L^j(t_{2Jk+2j-1})$ is the averaged center position of the markers in the left-view image at time $t_{2Jk+2j-1}$ as follows.

$$\bar{\boldsymbol{u}}_L^j(t_{2Jk+2j-1}) = \frac{1}{I_j}\sum_{i=0}^{I_j-1} \boldsymbol{u}_L^{ji}(t_{2Jk+2j-1}). \tag{5.9}$$

L6) Virtual synchronization with frame-interpolation

The center positions of the markers in the left-view image are virtually synchronized with those in the right-view image captured at time t_{2Jk+2j}. The virtually synchronized center positions $\tilde{\boldsymbol{u}}_L^{ji}(t_{2Jk+2j})$ $(i = 0, \cdots, I_j - 1)$ in the left-view image are estimated by frame interpolation using the center positions of the markers in the left-view images captured at time $t_{2Jk+2j-1}$ and $t_{2J(k+1)+2j-1}$ as follows:

$$\tilde{\boldsymbol{u}}_L^{ji}(t_{2Jk+2j}) = \frac{(J-1)\boldsymbol{u}_L^{ji}(t_{2Jk+2j-1}) + \boldsymbol{u}_L^{ji}(t_{2J(k+1)+2j-1})}{J} \tag{5.10}$$

Similarly, the pan and tilt angles $\tilde{\boldsymbol{\alpha}}_L(t_{2Jk+2j})$ of the left pan-tilt camera of the j-th virtual stereo tracking camera virtually synchronized with those of the right pan-tilt camera at time t_{2Jk+2j} are as follows:

$$\tilde{\boldsymbol{\alpha}}_L(t_{2Jk+2j}) = \frac{(J-1)\boldsymbol{\alpha}_L(t_{2Jk+2j-1}) + \boldsymbol{\alpha}_L(t_{2J(k+1)+2j-1})}{J} \tag{5.11}$$

where it is assumed that the center positions of the markers in the left-view images and the mirror angles of the virtual left pan-tilt camera vary in a linear manner for the interval $2J\Delta t$ during times $t_{2Jk+2j-1}$ and $t_{2J(k+1)+2j-1}$.

5.2.2 Right-view Process

In the right-view process, the subprocesses in steps R1–R5 are executed in a similar manner to those in steps L1–L5 in the left-view process, and the 3-D positions of the markers are estimated via stereo triangulation for the virtually synchronized pan-tilt cameras in step R6.

R1) Viewpoint switching

At time t_{2Jk+2j}, the pan and tilt angles have completed their movement to their desired value $\hat{\boldsymbol{\alpha}}(t_{2Jk+2j}; t_{2J(k-1)+2j})$, which is estimated at time $t_{2J(k-1)+2j}$ when capturing the right-view image for the next frame.

R2) Image acquisition and binarization

The right-view image $I(\boldsymbol{u}, t_{2Jk+2j})$ is captured at time t_{2Jk+2j} and it is converted into a binary image $B(\boldsymbol{u}, t_{2Jk+2j})$ with a threshold B_θ.

R3) Calculation of cell-based moment features

$B(\boldsymbol{u}, t_{2Jk+2j})$ is divided into $M'N'$ cells Γ_{ab} of $m \times n$ pixels and the cell-based moment features $M_{pq}(\Gamma_{ab}, t_{2Jk+2j})$ are calculated in a similar manner to those in step L3.

R4) Marker extraction with cell-based labeling

The connected regions $O_i(t_{2Jk+2j})$ $(i = 0, \ldots, I_j - 1)$ are labeled in $B(\boldsymbol{u}, t_{2Jk+2j})$ at time t_{2Jk+2j}, and their center positions $\boldsymbol{u}_R^{ji}(t_{2Jk+2j}) = \boldsymbol{u}(O_i(t_{2Jk+2j}))$ $(i = 0, \cdots, I_j - 1)$ are determined as those of the markers in the right-view image by accumulating the label-domain moment features $M_{pq}(O_i(t_{2Jk+2j}))$ $(p + q \leq 1)$ in a similar manner to those in step L4. In this process, it is assumed that all of the I_j markers in both the left- and right-view images of the j-th virtual stereo tracking camera can be extracted correctly with connected component labeling.

R5) Determination of mirror angles for the next frame

In a similar manner to step L5, the pan and tilt angles at time $t_{2J(k+1)+2j}$ are determined when capturing the right-view image for the next frame by using the pan and tilt angles $\boldsymbol{\alpha}(t_{2Jk+2j})$ measured at time t_{2Jk+2j}, as follows:

$$\hat{\boldsymbol{\alpha}}(t_{2J(k+1)+2j}; t_{2Jk+2j}) = -K(\bar{\boldsymbol{u}}_R^j(t_{2Jk+2j}) - \boldsymbol{u}_R^d) + \boldsymbol{\alpha}(t_{2Jk+2j}), \tag{5.12}$$

where $\bar{\boldsymbol{u}}_R^j(t_{2Jk+2j})$ is the averaged center position of the markers at time t_{2Jk+2j} and $\boldsymbol{u}_R^d$ is its desired value in the right-view image.

R6) Triangulation using virtual synchronized images

The positions of the markers for the left and right pan-tilt cameras, which are virtually synchronized at time t_{2k}, are used for stereo triangulation to compute the 3-D positions of the observed markers. It is assumed that the camera parameters for the virtual pan-tilt camera at arbitrary pan and tilt angles $\boldsymbol{\alpha}$ are initially given as the 3×4 camera matrices $\boldsymbol{P}(\boldsymbol{\alpha})$. First, the corresponding pairs of the marker positions in the left- and right-view images of the j-th stereo tracking camera, $\tilde{\boldsymbol{u}}_L^{ji}(t_{2Jk+2j})$ and $\boldsymbol{u}_R^{ji}(t_{2Jk+2j})$, are determined by considering the constraints on the epipolar geometry [141, 142], and the 3-D marker positions of the j-th stereo tracking camera at time t_{2Jk+2j}, $\boldsymbol{x}^{ji}(t_{2Jk+2j}) = (x^{ji}(t_{2Jk+2j}), y^{ji}(t_{2Jk+2j}), z^{ji}(t_{2Jk+2j}))$ $(i = 0, \cdots, I_j - 1)$ are estimated by stereo triangulation for the corresponding pairs of the marker positions in a similar manner to those in the standard stereo methods for multiple synchronized cameras, as follows:

$$\begin{aligned} \boldsymbol{x}^{ij}(t_{2Jk+2j}) = \quad & f_{tri}(\tilde{\boldsymbol{u}}_L^{ji}(t_{2Jk+2j}), \boldsymbol{u}_R^{ij}(t_{2Jk+2j}); \\ & \boldsymbol{P}(\tilde{\boldsymbol{\alpha}}_L(t_{2Jk+2j})), \boldsymbol{P}(\boldsymbol{\alpha}_R(t_{2Jk+2j})), \end{aligned} \tag{5.13}$$

where $f_{tri}(\boldsymbol{u}_L, \boldsymbol{u}_R; \boldsymbol{P}_L, \boldsymbol{P}_R)$ indicates the stereo triangulation function when the camera matrices of the left- and right cameras are $\boldsymbol{P}_L$ and $\boldsymbol{P}_R$, respectively. When two or more markers are located on the same epipolar line of a pair of left and right images, they correspond by considering the ordering consistency constraint so that the markers in the left and right images have the same order along the epipolar line.

5.2.3 Specifications

In this study, steps L2–L4 and R2–R4 with image size complexity in the order of $O(MN)$ were accelerated by a hardware implementation of the cell-based labeling circuit for multi-object extraction [112] on the user-specific FPGA of the IDP Express board. We used the cell-based labeling circuit to extract connected components with cell sizes of 4×4 pixels ($m = n = 4$) in a 512×512 image ($M = N = 512$). The other steps with

complexities in the order of $O(I_j)$ or less were implemented in software. The execution time when the number of tracked markers $I_j = 50$ ($j = 0, \cdots, J-1$) was evaluated as follows.

Step L1/R1 for viewpoint switching required $\tau_{mt} = 1$ ms for the settling time during mirror control. Steps L2–L4/R2–R4 from image acquisition to multi-marker extraction were executed within 0.25 ms by using the cell-based labeling circuit. In step L5/R5, the mirror angles were determined for the next left- or right-view frame within 0.003 ms. Step L6 required 0.004 ms for virtual synchronization of the marker positions and mirror angles in the virtual left pan-tilt camera, whereas stereo triangulation using 50 corresponding pairs of markers could be executed in 0.01 ms in step R6. The switching time for the left- and right-view images was set to $\Delta t = 2$ ms, so the settling time during mirror control was $\tau_{mt} = 1$ ms and the exposure time was 1 ms. By switching and processing a pair of the left- and right-view 512×512 images using a single camera operating at 500 fps, we confirmed that the 3-D motion tracking process could extract the 3-D positions of $50J$ markers in real time at $250/J$ fps when our system functioned as J virtual stereo tracking cameras. The camera matrices $\boldsymbol{P}_{lut}(\boldsymbol{\alpha}_{\xi\eta})$ at 52×31 mirror angles ($\xi = 0, \cdots, 51, \eta = 0, \cdots, 30$) were given as a look-up-table. The pan angle was in the ranges of -10 to -5 degrees and 5 to 10 degrees, and the tilt angle was in the range of -3 to 3 degrees at intervals of 0.2 degrees. The camera matrix $\boldsymbol{P}(\boldsymbol{\alpha})$ at the mirror angle $\boldsymbol{\alpha}$ was linearly interpolated with the look-up table matrices $\boldsymbol{P}_{lut}$ at the four nearest neighbor mirror angles around $\boldsymbol{\alpha}$.

5.3 Experiments

5.3.1 Stationary Marker at Different Depths

First, we measured the 3-D position of a stationary object at different depths. The object was a white circular marker with a diameter of 5 mm on a black background.

The *uv* centroids in both the virtual left- and right-view images, which corresponded to the apparent target positions, were controlled to their image centers $\boldsymbol{u}_L^d = \boldsymbol{u}_R^d = (255, 255)$. The thresholds for binarization and flag-map generation were set to $B_\theta = 70$ and $F_\theta =$

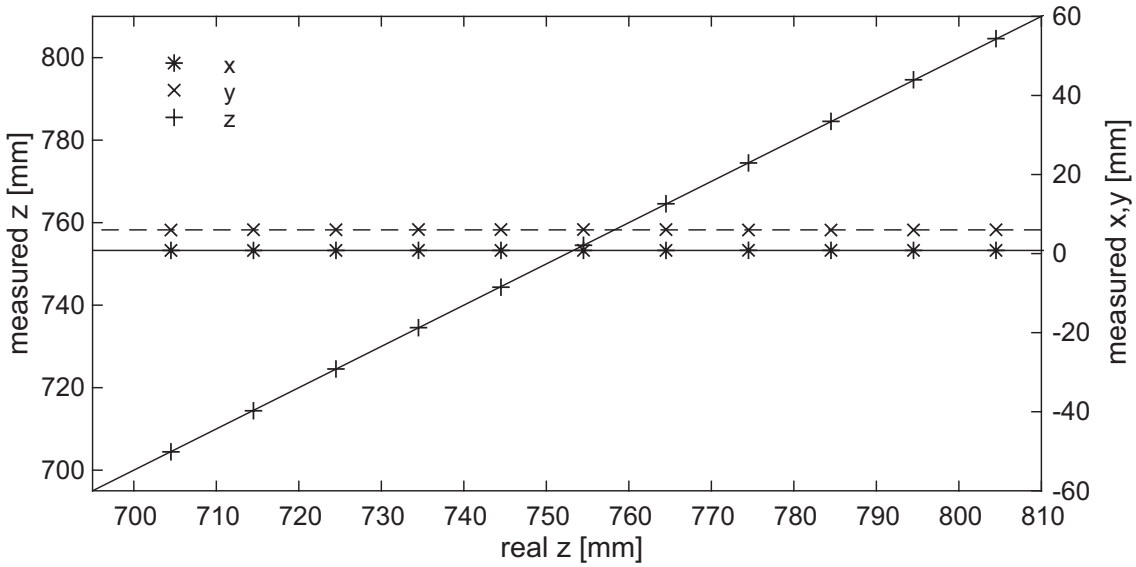
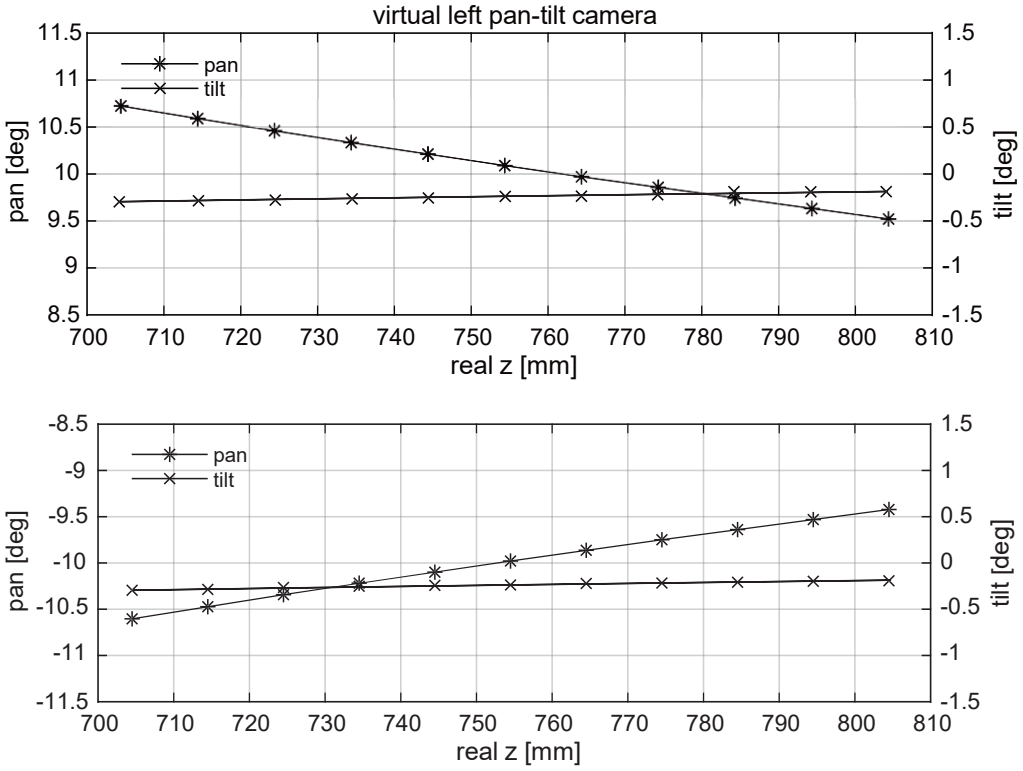


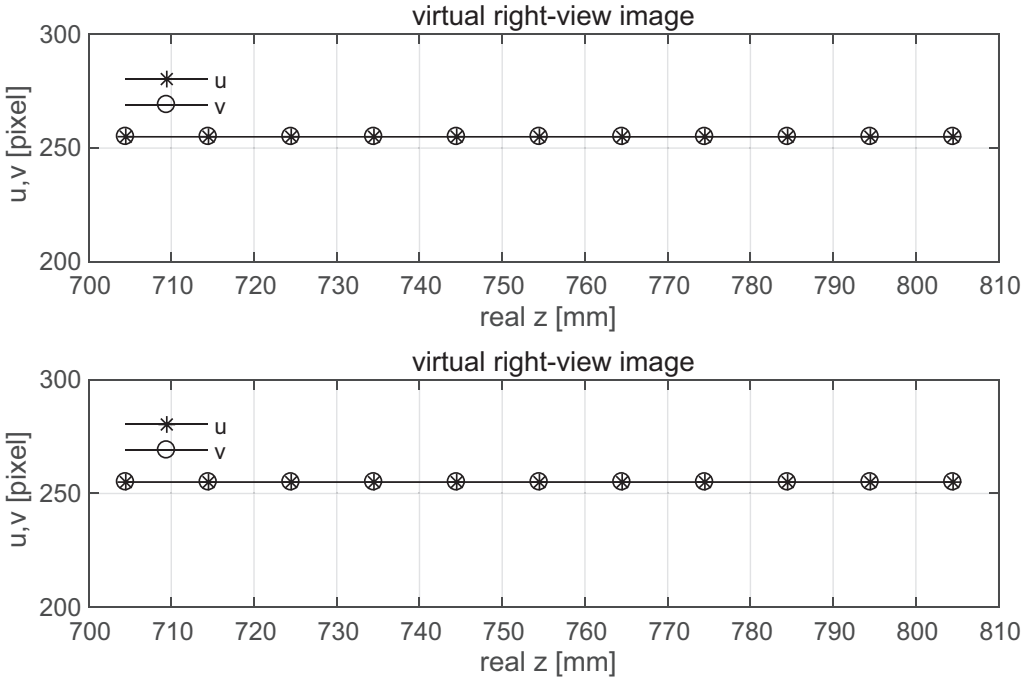
Figure 5.1: Measured 3-D positions of a stationary marker at different depths.

0 pixel, respectively. In the experiments, one pair of virtual stereo tracking cameras functioned, ($J = 1$), and the number of markers for extraction was $I_0 = 1$.

Figure 5.1 shows the 3-D position of the marker when the distance between the marker and the system was changed along a straight line of $x = 0.8$ mm and $y = 6.0$ mm from $z = 704.5$ to 804.5 mm at intervals of 10 mm. Figure 5.2 shows (a) the pan and tilt angles of the virtual left and right pan-tilt cameras, and (b) the *uv* centroids of the virtual left- and right-view images when the marker was located at different depths. The *xyz* coordinate values in Figure 5.1 were computed by considering the pan and tilt angles as well as the *uv* centroids of the virtual pan-tilt cameras. The pan angles in both the virtual left and right pan-tilt cameras decreased slightly as the distance between the marker and the system increased. Figure 5.1 shows that the *xyz* coordinate values almost agreed with the actual coordinate values, where the errors were within 0.2 mm at all of the measurement points. Thus, our catadioptric stereo tracking system could correctly measure the 3-D position of a stationary marker.



(a) Pan and tilt angles



(b) Image centroids

Figure 5.2: Pan and tilt angles, and image centroids of virtual left and right pan-tilt cameras when observing a stationary marker at different depths.

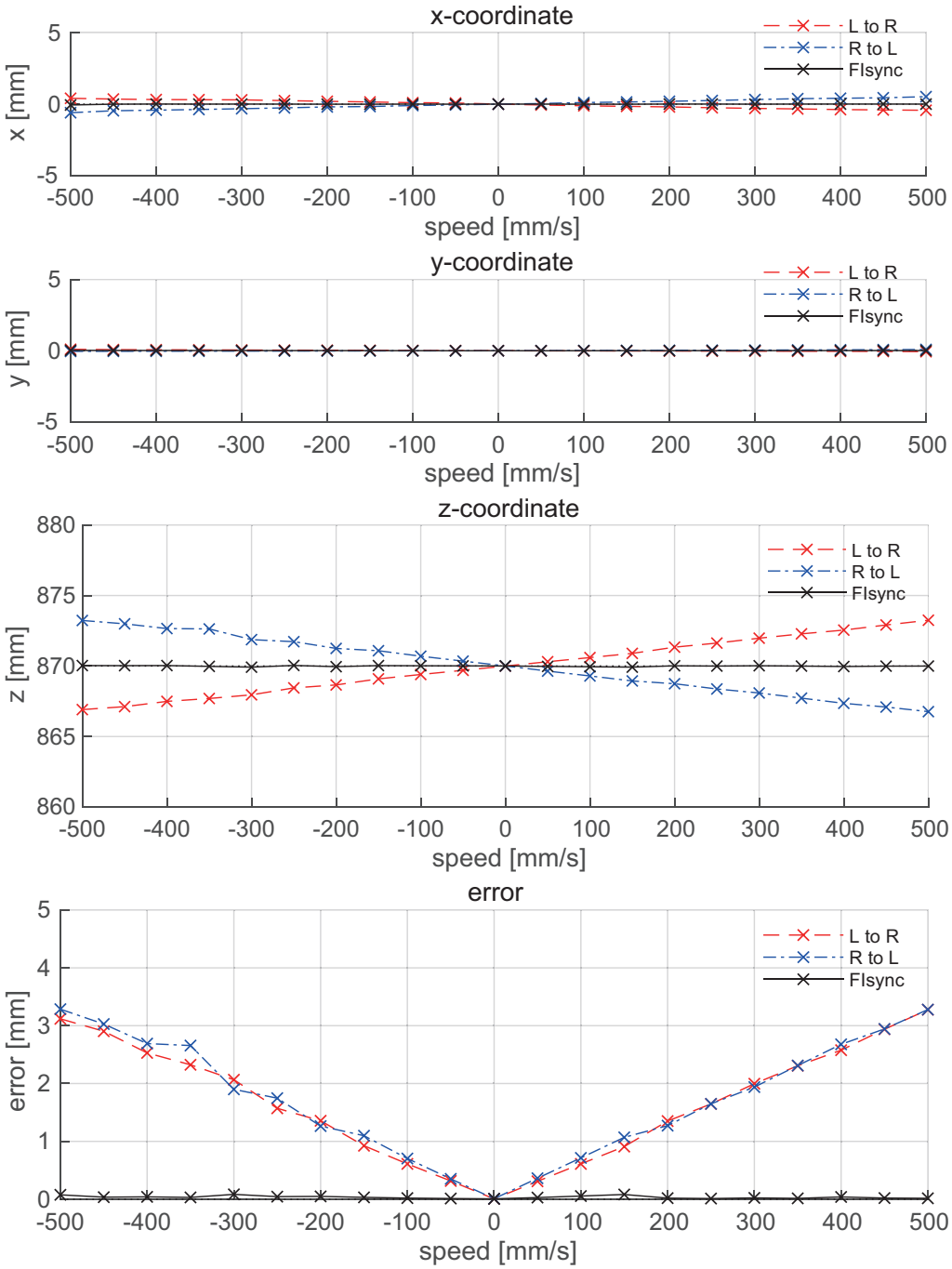


Figure 5.3: Measured 3-D positions and errors when a marker moved at different velocities in the x direction.

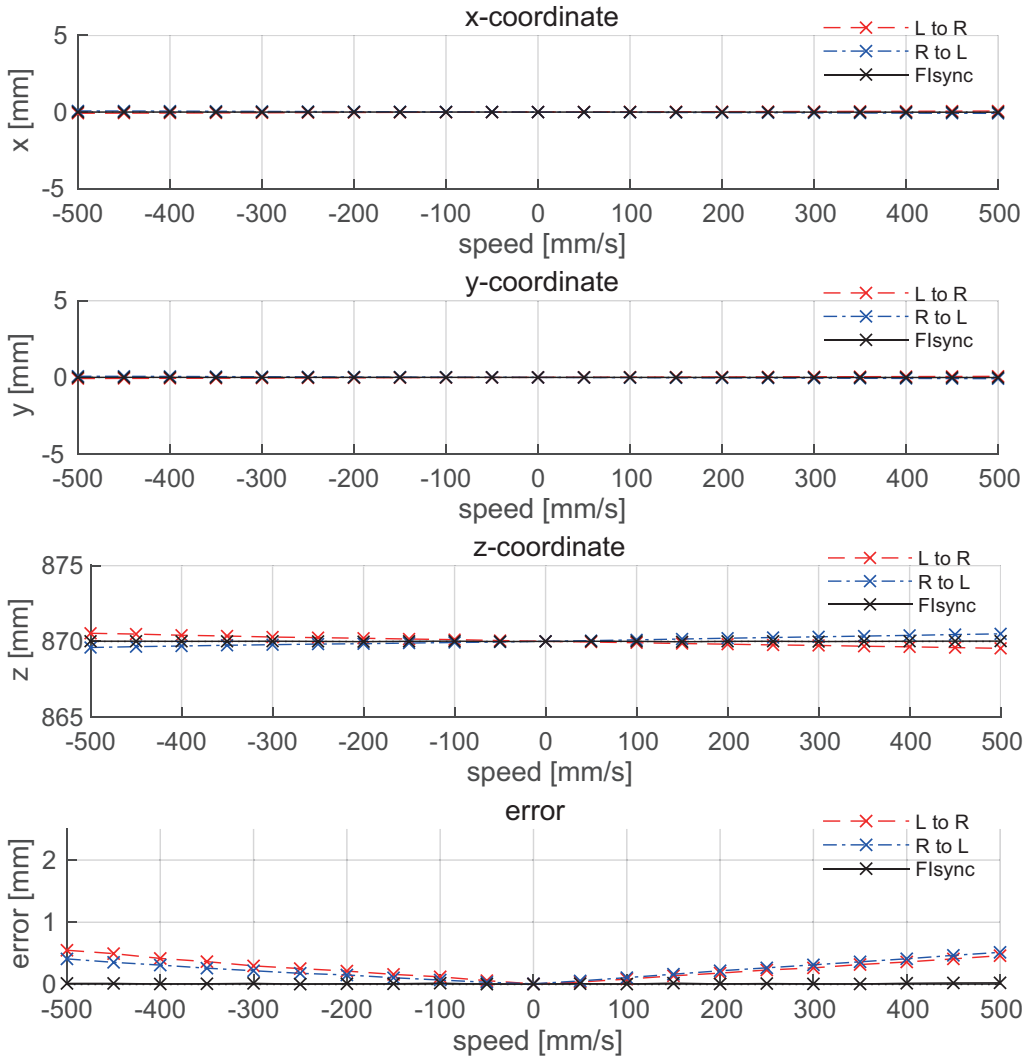


Figure 5.4: Measured 3-D positions and errors when a marker moved at different velocities in the y direction.

5.3.2 Unidirectional Moving Marker at Different Velocities

Next, the 3-D positions of a moving marker at different velocities were measured at (x, y, z) = (0, 0, 870 mm). The same marker mentioned in the previous subsection was conveyed by a linear slider. In the experiments, one pair of virtual stereo tracking cameras functioned, and the virtual left- and right-view images were tracked by setting the same parameters used in Subsection 5.3.1. Figures 5.3, 5.4, and 5.5 show the measured 3-D positions and their errors compared with the actual position (0, 0, 870 mm) when the marker moved at different speeds ranging from −500 to 500 mm/s in steps of 50 mm/s

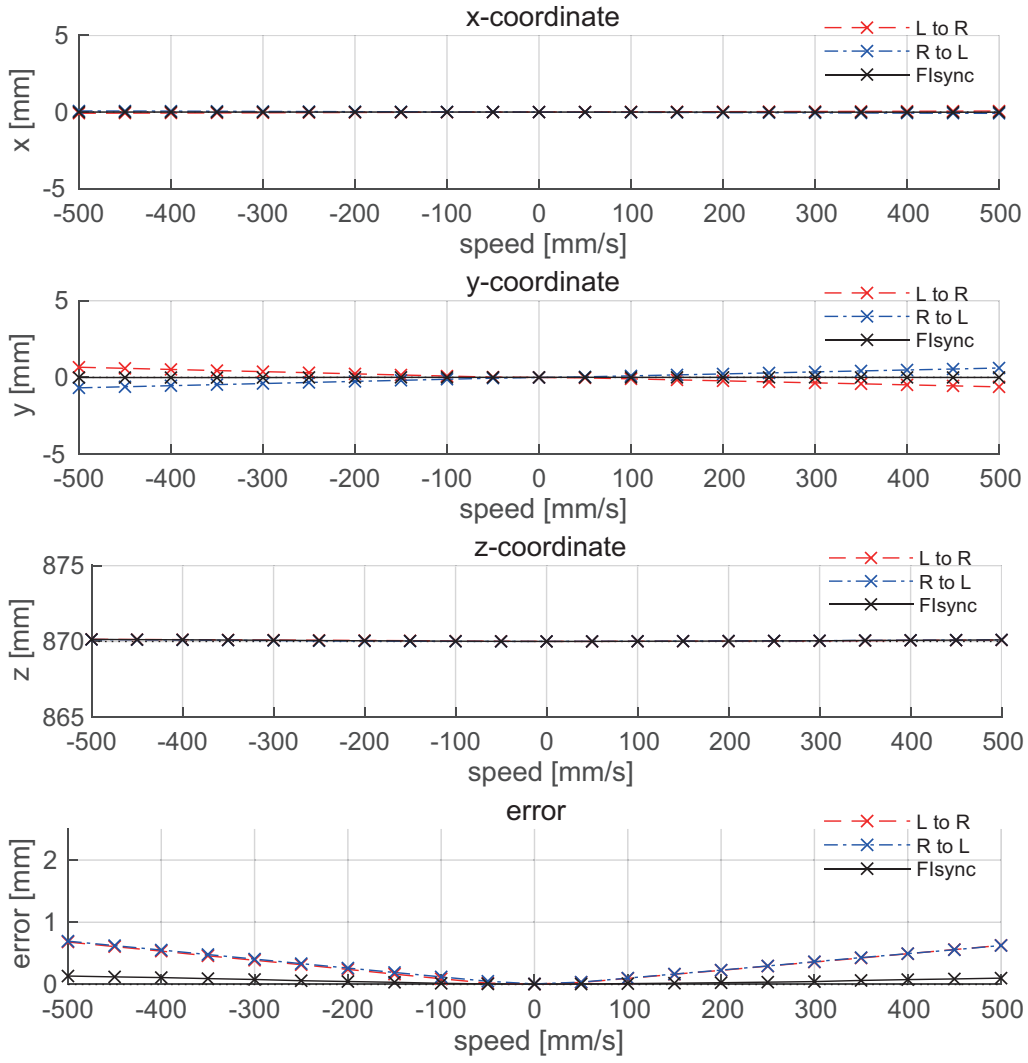


Figure 5.5: Measured 3-D positions and errors when a marker moved at different velocities in the z direction.

in the x-, y-, and z-directions, respectively. In these figures, the we compare the 3-D positions measured by virtual synchronization of the left- and right-views using frame interpolation ("FIsync"), and those measured without virtual synchronization when: (1) switching the left- to right-view image with a 2-ms delay ("L to R") and (2) switching the right- to left-view with a 2-ms delay ("R to L").

In all the experiments conducted at different speeds, in the y- and z-directions, there were minor differences between the 3-D positions measured by the "FI_Sync," "L to R," and "R to L" methods and the actual position (0, 0, 870 mm), whereas there were slight

deviations between the measured 3-D positions and the actual position at different speeds in the x direction according to the "L to R" and "R to L" measurements obtained without virtual synchronization. These differences can be explained by the marker's movement in the x direction leading to much larger synchronization errors between the virtual left and right pan-tilt cameras between frames compared with those when observing the marker moving in the y or z directions. Using ultrafast left- and right-view switching at 2 ms, the differences in the "L to R" and "R to L" measurements obtained at different speeds in the x direction were not as large according to Figure 5.3 and they increased slightly in proportion to the marker's speed, where the maximum deviation was 3.2 mm when the marker moved at 500 mm/s. By contrast, there were only small differences between the 3-D positions measured using the "FIsync" method and the actual position at different speeds in the x direction. This indicates that our catadioptric stereo tracking system could greatly reduce the synchronization errors after introducing virtual synchronization with frame interpolation as well as ultrafast left- and right-view switching.

5.3.3 Two Rotating Markers in 3-D Space

Next, we conducted experiments to determine the measured 3-D positions of two markers rotating at different speeds in 3-D space. The two markers ("M1" and "M2") had the same pattern as that mentioned in the previous subsections. The rotation center of "M1" was 185 mm distant from that of "M2". Both of "M1" and "M2" rotated on 33-mm radius circular trajectories on the common slanted plane at 2.67 rps and 0.56 rps, respectively; the plane of rotation was angled at 45.5 degrees with respect to the xz-plane. In the experiments, two pairs of virtual stereo tracking cameras functioned. ($J = 2$), and the number of markers for extraction was $I_0 = I_1 = 1$; the other parameters were set to the same parameters used in Subsection 5.3.1. The 3-D positions of markers could be extracted in real time at 125 fps. Figure 5.6 and Figure 5.7 shows the measured x, y, and z-coordinate values, the pan and tilt angles of the virtual pan-tilt cameras, and the image centroids of the left- and right-view images for $t = 0 – 1.8$ s. Figure 5.8 shows the 3-D trajectories for "M1" and "M2" for $t = 0 – 1.8$ s.

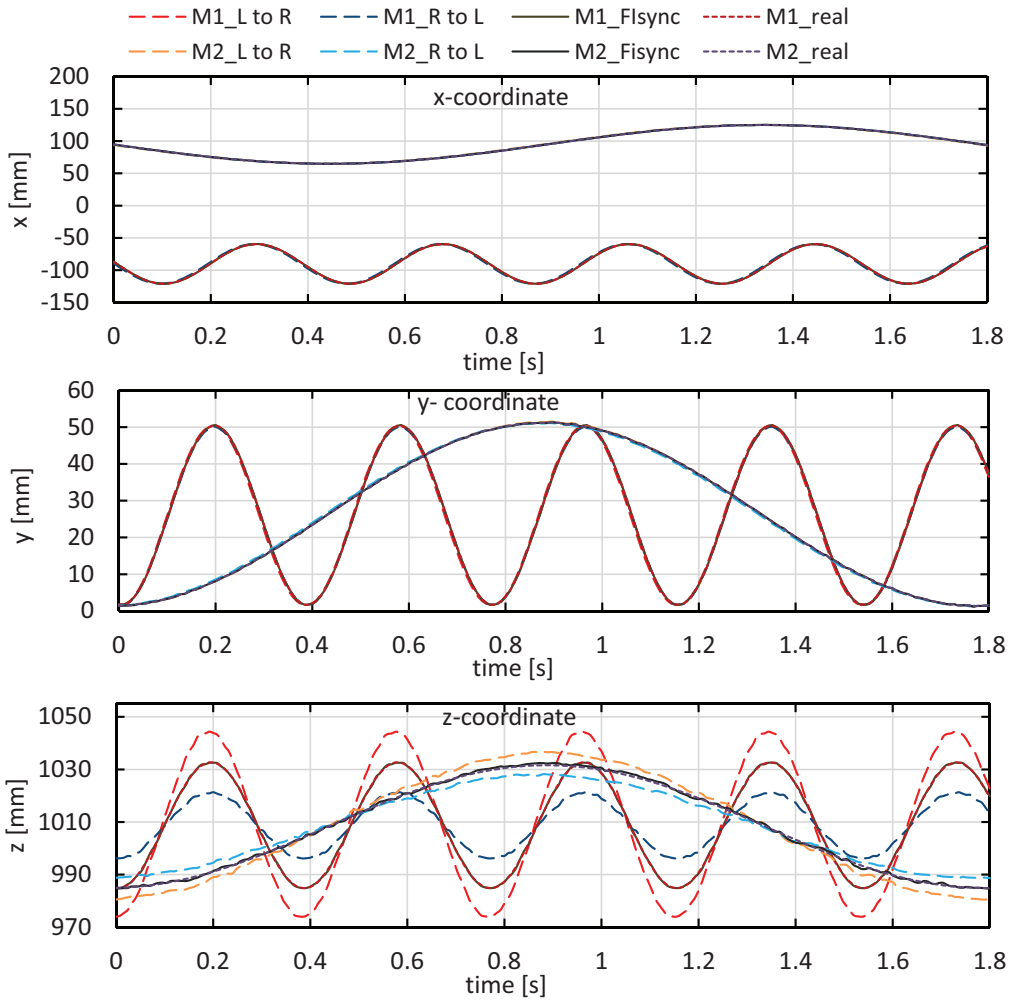
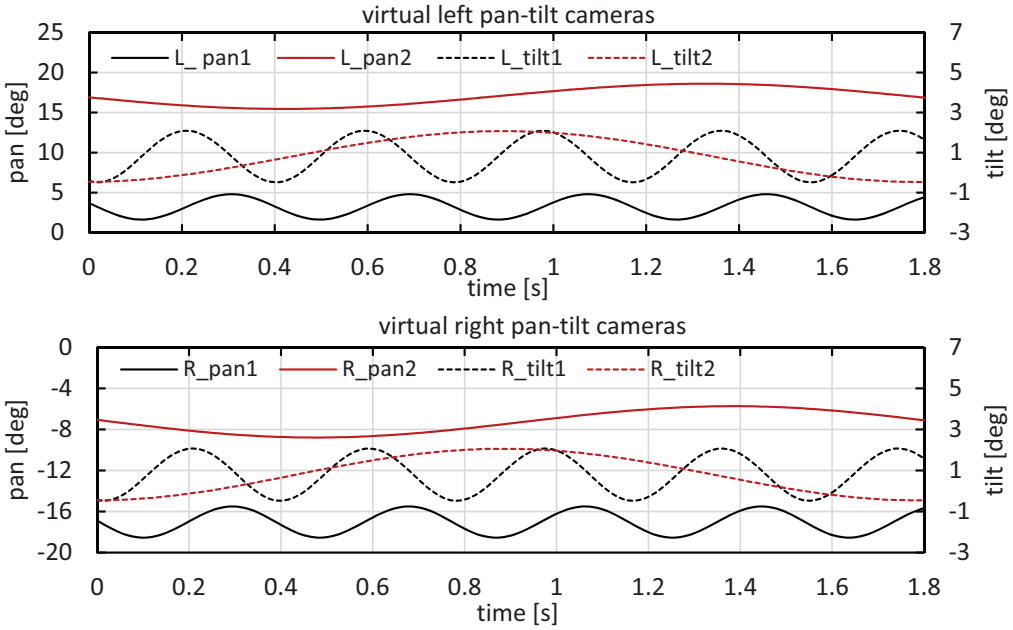
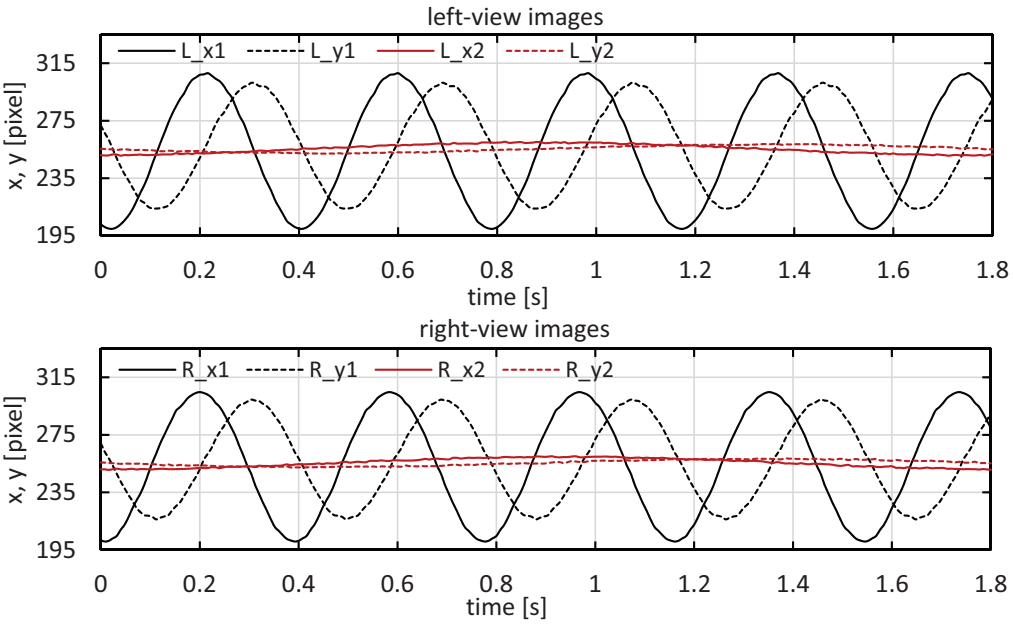


Figure 5.6: Measured x, y, and z-coordinate values for two rotating markers.

The image centroids of the left- and right-view images were tracked in the field of 512×512-pixel camera views by controlling the pan and tilt angles of the four virtual pan-tilt cameras, where they deviated from the image centers $(255, 255)$ and the maximum deviation of image centroids for "M1" and "M2" were 53.0 and 5.0 pixel, respectively; the deviation increased as the rotation speed became large. The x- and y-coordinate values measured using the "L to R" and "R to L" methods were similar to the actual values, but the measured z-coordinate value deviated slightly from its actual value. By contrast, there was little difference between the 3-D trajectories measured using the "FIsync" method and the actual trajectories of "M1" and "M2". Figure 5.9 shows the relationship between the velocity components in the x-, y-, and z-directions, and the deviations from the actual



(a) Pan and tilt angles



(b) Image centroids

Figure 5.7: Pan and tilt angles of the virtual left and right pan-tilt cameras, and image centroids for two rotating markers.

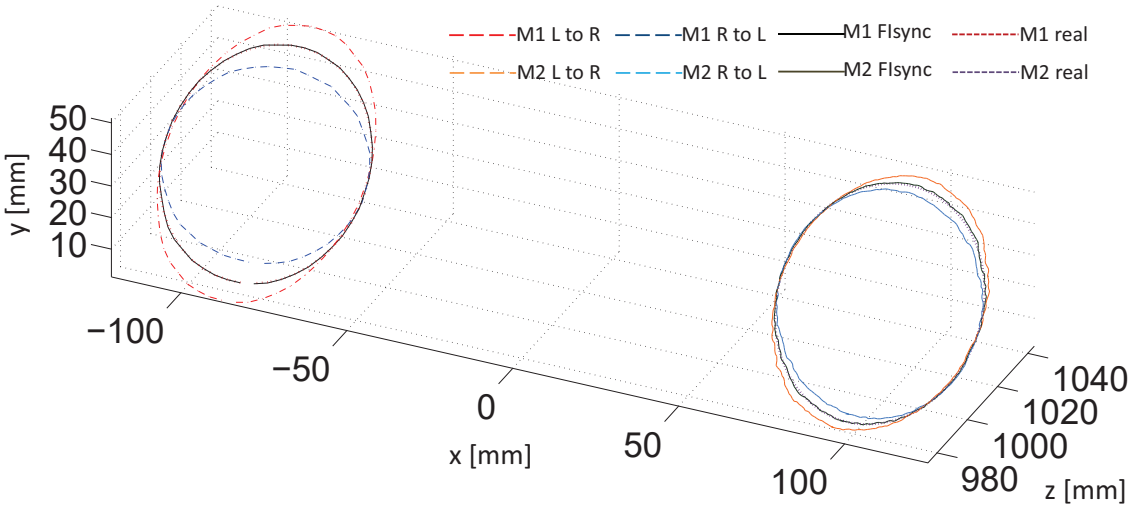


Figure 5.8: 3-D trajectories of two rotating markers.

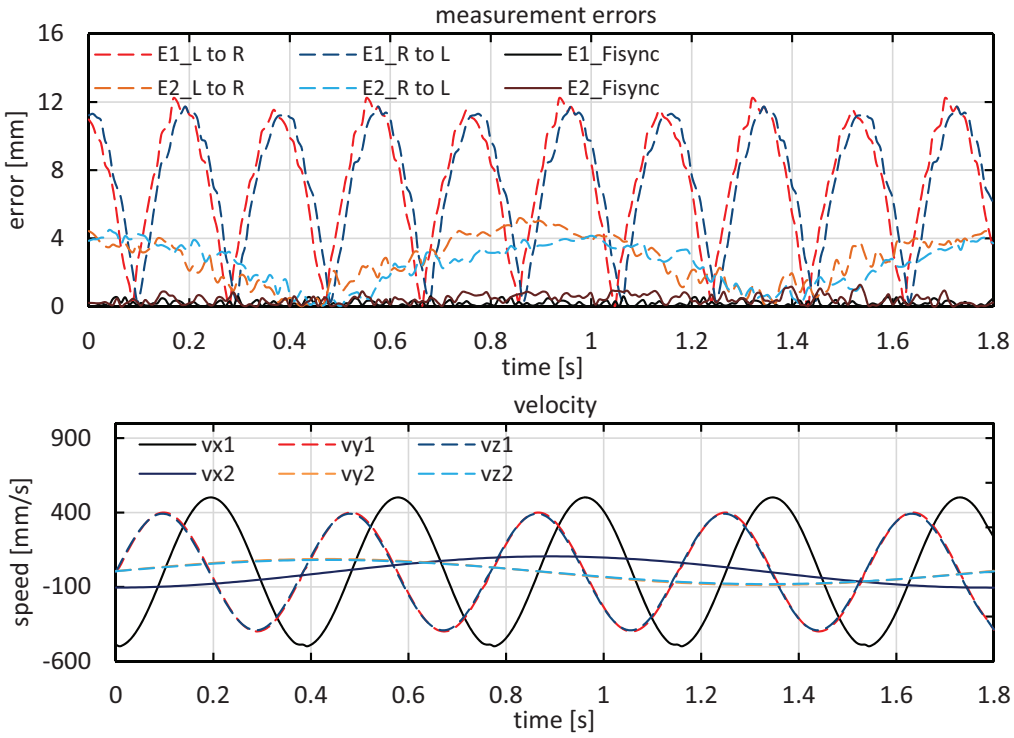


Figure 5.9: Relationships between deviations from the actual trajectories and the marker velocities.

trajectories for $t = 0 - 1.8$ s. The deviations in the "L to R" and "R to L" measurements increased when the amplitude of the velocity component in the x direction increased, but there was little deviation in the "FIsync" measurements. The maximum deviations using the "L to R", "R to L", and "FIsync" methods were 12.2, 11.8, and 0.6 mm for "M1", and 5.1, 4.5, and 0.8 mm for "M2", respectively. This indicates that our catadioptric stereo tracking system could accurately measure the 3-D positions of two markers rotating at different speeds without synchronization errors in real time when the moving markers were not tracked perfectly in the centers of left- and right-view images.

5.3.4 Two Dancing Dolls with Multiple Markers

Finally, two objects with multiple markers were simultaneously tracked, and the 3-D positions of markers were obtained via virtual synchronization ("FI" measurement). Two dancing horse models ("object 1" and "object 2") measuring 63×25×100 mm, as illustrated in Figure 5.10, were measured in the experiments; 18 white circular markers with a diameter of 3 mm were attached to each horse model. Two pairs of virtual stereo tracking cameras functioned, ($J = 2$), and the number of markers for extraction was $I_0 = I_1 = 18$.

The thresholds for binarization and flag-map generation were set to $B_\theta = 70$ and $F_\theta = 0$ pixel, respectively. The 3-D positions of markers could be extracted in real time at 125 fps. "Object 1" was placed at a distance of 170 mm from "object 2," as illustrated in Figure 5.10; the markers on the body centers of "object 1" and "object 2" were initially located on ${}^0P_1(-60, 45, 960 \text{ mm})$ and ${}^1P_1(110, 45, 940 \text{ mm})$, respectively, at time $t = 0$. "Object 1" moved without dancing at 100 mm/s in the z-direction from $z = 960$ to 1,060 mm by a linear slider, and "object 2" danced while shaking its body and legs at a frequency of approximately 2.0 Hz at the same place without any translation.

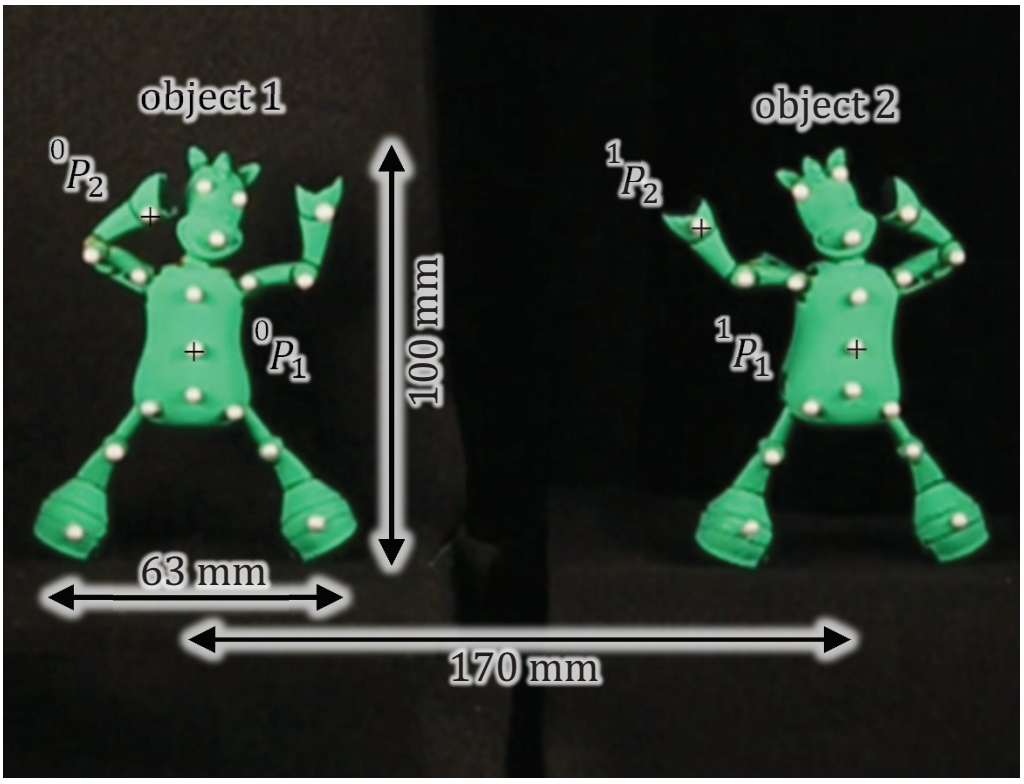
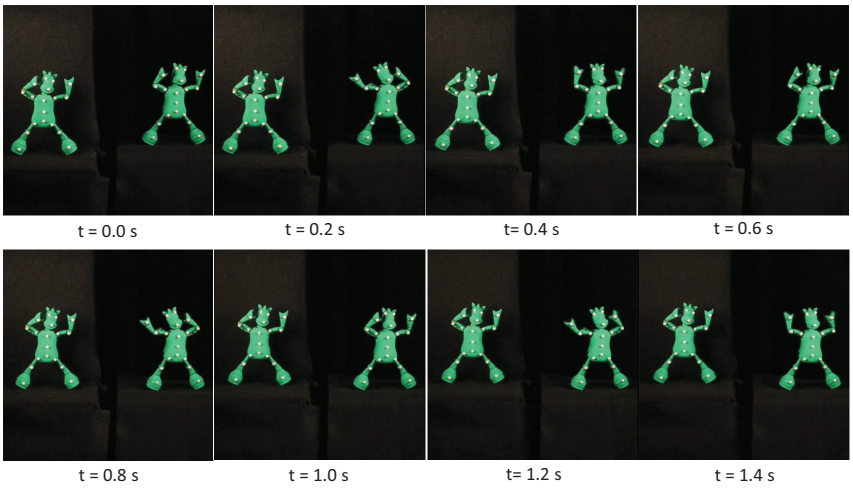
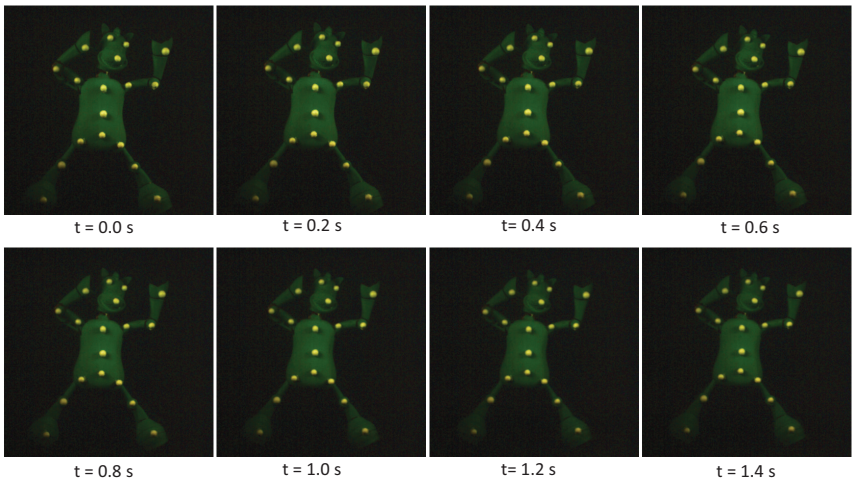


Figure 5.10: Observation of two horse models.

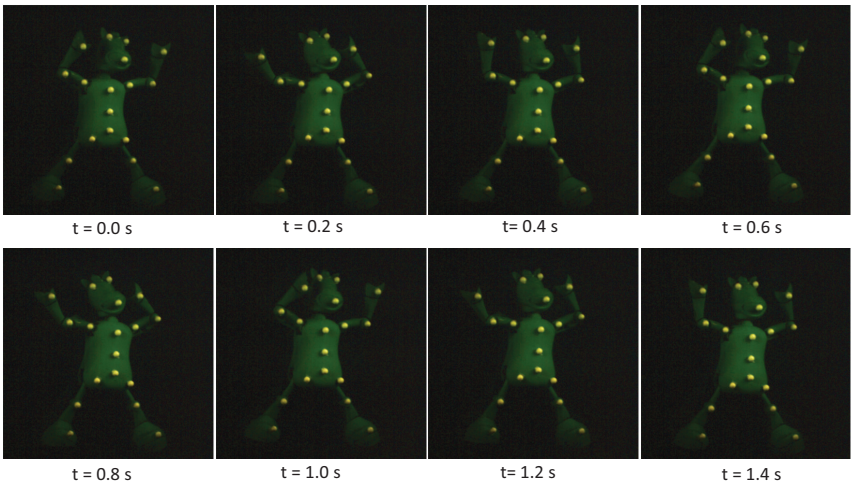
Figure 5.11 shows (a) an overview of the experimentally monitored images using a standard video camera at a fixed position, (b) the left-view images of "object 1", and (c) the left-view images of "object 2," which were taken at intervals of 0.2 s.



(a) Experimental overview



(b) Left-view images of “object 1”



(c) Left-view images of “object 2”

Figure 5.11: Experimental overview, and captured left-view images of “object 1” and “object 2”.

Figure 5.12 shows the measured 3-D positions of the markers attached to "object 1" and "object 2." Figure 5.13 shows the 3-D positions of the markers ${}^0P_1$, ${}^0P_2$, ${}^1P_1$, and ${}^1P_2$ during 1.4 s; markers, ${}^0P_1$ and ${}^1P_1$, were located on the body centers of the horse models, and markers, ${}^0P_2$ and ${}^1P_2$, were located on their legs, as shown in the Figure 5.10. Figures 5.14 and 5.15 show (a) the left and right centroid positions, and (b) pan and tilt angles of the left and right virtual pan-tilt cameras during 1.4 s, when observing "object 1" and "object 2", respectively.

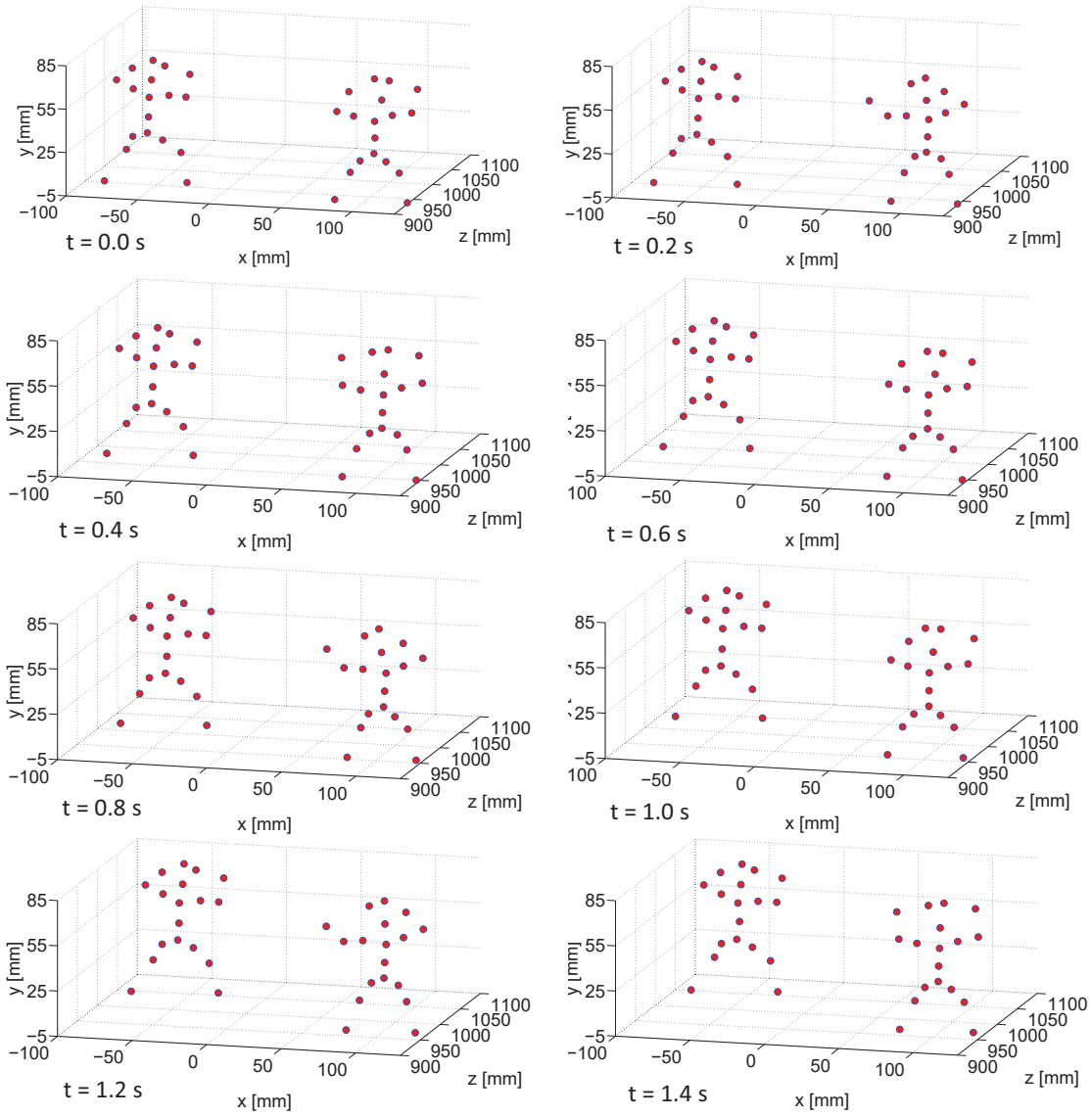


Figure 5.12: Measured 3-D positions of markers attached on "object 1" and "object 2".

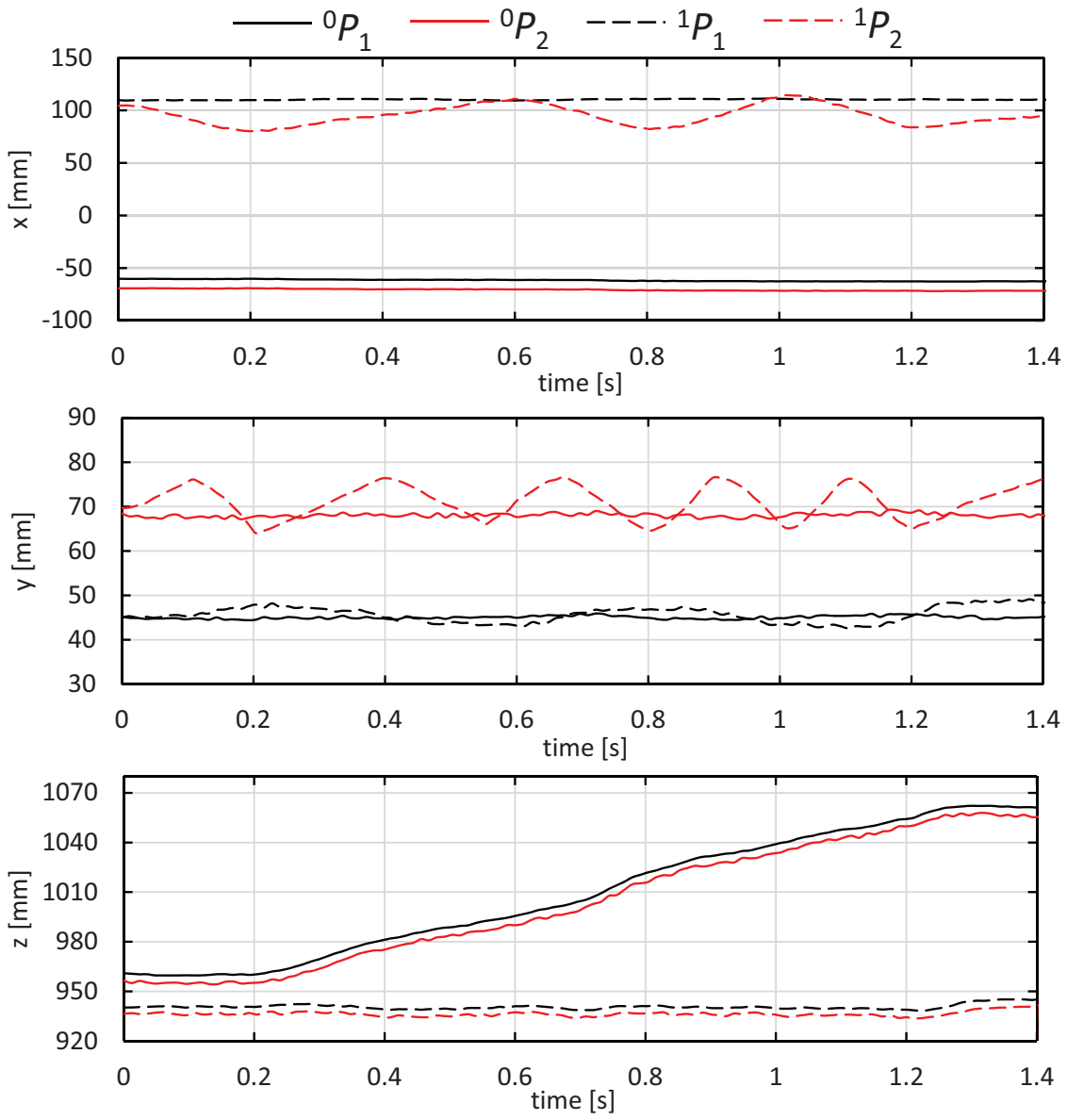
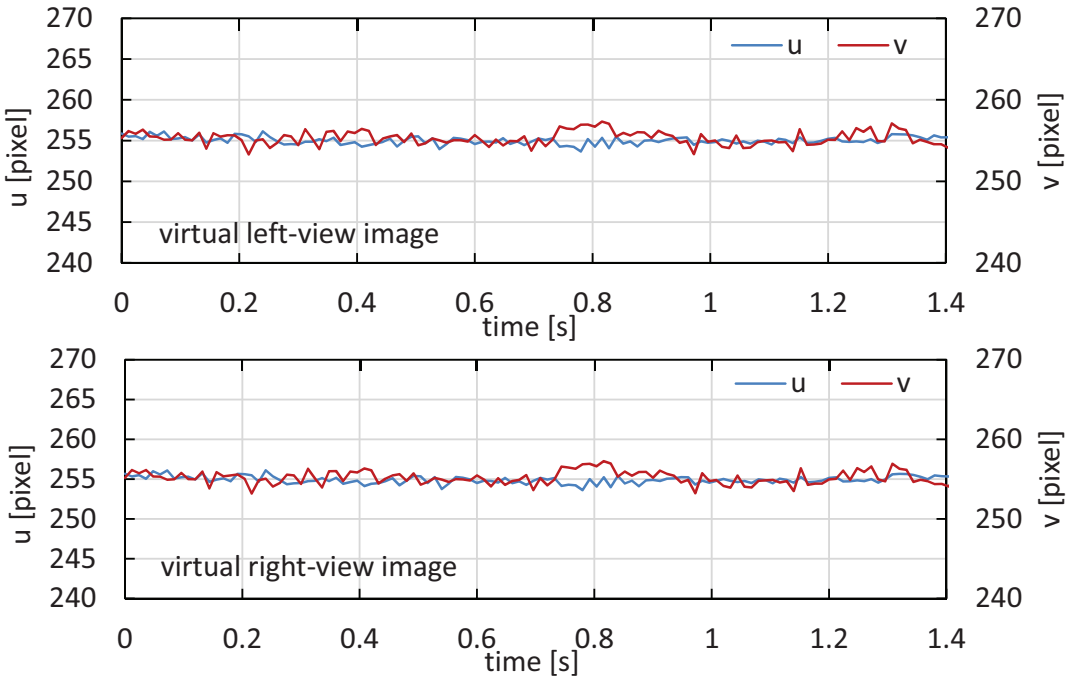
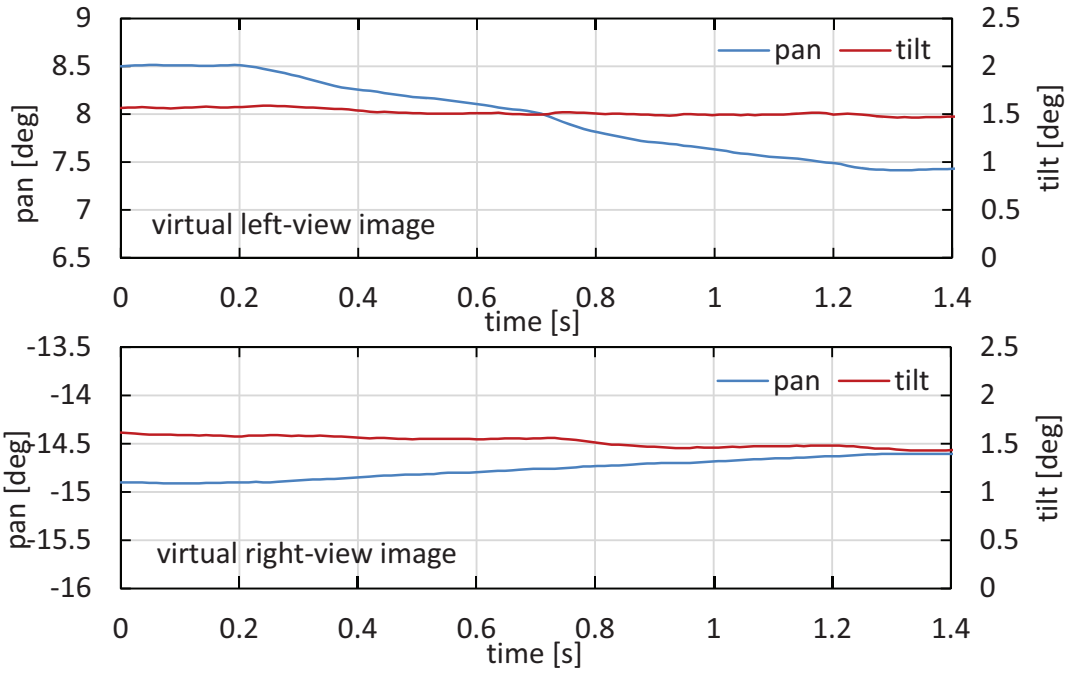


Figure 5.13: Measured 3-D positions of markers $^0P_1$, $^0P_2$, $^1P_1$, and $^1P_2$.

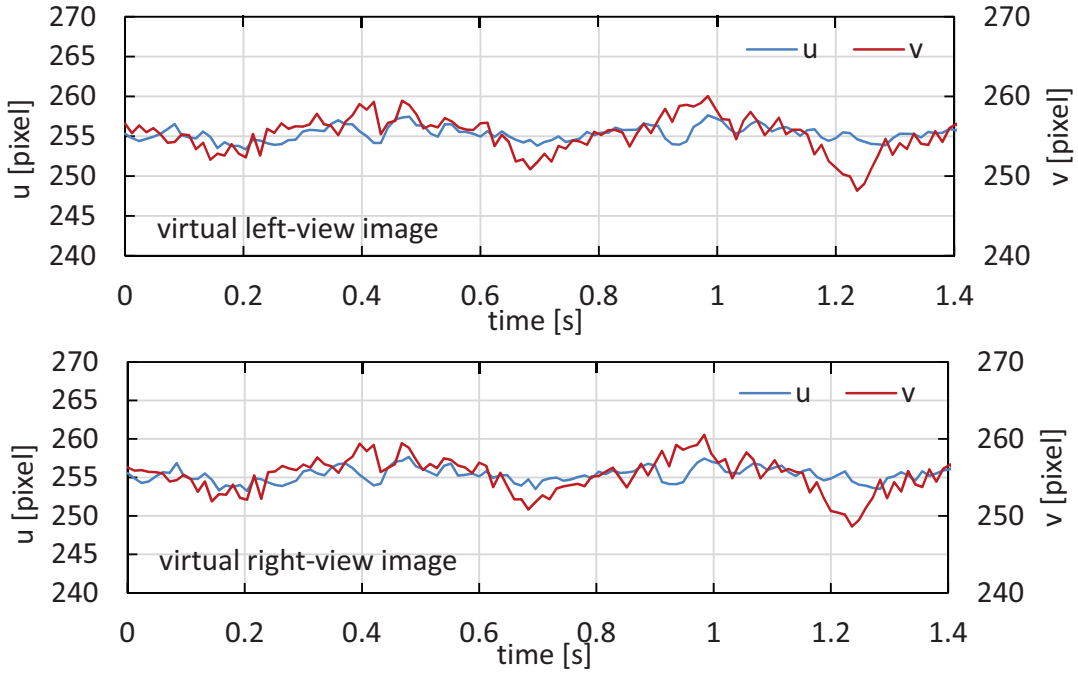


(b) Image centroids

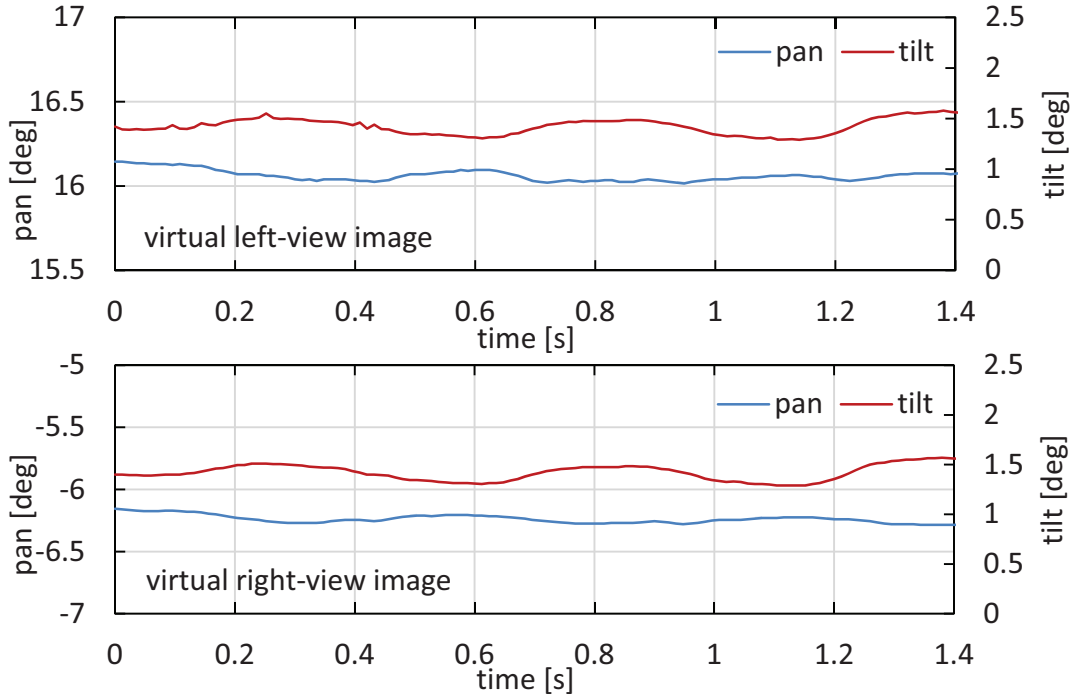


(c) Pan and tilt angles

Figure 5.14: Image centroids, and pan and tilt angles of the virtual left and right pan-tilt cameras for "object 1".



(b) Image centroids



(c) Pan and tilt angles

Figure 5.15: Image centroids, and pan and tilt angles of the virtual left and right pan-tilt cameras for "object 2".

It can be seen that both of "object 1" and "object 2" were always tracked in the left- and right-view images of two virtual stereo tracking cameras by controlling the pan and tilt angles so that the *uv* centroids were located around $(255, 255)$ and $(255, 255)$, respectively. The 3-D positions of the markers attached to the two horse models were measured accurately when "object 1" moved at 100 mm/s in the z-direction without dancing, whereas the body and legs of "object 2" moved rapidly at different speeds without translation. For "object 1", the z-coordinate values measured at points ${}^0P_1$ and ${}^0P_2$ varied from $z = 960$ to 1060 mm and $z = 956$ to 1056 mm, respectively, whereas the x- and y-coordinate values measured at points, ${}^0P_1$ and ${}^0P_2$, did not vary largely, because of no local body and leg motion. For "object 2", the z-coordinate values, measured at points, ${}^1P_1$ and ${}^1P_2$, were always around 940 mm and 936 mm respectively. The x- and y- coordinate values measured at point, ${}^1P_1$, were always around $x = 110$ mm and $y = 68$ mm, whereas those measured at point, ${}^1P_2$, varied periodically in the range of $x = 80$ to 115 mm and $y = 64$ to 76 mm, according to the left-and-right shaking of the legs.

Thus, our catadioptric stereo tracking system measured the 3-D positions of markers on two fast-moving objects in real time at 125 fps with zooming on the fields of view observed by two pairs of virtual left and right pan-tilt cameras, even when the markers were attached to time-varying-shape objects where the different parts moved at variable speeds.

5.4 Concluding Remarks

In this Chapter, we developed a real-time monocular 3-D tracking system that can function as J virtual stereo tracking cameras for capturing the 3-D positions of markers attached to multiple moving objects at $250/J$ fps by accelerating gaze control using an ultrafast catadioptric stereo tracking system. The virtual left and right pan-tilt cameras can be employed for stereo tracking with multithread gaze control by switching among 500 different views in one second via a catadioptric mirror system, with concurrent real-time video processing for marker extraction at 500 fps. Our experimental results demonstrated the effective performance of our catadioptric stereo tracking system for real-time monoc-

ular stereo tracking of multiple objects without decreasing measurement accuracy. At present, our catadioptric stereo tracking system is constrained by insufficient light intensity due to the small size of the mirror for ultrafast viewpoint switching. Thus, in order to facilitate sensitive stereo vision for real-world sensing applications, we plan to improve the efficiency of light collection by the optical system as well as implementing real-time monocular markerless 3-D motion capture functions.

Chapter 6

Monocular Wide Baseline Stereo Measurement Using High-speed Catadioptric System

6.1 Introduction

In this Chapter, we propose a novel monocular wide baseline stereo vision system based on an ultrafast mirror-drive pan-tilt active vision device that can simultaneously switch hundreds of different views in one second. By combining high-speed video-shooting and multithread ultrafast gaze control processing, one camera can be performed as two virtual cameras with different views. Stereo image pairs can be captured by two virtual cameras frame by frame at arbitrary viewpoints by adding auxiliary catadioptric equipment and designing suitable way of pan-tilt mirror device switching. In this study, two additional plane mirrors are used to make up a wide stereo system in a limited space such as indoor environment. Frame interpolation is used to eliminate the unsynchronization problem for moving object.

By switching 125 different views in a second with real-time image recording at 125 fps, the proposed system can function virtually as two cameras operating at 62.5 fps that can capture stereo pairs of 8-bit 512×512 color images. 3-D measurement experiment and analysis are demonstrated to verify the effectiveness of the proposed monocular wide baseline stereo system. In this Chapter, we developed a catadioptric monocular stereo vision system based on an ultrafast pan-tilt mirror device, which can function as two virtual stereo cameras with wide and easily adjusted baseline even in limited space. Experimental results obtained for depth estimation verify efficacy of the proposed system.

6.2 Geometry of Monocular Wide Baseline Stereo Vision System

The geometry of the proposed wide baseline stereo system that uses a pan-tilt device with single high-speed camera and two additional plane mirrors is described as shown in Figure 3.4. The positions and orientations of the two virtual cameras, as well as the value of baseline can be calculated using the geometrical relationship. The parameters of the virtual camera reflected on a pan-tilt mirror and a catadioptric mirror system can be described using mirror reflection. The relationship between the real camera and its virtual camera reflected by one planar mirror can be described as follows:

$$\begin{pmatrix} p_{i+1} \\ 1 \end{pmatrix} = \boldsymbol{P}_i \begin{pmatrix} p_i \\ 1 \end{pmatrix}, \tag{6.1}$$

where p_i is the optical center of the real camera, and p_{i+1} is the optical center of the mirrored virtual camera, which is the reflection of the real camera view on the mirror plane. $\boldsymbol{P}_i$ is the homogeneous transformation matrix made up of the normal vector of mirror $\boldsymbol{n}_i$ and one point $\boldsymbol{a}_i$ on the mirror plane. $\boldsymbol{P}_i$ can be expressed as follows:

$$\boldsymbol{P}_i = \begin{pmatrix} \boldsymbol{I} - 2\boldsymbol{n}_i\boldsymbol{n}_i^T & 2(\boldsymbol{n}_i\boldsymbol{n}_i^T)\boldsymbol{a}_i \\ 0 & 1 \end{pmatrix} \tag{6.2}$$

where $\boldsymbol{I}$ is the 3×3 identity matrix.

6.2.1 Geometrical Definitions of Mirrors

The pan-tilt mirror system has two movable mirrors in the pan and tilt directions and the geometrical relationship is shown as Figure 6.1. The center of the pan mirror is set to $\boldsymbol{a}_\alpha = (0, 0, 0)^T$, which is the origin of the xyz-coordinate system. The pan mirror can rotate around the z-axis, and its normal vector is given as $\boldsymbol{n}_\alpha = (-\cos\alpha, \sin\alpha, 0)^T$. The center of the tilt mirror (mirror 2) is located at $\boldsymbol{a}_\beta = (0, d, 0)^T$, where its distance from that of the pan mirror is represented by d. The tilt mirror can rotate around a straight line parallel to

the x-axis at a distance d, and its normal vector is given as $\boldsymbol{n}_\beta = (0, -\sin\beta, \cos\beta)^T$. α and β indicate the pan and tilt angles of the pan-tilt mirror system, respectively. The center of high-speed camera is installed to $\boldsymbol{p}_1 = (-l, 0, 0)^T$, where l is the distance from the center of the pan mirror.

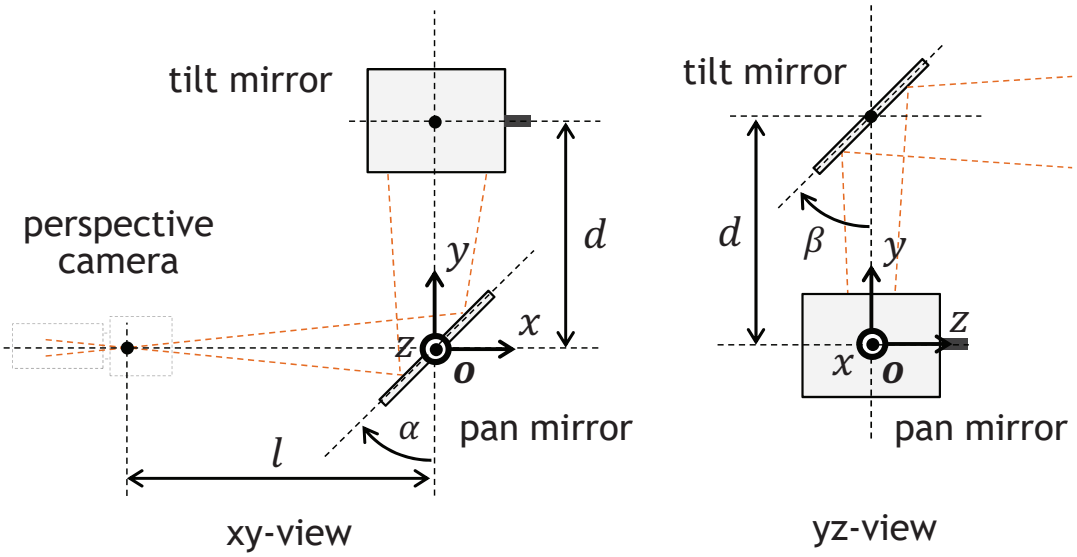


Figure 6.1: Geometries of pan-tilt mirror system.

The catadioptric mirror system with two planar mirrors is installed in front of the pan-tilt mirror paralleling to the y-axis. Figure 6.2 shows the xz-view of its geometry. The normal vectors of the mirror 1 and mirror 2 are given as $\boldsymbol{n}_1 = (\cos\theta_1, 0, -\sin\theta_1)^T$, $\boldsymbol{n}_2 = (-\cos\theta_2, 0, -\sin\theta_2)^T$, respectively. θ_1 and θ_2 are the angles formed by the xy-plane and the planes of mirrors 1 and 2. The planes of mirrors 1 and 2 are crossed on the yz-plane passing through the points $\boldsymbol{a}_1 = (0, d, h)(= \boldsymbol{a}_2)$ in front of the center of the tilt mirror.

6.2.2 Parameters of Virtual Cameras and Baseline

According to the geometric parameters of the pan-tilt mirror system and the two side mirrors, the optical center of the virtual cameras $\boldsymbol{p}_v$ can be expressed with reflection

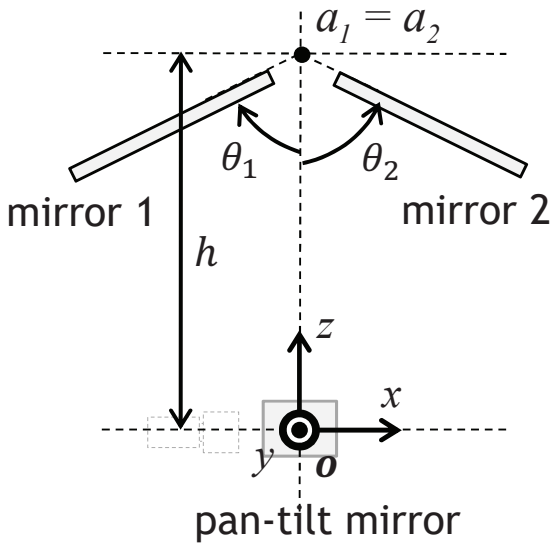


Figure 6.2: Geometries of catadioptric mirror system system.

transformation as follows:

$$\begin{pmatrix} \boldsymbol{p}_v \\ 1 \end{pmatrix} = \boldsymbol{P}_s \boldsymbol{P}_\beta \boldsymbol{P}_\alpha \begin{pmatrix} \boldsymbol{p}_1 \\ 1 \end{pmatrix}, \tag{6.3}$$

where $\boldsymbol{P}_\alpha, \boldsymbol{P}_\beta$ and $\boldsymbol{P}_s$ are the transformation matrix of pan, tilt and side mirrors, respectively.

The optical center $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ of the virtual camera 1 and 2 are described by the following functions related to the variable pan and tilt angles of α and β.

$$\boldsymbol{p}_1 = \begin{pmatrix} -l\cos 2\alpha \cos 2\theta_1 - h \sin 2\theta_1 - A \sin 2\beta \sin 2\theta_1 \\ d - A\cos 2\beta \\ -A\cos 2\theta_1 \sin 2\beta + 2h\sin^2\theta_1 + l\cos 2\alpha \sin 2\theta_1 \end{pmatrix}, \tag{6.4}$$

$$\boldsymbol{p}_2 = \begin{pmatrix} -l\cos 2\alpha \cos 2\theta_2 + h \sin 2\theta_2 + A \sin 2\beta \sin 2\theta_2 \\ d - A\cos 2\beta \\ -A\cos 2\theta_2 \sin 2\beta + 2h\sin^2\theta_2 - l\cos 2\alpha \sin 2\theta_2 \end{pmatrix}, \tag{6.5}$$

where $A = l\sin 2\alpha + d$.

The virtual baseline D_b can be calculated using the locations of the two virtual cameras $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ expressed as $D_b(\alpha,\beta) = |\boldsymbol{p}_2 - \boldsymbol{p}_1|$. The baseline D_b can be calculated easily by adjusting the pan and tilt mirrors angles when other parameters such as θ_1, θ_2 and h are fixed.

6.3 Implemented Algorithm and System Configuration

In this section, we implement the algorithm to calculate the 3-D images using the virtual stereo pair of images captured through the ultrafast pan-tilt mirror system. The flowchart of implemented algorithm is illustrated in Figure 6.3 including multithread gaze control for stereo pair acquisition and depth calculation.

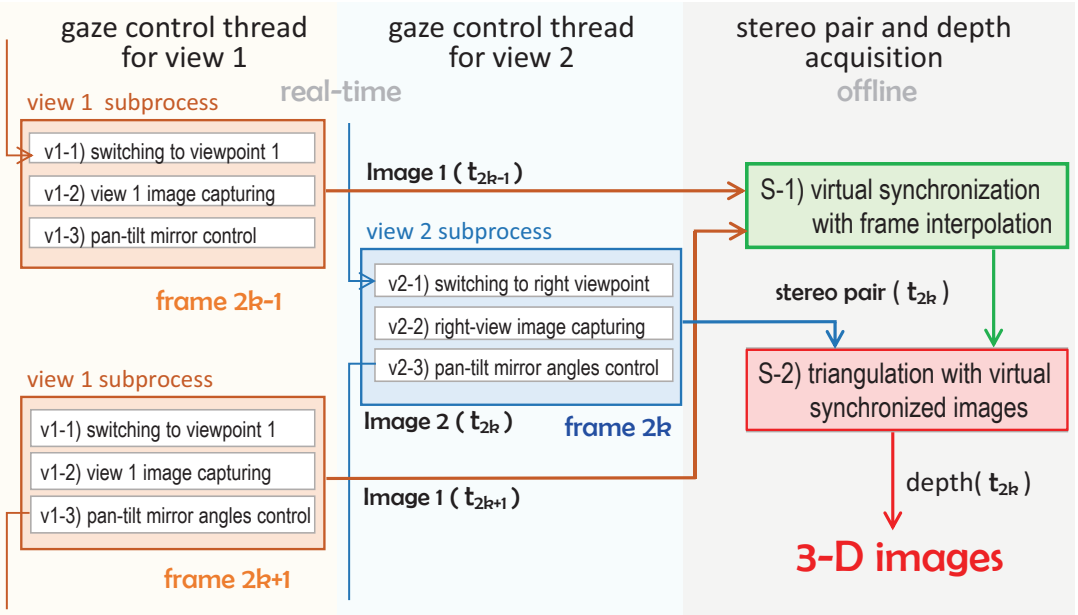


Figure 6.3: Flowchart of implemented algorithm.

6.3.1 Multithread Gaze Control for Stereo Pair Acquisition

The multithread gaze control processing is to coordinate high-speed video-shooting, processing and pan-tilt mirror switching. A fine temporal granularity for view 1 and view 2 is necessary to realize multiple virtual cameras using a single high-speed vision system. The accelerations of video-shooting and gaze control ensure the accurate and high-speed stereo pairs acquiring with sufficient large parallax. Compared with traditional

wide baseline stereo systems using two or more real cameras, the proposed system using virtual cameras has the advantages including no need to connect two real cameras, easier to adjust virtual camera positions, wide baseline stereo especially in limited space where two real cameras can't be placed and only one camera used.

Gaze directions for view 1 and view 2 are controlled according to the frame sequence of high-speed camera alternately. View 1 works for $t_{2k-1} - \tau_m \leq t < t_{2k} - \tau_m$, and view 2 works for $t_{2k} - \tau_m \leq t < t_{2k+1} - \tau_m$ (τ_m is the settling time in controlling the mirror angles of the pan-tilt mirror system). The time-division thread executes with a temporal granularity of Δt. $t_k = t_0 + k\Delta t$ (k: integer) indicates the image-capturing time of the high-speed vision system,

The images for view 1 and view 2 are captured at different timing leading to synchronization errors in stereo measurement especially when target object moves largely. In this study, we use frame interpolation technique [138] for virtual synchronization to eliminate unsynchronization caused by mirror switching. View 1 images were interpolated to correspond with view 2 images. Considering the view 2 image $I_2(t_{2k})$ captured at time t_{2k} as the standard image for virtual synchronization, the virtually synchronized view 1 image $\tilde{I}_1(t_{2k})$ at time t_{2k} can be estimated with frame interpolation using the two temporally neighboring view 1 images $I_1(t_{2k-1})$ at time t_{2k-1} and $I_1(t_{2k+1})$ at time t_{2k+1} as follows:

$$\tilde{I}_1(t_{2k}) = f_{FI}(I_1(t_{2k-1}), I_1(t_{2k+1})), \tag{6.6}$$

where $f_{FI}(I_a, I_b)$ indicates the frame interpolation function using two images I_a and I_b. Note that image horizontal flipping is essential because each side virtual camera captures images by three mirror reflections (pan-and-tilt mirrors and one side plane mirror), thus odd number of mirror reflection makes image mirroring.

6.3.2 Depth Estimation Using Virtual Synchronized Images

3-D depth information is estimated using virtually synchronized view 1 and view 2 images, $\tilde{I}_1(t_{2k})$ and $I_2(t_{2k})$, at time t_{2k} after mirroring. Figure 6.4 shows the flowchart

of depth estimation using stereo pairs. View 1 and view 2 stereo images are firstly rectified with stereo calibration parameters to make epipolar horizontal. Then, some standard stereo matching methods can be used to calculate the depth disparity map of the corresponding pixel. In this study, the rSGM method [139] are used as the stereo matching algorithm.

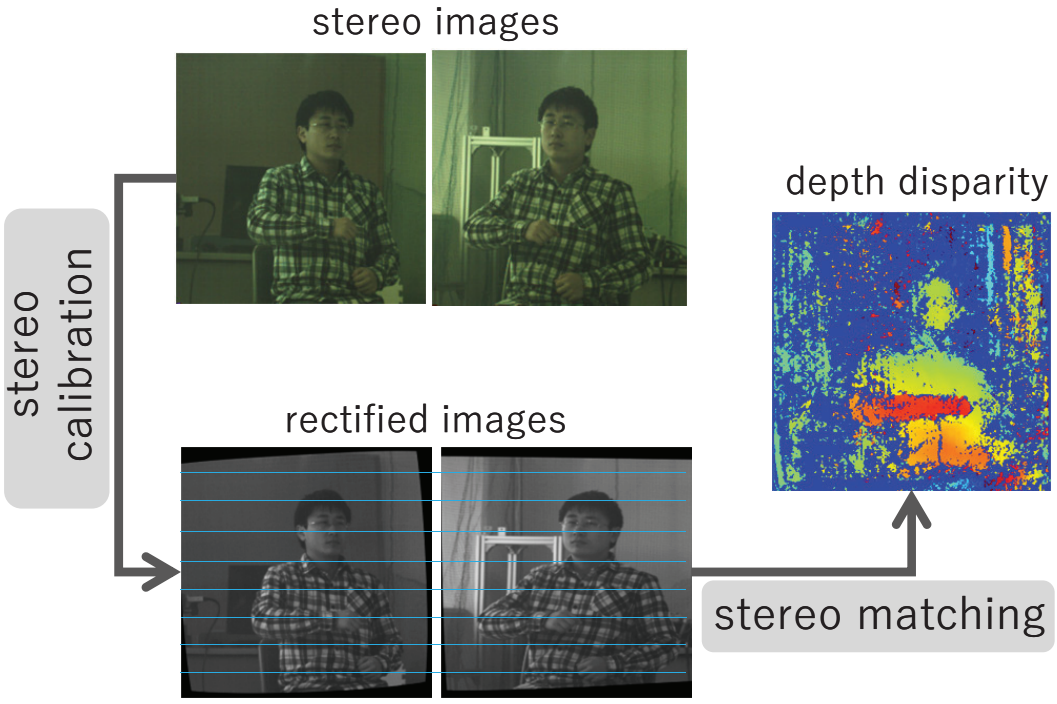


Figure 6.4: Flowchart of depth estimation using stereo pairs.

6.3.3 System Configuration

The proposed system is mainly made up of three parts: the high-speed vision part, catadioptric part and control system. The overview of the whole system is shown in Figure 4.3. The high-speed vision platform is IDP Express [78] consisting of a compact camera head and an FPGA (Xilinx XC3S5000) image processing board, which can capture 8-bit RGB images of 512×512 pixels at 2,000 fps. A f = 50 mm CCTV lens is attached to the camera head. The catadioptric system includes two-degrees-of-freedom (DOF) pan-tilt mirror and two additional side mirrors. The pan mirror was installed 25 mm in front of the CCTV lens, and the tilt mirror was installed 10 mm in front of the pan mirror.

The two side mirrors were vertically installed 5.4 m in front of the tilt mirror and the distances between two mirrors were 0.6 m and 2.9 m as shown in Figure 4.3. The control system includes a personal computer (PC) (Windows 7 Enterprise 64-bit OS, ASUS P6T7 WS Super Computer motherboard, Intel Core (TM) i7 3.20-GHz CPU, 6 GB memory) and A/D and D/A communication system. The D/A board (PEX-340416, Interface Inc, Japan) is used to send control signals to the pan-tilt mirrors (Cambridge Technology 6240H) and the A/D board (PEX-321216, Interface Inc, Japan) is used to receive signals of the pan and tilt angles. The pan and tilt mirrors are movable in the range of -10 to 10 degrees controlled within 2 ms.

6.4 Experiments

6.4.1 3-D Depth Measurement of Human Body

Figure 6.5 shows the human body 3-D measurement experiment using the proposed system. Human body and arm were moved during the measurement experiment and we selected images at the time of t=1.2 s, 2.5 s and 3.8 s. Figure 6.5 (a) is the overview of experiment monitored using a standard video camera. Figure 6.5 (b) and (c) show the captured images from view 1 and view 2 respectively after image mirroring and frame interpolation. Figure 6.5(d) shows the 3-D measurement results by using rSGM method after images rectification using stereo calibration parameters. 3-D depth of body and arm were measured even when the arm was moved. Experiment results demonstrate that the proposed monocular stereo system can work effectively by using high-speed camera and controlling the ultrafast pan-tilt mirror device switching.

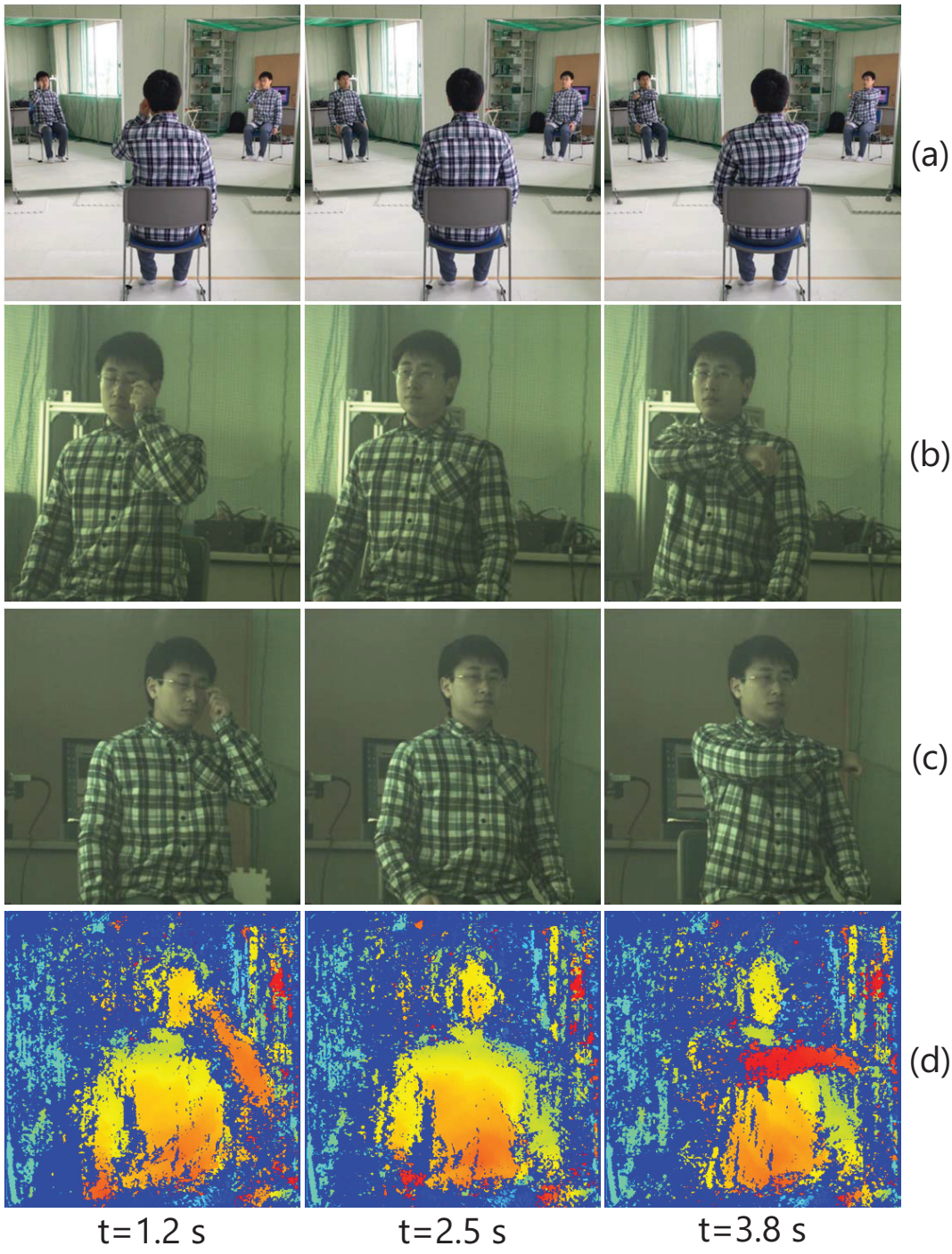


Figure 6.5: 3-D measurement experiment: (a) experiment overviews, (b) captured images from view 1, (c) captured images from view 2 and (d) depth images.

6.4.2 Experiment Analysis

Next, we compared the 3-D measurement experiment results with frame interpolation and without frame interpolation. Figure 6.6 shows the experiment results. In the experiment, two continuous captured view 1 images I_{1p} ($t = 2k - 1$) and I_{1n} ($t = 2k + 1$) were selected to make the frame interpolation image I_{FI} (virtually at $t = 2k$). View 2 image I_2 at $t = 2k$ were selected combining one of the three view 1 images as stereo pair to calculate the depth image.

In Figure 6.6, the top row shows three view 1 images: I_{1p}, I_{1n} and I_{FI}; the middle row is the view 2 images I_2 (the three images are the same one); the row below shows 3-D depth maps using different view 1 images (I_{1p}, I_{1n} and I_{FI}) and the same view 2 image I_2. The numbers of unmeasurable pixels in the 3-D images using images I_{1p} and I_{1n} were larger than that using frame interpolated image I_{FI}. This is because the deviation errors were larger without virtual synchronization when arm is moving. The experimental results indicate that the proposed wide baseline monocular stereo system can accurately measure the 3-D shapes of objects even in the movement case by using frame interpolation.

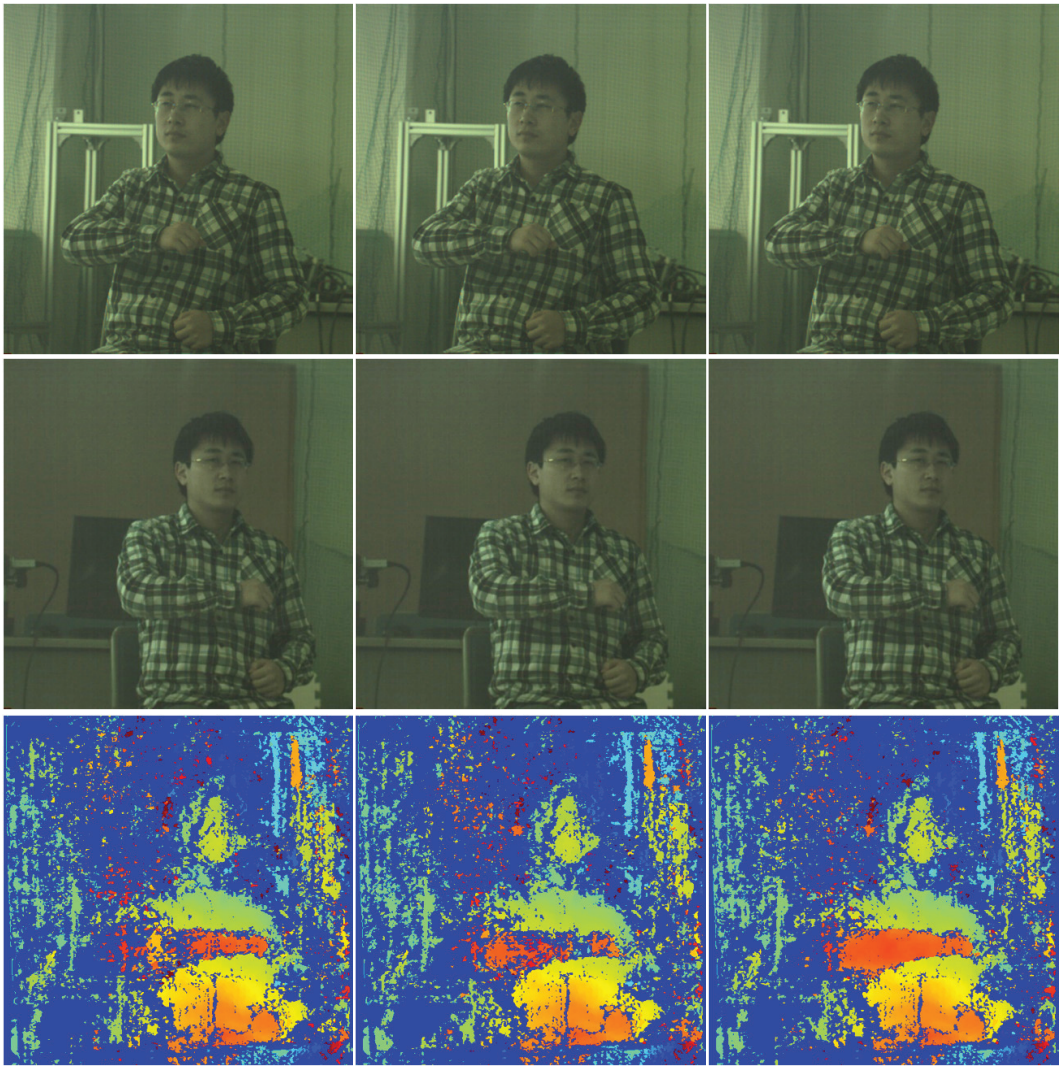


Figure 6.6: 3-D experiment with and without frame interpolation: (a) up: I_{1p} image; middle: I_2 image; down: depth result using I_{1p} and I_2, (b) up: I_{1n} image; middle: I_2 image; down: depth result using I_{1n} and I_2, (c) up: I_{FI} image; middle: I_2 image; down: depth result using I_{FI} and I_2.

6.5 Concluding Remarks

In this Chapter, we proposed a wide baseline catadioptric stereo system for monocular stereo measurement. High-speed vision system ad an ultrafast mirror-drive pan-tilt device switching different-view at 125 Hz in a second were used to capture 8-bit color

512×512 stereo pair images at 62.5 Hz. The geometry of the two virtual cameras, camera 1 and camera 2, for stereo measurement were presented to calculate their localizations. The baseline of two virtual cameras became wider and was able to adjust easily through pan and tilt angles. Frame interpolation strategy was proposed to reduce the synchronization errors caused by mirror switching in monocular stereo measurement. In the experiment, view 1 images were interpolated to be in coordinate with view 2 image. The 3-D measurement results and analysis were evaluated using the high-frame-rate videos with multithread gaze control, and verified the effectiveness of monocular stereo measurement using our monocular catadioptric stereo system with ultrafast viewpoint switching.

Chapter 7

Conclusion

In this study, we proposed the concept of novel high-speed monocular stereo vision system using the ultrafast pan-tilt catadioptric system. Based on this concept, we developed the monocular stereo vision system using only one camera to realize two cameras. Combining geometry relationship of the catadioptric system and flexible pan-tilt mirror device controlling, two or more active vision virtual stereo cameras can be obtained. In this way, target can be tracked and the stereo pair images are obtained at the same time by controlling the ultrafast pan-tilt mirror system. The parameters for three dimensional shape and reconstruction such as the positions and orientations of these virtual cameras are calculated using pan and tilt angles. Using this system, firstly we conducted experiment for 3-D shape measurement with tracking. In order to eliminate the synchronization errors generated by mirror switching, frame interpolation is introduced to make the stereo image pairs virtually captured at the same time. The stereo tracking and stereo image pair acquisition are running at the real-time, while the 3-D shape measurement result with frame interpolation is off-line because of the time consuming algorithms.

Then we applied the system to marker based 3-D motion tracking at the real-time. We expanded the system from one target stereo tracking to multiple objects stereo tracking by designing time sequence of the pan and tilt mirrors controlling. Hardware-based cell labelling algorithm was used to extract the positions of markers in each side image at the real-time. Interpolation is also needed for left and right switching synchronization, however frame interpolation is not suitable for real-time application. Thus we interpolated the centroid positions of each marker to make the marker positions virtually captured at

the same time. By using these strategies, 3-D motion of multiple targets can be tracked and obtained at real-time.

Finally, a wide baseline monocular stereo vision system was proposed to make it possible for the wider baseline stereo system realized even in the limited space. We redesigned the catadioptric system and calculated the geometry relationship of the mirrors. Larger mirrors were used to avoid insufficient light intensity due to the small-size mirror for ultrafast viewpoint switching. Stereo pair of human body movement was captured without additional illumination at real-time. Wide baseline stereo system can make the stereo measurement more accuracy.

In the future, we plan to expand the concept of virtual pan-tilt cameras to more applications. Currently, there are some problems to be solved such as offline 3-D image estimation due to the heavy computation required. To implement more sensitive and real-time stereo vision for real-world applications such as SLAM problems and large-scale mapping, we intend to improve both the optical system that maximizes the collection efficiency of light and the integrated monocular stereo algorithm that can be accelerated by parallel implementation of a local method with low computational complexity on GPUs and FPGAs. This is achieved by considering a trade-off between the complexity and accuracy in estimating stereo disparity maps on the basis of the monocular catadioptric stereo tracking system reported in this study.

Bibliography

[1] Scharstein, D.; Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42.

[2] Brown, M.Z.; Burschka, D.; Hager, G.D. Advances in computational stereo. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 993–1008.

[3] Lazaros, N.; Sirakoulis, G.C.; Gasteratos, A. Review of stereo vision algorithms: From software to hardware. *Int. J. Optomechatron.* **2008**, *2*, 435–462.

[4] Tombari, F.; Mattoccia, S.; Stefano, L.D.; Addimanda, E. Classification and evaluation of cost aggregation methods for stereo correspondence. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 24–26 June 2008; pp. 1–8.

[5] Herrera, P.J.; Pajares, G.; Guijarro, M.; Ruz, J.J.; Cruz, J.M. A stereovision matching strategy for images captured with fish-eye lenses in forest environments. *Sensors* **2011**, *11*, 1756–1783.

[6] Tippetts, B.; Lee, D.J.; Lillywhite, K.; Archibald, J. Review of stereo vision algorithms and their suitability for resource limited systems. *J. Real-Time Image Process.* **2013**, *11*, 5–25.

[7] Liu, J.; Li, C.; Fan, X.; Wang, Z. Reliable fusion of stereo matching and depth sensor for high quality dense depth maps. *Sensors* **2015**, *15*, 20894–20924.

[8] Hamzah, R.A.; Ibrahim, H. Literature survey on stereo vision disparity map algorithms. *J. Sensors* **2016**, *2016*, 8742920.

[9] Wang, L.; Yang, R.; Gong, M.; Liao, M. Real-time stereo using approximated joint bilateral filtering and dynamic programming. *J. Real-Time Image Process.* **2014**, *9*, 447–461.

[10] Sun, J.; Zheng, N.N.; Shum, H.Y. Stereo matching using belief propagation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 787–800.

[11] Yang, Q.; Wang, L.; Yang, R.; Wang, S.; Liao, M.; Nister, D. Real-time global stereo matching using hierarchical belief propagation. In Proceedings of the British Machine Vision Conference, Edinburgh, UK, 4–7 September 2006; pp. 989–998.

[12] Liang, C.K.; Cheng, C.C.; Lai, Y.C.; Chen, L.G.; Chen, H.H. Hardware-efficient belief propagation. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 525–537.

[13] Boykov,Y.; Veksler, O.; Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 1222–1239.

[14] Woodford, O.; Torr, P.; Reid, I.; Fitzgibbon, A. Global stereo reconstruction under second-order smoothness priors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 2115–2128.

[15] Yoon, K.J.; Kweon, I.S. Adaptive support-weight approach for correspondence search. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 650–656.

[16] Hosni, A.; Bleyer, M.; Gelautz, M. Secrets of adaptive support weight techniques for local stereo matching. *Comput. Vis. Image Underst.* **2013**, *117*, 620–632.

[17] Chen, D.; Ardabilian, M.; Chen, L. A fast trilateral filter-based adaptive support weight method for stereo matching. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 730–743.

[18] Veksler, O. Fast variable window for stereo correspondence using integral images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 16–22 June 2003; pp. 556–561.

[19] Xu, Y.; Zhao, Y.; Ji, M. Local stereo matching with adaptive shape support window based cost aggregation. *Appl. Opt.* **2014**, *53*, 6885–6892.

[20] McCullagh, B. Real-time disparity map computation using the cell broadband engine. *J. Real-Time Image Process.* **2012**, *7*, 87–93.

[21] Sinha, S.N.; Scharstein, D.; Szeliski, R. Efficient high-resolution stereo matching using local plane sweeps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 1582–1589.

[22] Yang, R.; Pollefeys, M. Multi-resolution real-time stereo on commodity graphics hardware. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 16–22 June 2003; pp. 211–217.

[23] Gong, M.; Yang, Y.H. Near Real-time reliable stereo matching using programmable graphics hardware. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; pp. 924–931.

[24] Grauer-Gray, S.; Kambhamettu, C. Hierarchical belief propagation to reduce search space Using CUDA for stereo and motion estimation. In Proceedings of the Workshop on Applications of Computer Vision, Snowbird, UT, USA, 7–8 December 2009; pp. 1–8.

[25] Humenberger, M.; Zinner, C.; Weber, M.; Kubinger, W.; Vincze, M. A fast stereo matching algorithm suitable for embedded real-time systems. *Comput. Vis. Image Underst.* **2010**, *114*, 1180–1202.

[26] Mei, X.; Sun, X.; Zhou, M.; Jiao, S.; Wang, H.; Zhang, X. On building an accurate stereo matching system on graphics hardware. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Barcelona, Spain, 6–13 November 2011; pp. 467–474.

[27] Perri, S.; Colonna, D.; Zicari, P.; Corsonello, P. SAD-based stereo matching circuit for FPGAs. In Proceedings of the International Conference on Electronics, Circuits and Systems, Nice, France, 10–13 December 2006; pp. 846–849.

[28] Gardel, A.; Montejo, P.; Garca, J.; Bravo, I.; Lzaro, J.L. Parametric dense stereovision implementation on a system-on chip (SoC). *Sensors* **2012**, *12*, 1863–1884.

[29] Zhang, X.; Chen, Z. SAD-Based Stereo Vision Machine on a System-on-Programmable-Chip (SoPC). *Sensors* **2013**, *13*, 3014–3027.

[30] Perez-Patricio, M.; Aguilar-Gonzalez, A. FPGA implementation of an efficient similarity-based adaptive window algorithm for real-time stereo matching. *J. Real Time-Image Process.* **2015**, *10*, 1–17.

[31] Krotkov, E.P. *Active Computer Vision by Cooperative Focus and Stereo*; Springer: New York, NY, USA, 1989; pp. 1–17, ISBN 13:978-1-4613-9665-9.

[32] Wan, D.; Zhou, J. Stereo vision using two PTZ cameras. *Comput. Vis. Image Underst.* **2008**, *112*, 184–194.

[33] Kumar, S.; Micheloni, C.; Piciarelli, C. Stereo localization using dual PTZ cameras. In Proceedings of the International Conference on Computer Analysis of Images and Patterns, Mnster, Germany, 2–4 September 2009; pp. 1061–1069.

[34] Kong, W.; Zhang, D.; Wang, X.; Xian, Z.; Zhang, J. Autonomous landing of an UAV with a ground-based actuated infrared stereo vision system. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; pp. 2963–2970.

[35] Ahuja, N.; Abbott, A.L. Active stereo: Integrating disparity, vergence, focus, aperture and calibration for surface estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1993**, *15*, 1007–1029.

[36] Kim, D.H.; Kim, D.Y.; Hong, H.S.; Chung, M.J. An image-based control scheme for an active stereo vision system. In Proceedings of the IEEE/RSJ International Con-

ference on Intelligent Robots and Systems, Sendai, Japan, 28 September–2 October 2004; pp. 3375–3380.

[37] Barreto, J.P.; Perdigoto, L.; Caseiro, R.; Araujo, H. Active stereo tracking of N¡=3 targets using line scan cameras. *IEEE Trans. Robot.* **2010**, *26*, 442–457.

[38] Kwon, H.; Yoon, Y.; Park, J.B.; Kak, A.C. Person tracking with a mobile robot using two uncalibrated independently moving cameras. In Proceedings of the IEEE International Conference on Robotics and Automation, Barcelona, Spain, 18–22 April 2005; pp. 2877–2883.

[39] Junejo, I.N.; Foroosh, H. Optimizing PTZ camera calibration from two images. *Mach. Vis. Appl.* **2012**, *23*, 375–389.

[40] Kumar, S.; Micheloni, C.; Piciarelli, C.; Foresti, G.L. Stereo rectification of uncalibrated and heterogeneous images. *Pattern Recognit. Lett.* **2010**, *31*, 1445–1452.

[41] Ying, X.; Peng, K.; Hou, Y.; Guan, S.; Kong, J.; Zha, H. Self-calibration of catadioptric camera with two planar mirrors from silhouettes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1206–1220.

[42] Wu, Z.; Radke, R.J. Keeping a pan-tilt-zoom camera calibrated. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *35*, 1994–2007.

[43] Schmidt, A.; Sun, L.; Aragon-Camarasa, G.; Siebert, J.P. The calibration of the pan-tilt units for the active stereo head. In *Image Processing and Communications Challenges 7*; Springer: Cham, Switzerland, 2016; pp. 213–221.

[44] Wan, D.; Zhou, J. Self-calibration of spherical rectification for a PTZ-stereo system. *Image Vis. Comput.* **2010**, *28*, 367–375.

[45] Micheloni, C.; Rinner, B.; Foresti, G.L. Video analysis in pan-tilt-zoom camera networks. *IEEE Signal Process. Mag.* **2010**, *27*, 78–90.

[46] Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334.

[47] Weng, J.; Huang, T.S.; Ahuja, N. Motion and structure from two perspective views: Algorithms, error analysis, and error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 451–476.

[48] Sandini, G.; Tistarelli, M. Active tracking strategy for monocular depth inference over multiple frames. *IEEE Trans. Pattern Anal. Mach. Intell.* **1990**, *12*, 13–27.

[49] Davision, A.J. Real-time simultaneous localisation and mapping with a single camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 16–22 June 2003; pp. 1403–1410.

[50] Adelson, E.H.; Wang, J.Y.A. Single lens stereo with a plenoptic camera. *IEEE Trans. Pattern Anal. Mach. Intell.* **1992**, *14*, 99–106.

[51] Fenimore, E.E.; Cannon, T.M. Coded aperture imaging with uniformly redundant arrays. *Appl. Opt.* **1978**, *17*, 337–347.

[52] Hiura, S.; Matsuyama, T. Depth measurement by the multifocus camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, USA, 23–25 June 1998; pp. 953–959.

[53] Mitsumoto, H.; Tamura, S.; Okazaki, K.; Kajimi, N.; Fukui, Y. 3D reconstruction using mirror images based on a plane symmetry recovery method. *IEEE Trans. Pattern Anal. Mach. Intell.* **1992**, *14*, 941–945.

[54] Zhang, Z.; Tsui, H. 3D reconstruction from a single view of an object and its image in a plane mirror. In Proceedings of the International Conference on Pattern Recognition, Brisbane, Australia, 16–20 August 1998; pp. 1174–1176.

[55] Goshtasby, A.; Gruver, W. Design of a single lens stereo camera system. *Pattern Recognit.* **1993**, *26*, 923–937.

[56] Gluckman, J.; Nayar, S.K. Catadioptric stereo using planar mirrors. *Int. J. Comput. Vis.* **2001**, *44*, 65–79.

[57] Pachidis, T.P.; Lygouras, J.N. Pseudostereo-vision system: A monocular stereo-vision system as a sensor for real-time robot applications. *IEEE Trans. Instrum. Meas.* **2007**, *56*, 2547–2560.

[58] Inaba, M.; Hara, T.; Inoue, H. A stereo viewer based on a single camera with view-control mechanism. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Yokohama, Japan, 26–30 July 1993; pp. 1857–1865.

[59] Mathieu, H.; Devernay, F. Système de miroirs pour la stéréoscopie. In Technical Report 0172, INRIA Sophia-Antipolis, French, 1995; pp. 14.

[60] Yu, L.; Pan, B. Structure parameter analysis and uncertainty evaluation for single-camera stereo-digital image correlation with a four-mirror adapter. *Appl. Opt.* **2016**, *55*, 6936–6946.

[61] Lee, D.H.; Kweon, I.S. A novel stereo camera system by a biprism. *IEEE Trans. Rob. Autom.* **2001**, *16*, 528–541.

[62] Xiao, Y.; Lim, K.B. A prism-based single-lens stereovision system: From trinocular to multi-ocular. *Image Vis. Comput.* **2007**, *25*, 1725–1736.

[63] Southwell, D.; Basu, A.; Fiala, M.; Reyda, J. Panoramic stereo. In Proceedings of the IEEE International Conference on Pattern Recognition, Vienna, Austria, 25–19 August 1996; pp. 378–382.

[64] Peleg, S.; Ben-Ezra, M. Stereo panorama with a single camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Collins, CO, USA, 23–25 June 1999; pp. 395–401.

[65] Yi, S.; Ahuja, N. An omnidirectional stereo vision system using a single camera. In Proceedings of the IEEE International Conference on Pattern Recognition, Hong Kong, China, 20–24 August 2006; pp. 861–865.

[66] Li, W.; Li, Y.F. Single-camera panoramic stereo imaging system with a fisheye lens and a convex mirror. *Opt. Exp.* **2011**, *19*, 5855–5867.

[67] Xiang, Z.; Sun, B.; Dai, X. The camera itself as a calibration pattern: A novel self-calibration method for non-central catadioptric cameras. *Sensors* **2012**, *12*, 7299–7317.

[68] Jaramillo, C.; Valenti, R.G.; Guo, L.; Xiao, J. Design and analysis of a single-camera omnistereo sensor for quadrotor micro aerial vehicles (MAVs). *Sensors* **2016**, *16*, 217.

[69] T.M. Bernard, B.Y. Zavidovique, and F.J. Devos, "A programmable artificial retina," *IEEE J. of Solid-State Circuits*, Vol. 28, No. 7, pp. 789–797, 1993.

[70] J.E. Eklund, C. Svensson, and A. Astrom, "VLSI implementation of a focal plane image processor - A realization of the near-sensor image processing concept," *IEEE Trans. on VLSI Systems*, Vol. 4, No. 3, pp. 322–335, 1996.

[71] T. Komuro, I. Ishii, and M. Ishikawa, "Vision chip architecture using general-purpose processing elements for lms vision system," *Proc. of IEEE Int. Workshop on Computer Architecture for Machine Perception*, pp. 276–279, 1997.

[72] M. Ishikawa, K. Ogawa, T. Komuro, and I. Ishii, "A cmos vision chip with simd processing element array for lms image processing," *Proc. IEEE Int. Solid-State Circuits Conf.*, pp. 206–207, 1999.

[73] T. Komuro, S. Kagami, and M. Ishikawa, "A Dynamically Reconfigurable SIMD Processor for a Vision Chip," *IEEE J. of Solid-State Circuits*, Vol. 39, No. 1, pp. 265–268, 2004.

[74] I. Ishii, K. Yamamoto, and M. Kubozono, "Higher order autocorrelation vision chip," *IEEE Trans. on Electron Devices*, Vol. 53, No. 8, pp. 1797–1804, 2006.

[75] S. Hirai, M. Zakoji, A. Masubuchi, and T. Tsuboi, "Realtime FPGA-based vision system," *J. of Robotics and Mechatronics*, Vol. 17, No. 4, pp. 401–409, 2005.

[76] Y. Watanabe, T. Komuro, and M. Ishikawa, "955-fps real-time shape measurement of a moving/deforming object using high-speed vision for numerous-point analysis," *Proc. IEEE Int. Conf. Robot, and Autom.*, pp. 3192–3197, 2007.

[77] I. Ishii, T. Taniguchi, R. Sukenobe, and K. Yamamoto, "Development of high-speed and real-time vision platform, H^3 Vision," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 3671–3678, 2009.

[78] I. Ishii, T. Tatebe, Q. Gu, Y. Moriue, T. Takaki, and K. Tajima, "2000 fps real-time vision system with high-frame-rate video recording," *Proc. IEEE Int. Conf. Robot, and Autom.*, pp. 1536–1541, 2010.

[79] Gluckman, J.; Nayar, S.K. Rectified catadioptric stereo sensors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 224–236.

[80] Shimizu, M.; Okutomi, M. Calibration and rectification for reflection stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 24–26 June 2008; pp. 1–8.

[81] Zhu, L.; Weng, W. Catadioptric stereo-vision system for the real-time monitoring of 3D behavior in aquatic animals. *Physiol. Behav.* **2007**, *91*, 106–119.

[82] Gluckman, J.; Nayar, S.K.; Thoresz, K.J. Real-Time omnidirectional and panoramic stereo. *Comput. Vis. Image Underst.* **1998**, *1*, 299–303.

[83] Koyasu, H.; Miura, J.; Shirai, Y. Real-time omnidirectional stereo for obstacle detection and tracking in dynamic environments. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Maui, HI, USA, 29 October–3 November 2001; pp. 31–36.

[84] Voigtlnder, A.; Lange, S.; Lauer, M.; Riedmiller, M.A. Real-time 3D ball recognition using perspective and catadioptric cameras. In Proceedings of the European Conference on Mobile Robotics, Freiburg, Germany, 19–21 September 2007.

[85] Lauer, M.; Schönbein, M.; Lange, S.; Welker, S. 3D-object tracking with a mixed omnidirectional stereo camera system. *Mechatronics* **2011**, *21*, 390–398.

[86] Hmida, R.; Ben Abdelali, A.; Comby, F.; Lapierre, L.; Mtibaa, A.; Zapata, R. Hardware implementation and validation of 3D underwater shape reconstruction algorithm using a stereo-catadioptric system. *Appl. Sci.* **2016**, *6*, 247.

[87] Liang, C.K.; Lin, T.H.; Wong, B.Y.; Liu, C.; Chen, H.H. Programmable aperture photography: Multiplexed light field acquisition. *ACM Trans. Graph.* **2008**, *27*, doi:10.1145/1360612.1360654.

[88] Moriue, Y.; Takaki, T.; Yamamoto, K.; Ishii, I. Monocular stereo image processing using the viewpoint switching iris. In Proceedings of the IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May 2009; pp. 2804–2809.

[89] T.B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[90] A.W.M. Smeulders, D.M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: an experimental survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 36, no. 7, pp. 1442–1468, 2014.

[91] S.N. Purkayastha, M.D. Byrne, and M.K. O'Malley, "Human-scale motion capture with an accelerometer-based gaming controller," *J. Robot. Mechatron.*, vol. 25, no. 3, pp. 458–465, 2013.

[92] O. Mendels, H. Stern, and S. Berman, "User identification for home entertainment based on free-air hand motion signatures," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 11, pp. 1461–1473, 2014.

[93] C. Bregler, "Motion capture technology for entertainment," *IEEE Sign. Proc. Mag.*, vol. 24, no. 6, pp. 156–158, 2007.

[94] W.M. Hu, T.N. Tan, L. Wang, and S.J. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 3, pp. 334–352, 2004.

[95] K.A. Joshi and D.G. Thakore, "A survey on moving object detection and tracking in video surveillance system," *Int. J. Soft Comput. Eng. (IJSCE)*, vol. 2, no. 3, pp. 44–48, 2012.

[96] B. Noureddin, P.D. Lawrence, and C.F. Man, "A non-contact device for tracking gaze in a human computer interface," *Comput. Vis. Image Underst.*, vol. 98, no. 1, pp. 52–82, Apr. 2005.

[97] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey," *Comput. Vis. Image Underst.*, vol. 108, no. 1, pp. 116–134, 2007.

[98] M. Field, D. Stirling, F. Naghdy, and Z. Pan, "Motion capture in robotics review," in *IEEE International Conference on Control and Automation (ICCA)*, Christchurch, New Zealand, Dec. 2009, pp. 1697–1702.

[99] H. Audren, J. Vaillant, A. Kheddar, A. Escande, K. Kaneko, and E. Yoshida, "Model preview control in multi-contact motion application to a humanoid robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, May, 2014, pp. 4030–4035.

[100] K. Dorfmuller, "Robust tracking for augmented reality using retroreflective markers," *Comput. Graph.*, vol. 23, pp. 795–800, 1999.

[101] U.C. Ugbolue, E. Papi, K.T. Kaliarntas, A. Kerr, L. Earl, V.M. Pomeroy, and P.J. Rowe, "The evaluation of an inexpensive, 2D, video based gait assessment system for clinical use," *Gait Posture*, vol. 38, pp. 483–489, 2013.

[102] M. Myint, K. Yonemori, A. Yanou, K.N. Lwin, M. Minami, and S. Ishiyama, "Visual Servoing for Underwater Vehicle Using Dual-Eyes Evolutionary Real-Time Pose Tracking," *J. Robot. Mechatron.*, vol. 28, no. 4, pp. 543–558, 2016.

[103] A. Censi, J. Strubel, C. Brandli, T. Delbruck, and D. Scaramuzza, "Low-latency localization by Active LED Markers tracking using a Dynamic Vision Sensor," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, Nov. 2013, pp. 891–898.

[104] C. Krishnan, E.P. Washabaugh, and Y. Seetharaman, "A low cost realtime motion tracking approach using webcam technology," *J. Biomech*, vol. 48, no. 4, pp. 544–548, 2015.

[105] K. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, vol. 1, June, 2003, pp. 77–84.

[106] J.L. Martinez, A. Mandow, J. Morales, S. Pedraza, and A. Garcia-Cerezo, "Approximating kinematics for tracked mobile robots," *Int. J. Robot. Res.*, vol. 24, no. 10, pp. 867–878, 2005.

[107] C. Theobalt, I. Albrecht, J. Haber, M. Magnor, and H.-P. Seidel, "Pitching a baseball: Tracking high-speed motion with multi-exposure images," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 540–547, 2004.

[108] A. Aristidou and J. Lasenby, "Real-time marker prediction and CoR estimation in optical motion capture," *The Visual Computer*, vol. 29, no. 1, pp. 7–26, 2013.

[109] S. Hu, Y. Matsumoto, T. Takaki, and I. Ishii, "monocular stereo measurement using high-speed catadioptric tracking," *Sensors*, vol. 17, no. 8, pp. 1839, 2017.

[110] T. Yamazaki, H. Katayama, S. Uehara, A. Nose, M. Kobayashi, S. Shida, M. Odahara, K. Takamiya, Y. Hisamatsu, S. Matsumoto, L. Miyashita, Y. Watanabe, T. Izawa, Y. Muramatsu, and M. Ishikawa, "A 1ms high-speed vision chip with 3D-stacked 140GOPS column-parallel PEs for spatio-temporal image processing," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 82–83.

[111] I. Ishii, T. Taniguchi, K. Yamamoto, and T. Takaki, "High-frame-rate optical flow system," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 22, no. 1, pp. 105–112, 2012.

[112] Q. Gu, T. Takaki, and I. Ishii, "Fast FPGA-based multi-object feature extraction," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 23, no. 1, pp. 30–45, 2013.

[113] I. Ishii, T. Ichida, Q. Gu, and T. Takaki, "500-fps face tracking system," *J. Real-time Image Proc.*, vol. 8, no. 4, pp. 379–388, 2013.

[114] A. Namiki, Y. Imai, M. Kaneko, and M. Ishikawa, "Development of a high-speed multifingered hand system and its application to catching," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2003, pp. 2666–2671.

[115] M. Jiang, T. Aoyama, T. Takaki, and I. Ishii, "Pixel-level and robust vibration source sensing in high-frame-rate video analysis," *Sensors*, vol. 16, no. 11, pp. 1842, 2016.

[116] M. Jiang, Q. Gu, T. Aoyama, T. Takaki, and I. Ishii, "Real-time vibration source tracking using high-speed vision," *IEEE Sensors J.*, vol. 17, no. 11, pp. 1513–1527, 2017.

[117] Q. Gu, T. Aoyama, T. Takaki, and I. Ishii, "Simultaneous vision-based shape and motion analysis of cells fast-flowing in a microchannel," *IEEE Trans. Automat. Sci. Eng.*, vol. 12, no. 1, pp. 204–215, 2015.

[118] Q. Gu, T. Kawahara, T. Aoyama, T. Takaki, I. Ishii, A. Takemoto, and N. Sakamoto, "LOC-based high-throughput cell morphology analysis system," *IEEE Trans. Automat. Sci. Eng.*, vol. 12, no. 4, pp. 1346–1356, 2015.

[119] H. Yang, Q. Gu, T. Aoyama, T. Takaki, and I. Ishii, "Dynamics-based stereo visual inspection using multidimensional modal analysis," *IEEE Sensors J.*, vol. 13, no. 12, pp. 4831–4843, 2013.

[120] K. Okumura, K. Yokoyama, H. Oku, and M. Ishikawa, "1ms auto pantilt– video shooting technology for objects in motion based on saccade mirror with background subtraction," *Advan. Robot.*, vol. 29, no. 7, pp. 457–468, 2015.

[121] L. Li, T. Aoyama, T. Takaki, I. Ishii, H. Yang, C. Umemoto, H. Matsuda, M. Chikaraishi, and A. Fujiwara, "Vibration distribution measurement using a high-speed multithread active vision," in *IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, 2017, pp. 400–405.

[122] B. Tippetts, D. J. Lee, K. Lillywhite, and J. Archibald, "Review of stereo vision algorithms and their suitability for resource-limited systems," *J. Real-Time Image Process.*, vol. 11, no. 1, pp. 5–25, 2016.

[123] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efcient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, 2010.

[124] D. Mishkin, J. Matas, M. Perdoch, and K. Lenc, "Wxbs: Wide baseline stereo generalizations", in *Proc. BMVC*, 2015, pp. 1–12.

[125] D. Zhang, Y. Wang, W. Tao, and C. Xiong, "Epipolar geometry estimation for wide baseline stereo by clustering pairing consensus," *Pattern Recognit. Lett.*, vol. 36, no. 1, pp. 1–9, 2014.

[126] M. Galun, T. Amir, T. Hassner, R. Basri and Y. Lipman, "Wide baseline stereo matching with convex bounded distortion constraints", *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2228–2236.

[127] L. Wolf and A. Zomet, "Wide Baseline Matching between Unsynchronized Video Sequences," *Int. J. Computer Vision*, vol. 68, no. 1, pp. 43–52, 2006.

[128] S. Hu, J. Ming, T. Takaki, and I. Ishii, "Real-time Monocular Three-dimensional Motion Tracking Using a Multithread Active Vision System," *J. Robot. Mechatron.*, vol. 30, no. 3, 2018 (in press).

[129] Chen, S.E.; Williams, L. View interpolation for image synthesis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 218–226.

[130] McMillan, L.; Bishop, G. Plenoptic Modeling: An image-based rendering system. In Proceedings of the ACM SIGGRAPH, New York, NY, USA, 6–11 August 1995; pp. 39–46.

[131] Wexler, Y.; Sashua, A. On the synthesis of dynamic scenes from reference views. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, SC, USA, 13–15 June 2000; pp. 576–581.

[132] Vedula, S.; Baker, S.; Kanade, T. Image-based spatio-temporal modeling and view interpolation of dynamic events. *ACM Trans. Graph.* **2005**, *24*, 240–261.

[133] Beier, T.; Neely, S. Feature-based image metamorphosis. In Proceedings of the ACM SIGGRAPH Computer Graphics, Chicago, IL, USA, 27–31 July 1992; pp. 35–42.

[134] Wolberg, G. Image morphing: A survey. *Vis. Comput.* **1998**, *14*, 360–372.

[135] Schaefer, S.; McPhail, T.; Warren, J. Image deformation using moving least squares. In Proceedings of the ACM SIGGRAPH Computer Graphics, Boston, MA, USA, 30 July–3 August 2006; pp. 533–540.

[136] Chen, K.; Lorenz, D.A. Image sequence interpolation using optimal control. *J. Math. Imaging Vis.* **2011**, *41*, 222–238.

[137] Fortun, D.; Bouthemy, P.; Kervrann, C. Optical flow modeling and computation: A survey. *Comput. Vis. Image Underst.* **2015**, *134*, 1–21.

[138] Meyer, S.; Wang, O.; Zimmer, H.; Grosse, M.; Sorkine-Hornung, A. Phase-based frame interpolation for video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1410–1418.

[139] Spangenberg, R.; Langner, T.; Adfeldt, S.; Rojas, R. Large scale semi-global matching on the CPU. In Proceedings of the Conference on IEEE Intelligent Vehicles Symposium Proceedings, Dearborn, MI, USA, 8–11 June 2014; pp. 195–201.

[140] Q. Gu, T. Takaki, and I. Ishii, "A fast multi-object extraction algorithm based on cell-based connected components labeling," *IEICE Transactions on Information and Systems*, vol. 95, no. 2, pp. 636–645, 2012.

[141] A. Gaschler, D. Burschka, and G. Hager, "Epipolar-based stereo tracking without explicit 3D reconstruction," in *IEEE International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, 2010, pp. 1755–1758.

[142] W. Xu, R. Foong, and H. Ren, "Maker based shape tracking of a flexible serpentine manipulator," in *IEEE International Conference on Information and Automation*, Aug. 2015, pp. 637–642.

Acknowledgment

Firstly, I wish to appreciate my advisor, Prof. Idaku Ishii, who provided me an opportunity to join our laboratory and led me into academic circle. The harvest of my overseas study should owe to his patient instruction both in concrete research and professional attitude. His strict attitude of work and endless passion of pursuing innovation have significantly affected my attitude facing trouble both in work and life.

Besides my advisor, I would like to express my gratitude to Dr. Takeshi Takaki, Dr. Tadayoshi Aoyama, Dr. Qingyi Gu, Dr. Yuji Matsumoto and Dr. Mingjun Jiang. Their invaluable suggestions helped me overcome the unfamiliarity with fresh experimental environment when I initially joined our laboratory.

I would also like to express my heartfelt gratitude to Ms. Yukari Kaneyuki (educational administrator), Ms. Rumi Horiuchi and Ms. Etsuko Yokoyama (laboratory secretary). During my lonely and tough overseas life in Japan, they were my most reliable staffs in our institution, I received considerate attention both in my study and life from them.

I would also like to express my sincere thanks to the bachelor, master and doctoral students in Robotics Laboratory for their help in life and my research.

Last but not the least, I would like to express my profound gratitude to my family. They devoted all their spiritual energy to my education in the past two decades and supported me spiritually throughout writing this thesis and my life in general.

July, 2018

Shaopeng Hu