Free Viewpoint Video Synthesis and Presentation for Dynamic Events

2006

Naho Inamoto

Abstract

Development of image-based modeling and rendering techniques enables to create photorealistic visualizations of dynamic scenes in the real world. The image acquisition is often performed in prepared environment, and the reconstruction target is limited to the foreground objects. If the entire scene can be presented from arbitrary viewpoints using captured images in unprepared environment, these techniques will be widely used in various situations. The immersive and interactive visualization of dynamic events is preferable.

This thesis presents a novel approach for free viewpoint video synthesis and presentation of dynamic events in a large space. The entire scene of sporting match that is captured using multiple cameras in a stadium is represented from a novel viewpoint. A method of virtual view synthesis for dynamic events is proposed, and then free viewpoint video presentation systems are introduced to allow viewers to fly through in real sporting scenes interactively. The technique of view synthesis is expanded to the field of Mixed Reality for creating a new framework for enjoying sporting events such that the viewer can watch the sporting match overlaid onto the real world.

In a large space such as a stadium, it is almost impossible to reconstruct an accurate 3D model of an object because strong calibration of multiple cameras needs many efforts and movements of players are complicated. Instead of constructing a 3D model, projective geometry between cameras is utilized for the view synthesis. Since just corresponding natural feature points in images are required for obtaining projective geometry, which is termed weak calibration, our approach can be easily applied to even large-scale events.

The appearance of the object at virtual viewpoint is generated through transfer of dense correspondence among real cameras. The object scene is segmented into several regions according to the geometric property of the scene. The correspondence is automatically obtained by applying projective geometry to each region. Superimposing virtual view images synthesized in every region completes the appearance of the entire scene. Free viewpoint videos are synthesized by selecting reference cameras, interpolating weight and zoom ratio at each frame in an image sequence. The effectiveness of the proposed method is demonstrated by producing realistic fly-through videos where the entire scenes are naturally reconstructed from virtual viewpoints.

In addition, viewpoint on-demand system and mixed reality presentation systems are introduced as applications for free viewpoint replays of sporting match. The first system allows viewer to select his/her viewpoint on GUI. The second systems overlay a sporting match onto a desktop stadium model in the real world using a head mounted display or a handheld display with a web camera attached to. When overlaying a virtual object on the real world, 3D positional relationships among virtual object, real world, and viewer's position are generally required for registration. The conventional method cannot be used for the proposed method where the above 3D information is not available. Two approaches are proposed to achieve image-based registration between virtual view images of players and a desktop stadium in the real world: feature-based registration and marker-based registration. Viewer's position is calculated using an image in which the real environment is captured, and the positions of players in the desktop stadium are obtained through projective geometry of the ground plane of the stadium. The mixed reality presentation systems are demonstrated; a virtual soccer match is replayed on a small desktop stadium using multiple soccer videos. The impression is given to the viewer as if the event is taken place in front of him/her. The proposed systems are based on the condition that objects move on a planar area. In addition, it does not require strong calibration of multiple cameras that capture the dynamic event. Therefore our approach can be applied to entertainments as well as sporting events.

Acknowledgments

I wish to sincerely thank my advisor Associate Professor Hideo Saito for his guidance, encouragement, and support throughout my graduate studies. His advice, ideas, and suggestions on my research have been a tremendous influence and an invaluable resource for me. I would like to thank the other members of my thesis committee: Professor Nozomu Hamada, Professor Yoshio Ohno, and Professor Shinji Ozawa for their comments and suggestions on this thesis.

My work for this thesis is supported in part by a Grant in Aid for the 21st century Center of Excellence (COE) for Optical and Electronic Device Technology for Access Network from the Ministry of Education, Culture, Sports, Science, and Technology in Japan, and additionally by Japan Society for the Promotion of Science Research Fellowship for Young Scientists. I deeply thank leaders of the COE program Professor Toshiaki Makabe and Professor Minoru Obara for all their support. The advanced lectures and on-going high-level research activities helped me a lot obtain multi-disciplinary knowledge, deep understanding, creativity, and program-solving skills for the creation of innovative information technology in the feature. Furthermore, I have been benefited from interactions with COE research assistants through the program. They have been great colleagues and friends for me.

I gained many valuable experiences in my international research internship at the Robotics Institute, Carnegie Mellon University (CMU). The study under Professor Takeo Kanade made a great impact on my work and me. I am very fortunate to have had the opportunity to work with him during my time at CMU. The Vision Autonomous Systems Center (VASC) members have helped me a lot both in my study

and in everyday life. I deeply appreciate Professor Kanade and all VASC members at CMU.

Finally, I would like to express my gratitude to all members of Hideo Saito laboratory and Ozawa and Sato laboratory at Keio University. Most of all, I am deeply grateful to my parents for supporting my study for more than twenty years.

February 2006
Naho Inamoto

Contents

Chapter 1: Introduction 1
1.1 Background and Motivation 2
1.2 Overview of the Approach 5
1.3 Thesis Outline 8

Chapter 2: Related Work 9
2.1 Image Based Rendering 10
2.1.1 Rendering without Geometry 11
2.1.2 Rendering with Explicit Geometry 12
2.1.3 Rendering with Implicit Geometry 13
2.2 Free Viewpoint Video for Virtual or Mixed Reality Environment . . . 15
2.3 Sports Video Processing 17

Chapter 3: Virtual View Synthesis 19
3.1 View Interpolation for Dynamic Events in a Large Space 20
3.2 Estimation of Projective Geometry 24
3.2.1 Fundamental Matrix 24
3.2.2 Homography 25
3.3 Background Image Generation 26
3.4 View Synthesis for Static Regions 30
3.4.1 Field Regions 30
3.4.2 Background Region 31

3.5 View Synthesis for Dynamic Regions 33
3.5.1 Overview 33
3.5.2 Extraction of Dynamic Regions 33
3.5.3 Segmentation of Dynamic Regions 35
3.5.4 Shadow Region 36
3.5.5 Player/Ball Region 36
3.6 Experimental Results 39
3.6.1 Image Acquisition 39
3.6.2 Results on Real Images 41
3.6.3 Results on Synthetic Images 47
3.6.4 Discussion 48

Chapter 4: Free Viewpoint Video Generation 51
4.1 Free Viewpoint Video 52
4.2 Viewpoint on Demand System 56
4.2.1 System Overview 56
4.2.2 Offline Process 56
4.2.3 Online Process 57
4.2.4 User Interface 59
4.2.5 Experimental Results 60
4.2.6 Discussion 62

Chapter 5: Mixed Reality Presentation 64
5.1 Expansion toward Mixed Reality 65
5.1.1 Overview of Mixed Reality 65
5.1.2 Instruments 67
5.1.3 Registration Techniques 72
5.1.4 Vision-Based Tracking 75
5.2 Feature-Based MR Presentation System 78
5.2.1 System Overview 78
5.2.2 Calculation of Viewpoint 80
5.2.3 Overlay of Dynamic Objects 83

5.2.4 Experimental Results 86
5.2.5 Discussion 93
5.3 Marker-Based MR Presentation System 94
5.3.1 System Overview 94
5.3.2 Camera Calibration 96
5.3.3 Calculation of Viewpoint 97
5.3.4 Overlay of Dynamic Objects 102
5.3.5 Experimental Results 105
5.3.6 Discussion 110

Chapter 6: Conclusions 111
6.1 Summary 112
6.2 Contributions 113
6.3 Future Work 115
6.4 Prospect for Applications 116

Appendices 118
A Projective Geometry between Cameras 119
A.1 Epipolar Geometry 119
A.2 Epipolar Constraint 120
A.3 Estimation of Fundamental Matrix 121
B Projective Geometry between Cameras and a World Plane 123

References 125

List of Figures

1.1 Overview of the proposed approach. 5

2.1 Image-based modeling and image-based rendering. 10

3.1 View interpolation. 20
3.2 View synthesis for soccer scenes. 22
3.3 Fundamental matrix. 24
3.4 Homography. 25
3.5 Background generation from the image sequence. 27
3.6 Variation in pixel value in the image sequence and the histogram. . . . 28
3.7 Examples of virtual view images for the field regions. 30
3.8 Image mosaicing. 31
3.9 Examples of virtual view images for the background. 32
3.10 Process flow of the view synthesis for the dynamic regions. 33
3.11 Extraction of the dynamic regions. 34
3.12 Segmentation of the dynamic regions. 35
3.13 Silhouette correspondence. 36
3.14 Silhouette correspondence in the case of occlusion. 37
3.15 Pixel correspondence. 38
3.16 Reconstruction of the player from different angles. 38
3.17 Camera configuration in the stadiums. 39
3.18 Examples of multiple view images captured in the stadiums. 40
3.19 Synthesized virtual view images at one frame for the entire soccer scene from real camera images in the Edogawa Athletic Stadium. 42

3.20 Close-up views of the previous figure. 43
3.21 Comparison between the virtual and real camera images on the real soccer scene. 44
3.22 View interpolation among three views for the soccer scene captured in the Kashima Stadium. 45
3.23 Comparison of the synthesized virtual view images with/without segmentation of player regions and shadow regions in view interpolation. 46
3.24 Comparison between the virtual view image and ground truth image on the synthetic scene. 47
3.25 Epipolar lines between neighboring cameras. 48

4.1 Process flow of the free viewpoint video generation. 53
4.2 Fly-through view image sequence. 54
4.3 Visual effect of the freeze-and-rotate camera motion. 55
4.4 Offline and online processes in the Viewpoint on Demand System. . . 58
4.5 The interface of the Viewpoint on Demand System. 59
4.6 Examples of the image window of the Viewpoint on Demand System. . 60
4.7 Example of free viewpoint replay for the sequence including shadows in the Viewpoint on Demand System. 61
4.8 Correlation between the processing time and number of the objects in the Viewpoint on Demand System. 62

5.1 Example images of mixed reality presentation of a soccer match. . . . 65
5.2 Different types of displays in MR environment. 67
5.3 Video see-through HMD and optical see-through HMD. 68
5.4 System configuration using an HMD. 70
5.5 System configuration using a web camera and a handheld display. . . . 70
5.6 Comparison of the registration techniques. 73
5.7 Process flow of the feature-based MR presentation system. 78
5.8 Examples of edge images (left) and detected feature lines (right). . . . 80
5.9 Comparison of the location of vanishing points between in the MR camera image and in the stadium camera images for feature-based presentation. 81

5.10 Determination of the player and ball positions in the MR camera image. 83
5.11 Representation of the shadows on the stadium model. 85
5.12 Watching a soccer match in the viewer's environment using an HMD. 86
5.13 Result image of feature-based MR presentation and the original soccer scene. 89
5.14 Result image sequence in the feature-based MR presentation system. . 90
5.15 Result image sequence including shadows in the feature-based MR presentation system. 91
5.16 Close-up view of the feature-based MR presentation of a soccer match. 91
5.17 Comparison of the player positions between in the case of using the centroid and the foot position for the registration. 92
5.18 Process flow of the marker-based MR presentation system. 94
5.19 The checker pattern used for camera calibration. 96
5.20 Process flow of the marker detection in ARToolkit. 97
5.21 The world coordinate system defined in the proposed method. 98
5.22 Examples of the projected feature lines using estimated camera position and orientation. 99
5.23 Comparison of the location of vanishing points between in the MR camera image and in the stadium camera images for marker-based presentation. 100
5.24 Determination of the player and ball positions in the MR camera image via the desktop stadium model. 103
5.25 Examples of a 2D marker and a desktop stadium model. 105
5.26 Result image #1 of marker-based MR presentation and the original soccer scene. 108
5.27 Result image #2 of marker-based MR presentation and the original soccer scene. 108
5.28 Result image sequence #1 in the marker-based MR presentation system. 109
5.29 Result image sequence #2 in the marker-based MR presentation system. 109

A.1 Epipolar geometry. 119
B.1 The homography induced by a plane. 123

Chapter 1:

Introduction

1.1 Background and Motivation

The traditional goal of Computer Graphics research has been developing algorithms for realistic rendering of synthetic scenes from arbitrary viewpoints. On the other hand, the focus of Computer Vision is the inverse process of extracting a model of a given real world scene using information obtained with cameras. Recently, a convergence between the fields of Computer Graphics and Computer Vision has resulted in an emerging research area known as image-based modeling and rendering. Development of image-based modeling and rendering techniques enables to create photo-realistic visualizations of real world scenes in a computer not by designing models of shape and appearance, but by reconstructing these models from photographic or video data of the real world.

The challenging topic in this interdisciplinary research area is free viewpoint video, which enables the selection of an arbitrary viewpoint onto a dynamic scene, thereby creating a feeling of immersion into the event.

Virtualized reality [53], which is a pioneering project in this field, has achieved such realistic visualization of dynamic scenes from arbitrary viewpoints. A three-dimensional model of an object in a target scene is reconstructed from multiple view images. The colors in real images are used to synthesize the texture of the 3D model. Using conventional rendering techniques, new view images are generated from the color-textured 3D model.

Free viewpoint video effects can be seen frequently in recent motion pictures or television broadcastings. Freeze-and-rotate camera shots around actors are included in movies such as "The Matrix" series [114]. These effects are made possible by recording the actor with tens to hundreds of cameras placed around the set and switching the cameras continuously. Another example of these effects in TV broadcasting is the "Eye Vision" system [29] that was used for the Super Bowl XXXV broadcast by CBS. Multiple video streams are captured using more than 30 cameras. The sequences of video images from different angles are then used to create virtual camera movements such that the viewpoint revolves around the object event at a temporally frozen moment. Both systems create free viewpoint videos by simply switching the video camera images. The object scene can be presented only from real camera view-

points. A large number of cameras are required for free viewpoint movement. It is preferable to produce the same effect using a small number of cameras. In addition to special effects created by producers, interactive visualization can increase viewer's experience greatly.

In this thesis, we firstly propose a method for synthesizing free viewpoint video of dynamic events in the real world. Especially, we focus on sporting scene such as soccer match captured in a stadium. Virtual viewpoint images are synthesized from reference camera images taken with uncalibrated multiple cameras. Without constructing a 3D model, the projective geometry between neighboring cameras is only used to synthesize new view images [47]. View interpolation [18] is utilized for presenting the entire soccer scene from any intermediate viewpoints among the real cameras. Furthermore, we introduce free viewpoint video presentation system: "Viewpoint on Demand System", which enables viewers to select their favorite viewpoints while observation through standard GUI [48].

Secondly, we expand our view synthesis method to the field of Mixed Reality. In recent years, Mixed Reality/Virtual Reality technologies have been used for enhancement of sport coverage on television and on the Internet. One example of those enhancements is real time augmentation of broadcast video. A virtual advertisement is inserted or replaced in a scene [81, 90, 101, 112, 122]. Alternatively, virtual objects such as virtual first-down line in football [90], virtual offside line representing the last defender line in soccer [81], and virtual record in track and field races or swimming [81], are inserted in live video. Some other applications have been also developed. The FoxTrax system [16] highlights the location of a hard-to-see hockey puck as it moves rapidly across the ice. The VideoFinish system [51] merges the videos depicting the performances of two athletes into a single video stream as if the athletes were actually competing together at the same time.

Another type of example is virtual replay to enhance sport coverage. The LucentVision system [64, 84] produces a virtual replay that allows to visualize the trajectory of the tennis ball during serves from any point of view. Orad [81] has introduced the Virtual Replay and VirtualLive system which are respectively static and dynamic 3D rendering applications for a soccer match animation. In the PISTE (Personalized Immersive Sports TV Experience) project [85, 66] that was established for developing a set of mixed-reality application for interactive sports television, the scenario of 3D

modeling of the moving athlete was described.

Although many applications and products have been developed to enhancement of sport coverage, the conventional visualization just inserts virtual lines/objects into live video, or replays the scenes into graphical animation. The augmented video is presented to viewers typically via a television screen or a computer screen. It does not give enough immersive impression or interactivity to the viewer because the viewer's environment is not taken into account.

In latter part of this thesis, we introduce a new type of system that embeds a sporting event into viewer's environment. The soccer match is replayed not in the original stadium but in a desktop stadium that is located in front of the viewer. This requires generating an appropriate view of the scene according to the viewer's position. The synthesized scene should be naturally inserted into the viewer's environment. For visualization of a sporting event in the viewer's environment, we propose a novel method for registration between the original stadium and the desktop stadium based on projective geometry between cameras [49, 50]. Feature-based and marker-based mixed reality presentations are demonstrated. The proposed systems allow the viewer to watch a soccer match in the real world via a head mounted display (HMD) or a handheld display with a web camera attached to from favorite viewpoints.

1.2 Overview of the Approach

This thesis proposes a novel method of virtual view synthesis for dynamic events in a large space. Free viewpoint video presentation systems are then introduced to allow viewers to fly through in real sporting scenes. The technique of view synthesis is expanded to the field of Mixed Reality for creating a new framework for enjoying sporting events such that viewers can watch a sporting match overlaid onto the real world.

The overview of the proposed approach is presented in Figure 1.1. A soccer match is captured using uncalibrated and fixed multiple cameras in a real stadium. The soccer scene synthesized by applying view synthesis method is presented from the chosen viewpoint by a viewer.

Images of virtual viewpoints are generated by view interpolation among real camera images. Two or three cameras near the virtual viewpoint chosen by the viewer that correspond to reference cameras are selected from multiple cameras. The virtual viewpoint image is generated through correspondence among the selected reference cameras. As the target is a dynamic event in a large space, the object scene is segmented into dynamic regions (foreground objects) and static regions (background). View interpolation is then performed for each region independently. After virtual view

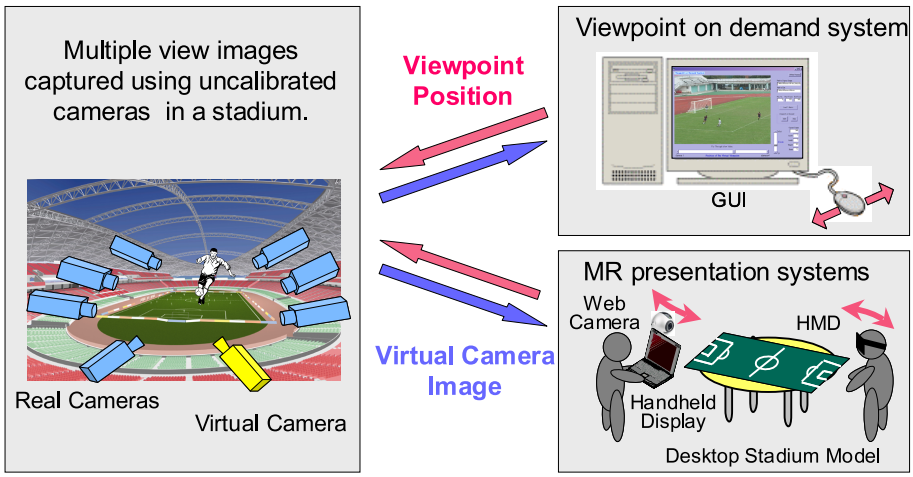


Figure 1.1: Overview of the proposed approach.

images are synthesized in every region, superimposing them visualizes the entire scene of the target from the selected viewpoint. The projective geometry between cameras that is used for the view interpolation is estimated in captured multiple view images in advance.

In Viewpoint on Demand System, virtual view images of background including the stadium and its shadow are generated beforehand. Once the viewer selects a viewpoint position on GUI, virtual view images of foreground objects such as players, ball, and their shadows are synthesized from reference cameras. The generated images are then overlaid on the virtual view image of the original soccer stadium at the corresponding viewpoint.

For Mixed Reality (MR) presentation, free viewpoint video presentation is performed on a desktop stadium model in front of a viewer using an HMD or a handheld display with a web camera attached to. Two kinds of visualization systems are introduced. In the first system, registration is performed by tracking natural features. A camera captures the desktop stadium model in the real environment. The viewpoint position is determined by tracking feature lines in the camera image. Virtual view images of dynamic objects are synthesized from reference cameras in the same way as in the Viewpoint on Demand System, and then overlaid on the desktop stadium model according to the viewpoint position. If the original soccer scene contains shadows, the shadows of players and ball are represented on the stadium model using the relationship between the original stadium and the stadium model.

In the second system, marker-based tracking is utilized for achieving an online MR presentation. A camera takes the real environment including the stadium model where a 2D square marker is attached. According to the camera position and orientation obtained by marker tracking, virtual view images of the dynamic objects are overlaid

Table 1.1: Configuration of the proposed systems.

	Viewpoint on Demand System	MR presentation systems
Viewpoint selection	GUI	Camera control
View synthesis	Entire scene	Foreground objects
Presentation	Original stadium	Desktop stadium model

at the appropriate position onto the stadium model. This system enables online observation of a soccer match.

Table 1.1 describes the configuration of the proposed systems. Visualization process in all systems consists of three stages: viewpoint selection, virtual view synthesis, and presentation of the synthesized scene at the desired viewpoint. In Viewpoint on Demand System, the viewer directly selects a viewpoint on GUI, and then entire scene is presented with the original stadium. On the other hand, the viewpoint is determined according to the position and pose of the camera, and then just foreground objects are overlaid onto the stadium model in MR presentation systems. The methods for the viewpoint selection and registration between original stadium and desktop stadium are proposed for MR presentation.

1.3 Thesis Outline

The outline of this thesis is as follows. In Chapter 2, related work is discussed. Image-based rendering techniques are explained in three categories according to how much geometric information is used. Techniques for free viewpoint video generation and video processing targeting sporting scenes are then described.

Chapter 3 presents a novel approach for virtual view synthesis of dynamic scenes in a large-scale space. After reviewing the theory of projective geometry between cameras used in the proposed approach, virtual view synthesis methods for each segmented region of soccer scene are described. Subsequently, experimental results show realistic soccer scenes at virtual viewpoints generated by uncalibrated camera images. The effectiveness of the proposed method is demonstrated by comparing virtual camera image with real camera image at the same position using both real image and synthetic image.

Chapter 4 introduces the method for generating free viewpoint videos. An interactive visualization system "Viewpoint on Demand System" enables viewers to watch the dynamic events from their favorite viewpoints.

Chapter 5 expands the free viewpoint video generation technique to the field of Mixed Reality. An image-based registration technique is proposed to achieve insertion of sporting events into the real world. A viewer can virtually watch the scenes overlaid on a desktop stadium model. The prototype system is demonstrated where a virtual soccer match is replayed on a small desktop stadium using multiple soccer videos.

Chapter 6 summarizes contributions of this work, in addition to looking at future directions.

Chapter 2:

Related Work

2.1 Image Based Rendering

Over the last decade, various image-based modeling and image-based rendering (IBR) techniques have been developed [106]. They recover model information and render novel views from images directly. IBR techniques can be classified into three categories according to how much geometric information is used: rendering without geometry, rendering with explicit geometry (either with approximate or accurate geometry), and rendering with implicit geometry (i.e. correspondence). The second one using explicit geometry is commonly termed model-based approach while the third one using implicit geometry is termed transfer-based approach. The categories of these techniques are indicated in Figure 2.1.

The proposed method utilizes rendering with implicit geometry, which is included in the transfer-based approach. Novel views are synthesized and presented using projective geometry between cameras without constructing a 3D model. We describe the reason why the transfer-based approach is appropriate for visualization of dynamic events in a large space while explaining the methods already proposed in each category.

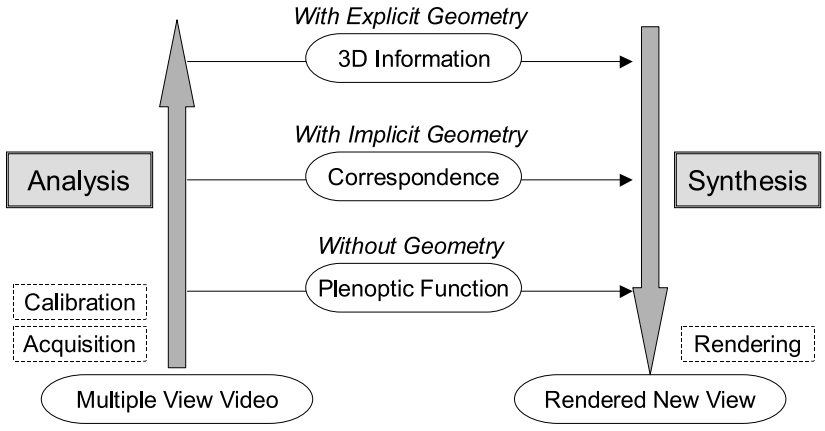


Figure 2.1: Image-based modeling and image-based rendering.

2.1.1 Rendering without Geometry

The techniques for rendering without geometry rely on the characteristics on plenoptic function, which describes all the radiant energy perceived by an observer at any point in space and time. They use many images, but do not require any geometric information or correspondence for creating novel views.

In its most general forms, the plenoptic function is a seven-dimensional function. Due to its high dimensional nature, data reduction or compression of the plenoptic function is essential. The light field of Levoy and Hanrahan [61] and the Lumigraph of Gortler et al. [39] simplified the function to four dimensions. Shum and He [105] proposed a new IBR technique termed concentric mosaic for virtual reality applications. Unlike light field and Lumigraph where cameras are placed on a 2D grid, the concentric mosaics representation reduces the amount of data by capturing a sequence of images along a circle path.

This approach provides much better image quality and lower computational requirement for rendering virtual views than model-based approach described in the next section. However, many input images taken at slightly different positions are necessary. A large number of cameras are required in case that the target is a sporting event held in a stadium. The data volume to deal with all radiant energy in a large-scale space becomes huge. Therefore it is not efficient to apply this approach to large-scale events.

Meanwhile, plenoptic function is also used for image mosaicing. A complete plenoptic function at a fixed viewpoint can be constructed from incomplete samples. Specifically, a panoramic mosaic is constructed by registering multiple regular images. Many systems have been built to construct cylindrical and spherical panoramas by stitching multiple images together [19, 72, 113]. Szeliski and Shum [105] presented a complete system for constructing panoramic image mosaics from sequences of images. Their mosaic representation associates a transformation matrix with each input image rather than explicitly projecting all of the images onto a common surface, such as a cylinder. A part of the proposed approach in this thesis makes use of the idea of panoramic mosaic. Distant views of a stadium from virtual viewpoints are synthesized by stitching multiple regular images.

2.1.2 Rendering with Explicit Geometry

The techniques in this category (i.e. model-based approach) use direct 3D information. The key technologies are shape reconstruction of objects from multiple view images and rendering virtual viewpoint images. The more traditional 3D modeling and texture-mapping [74, 119] belong to this category. The methods proposed recently are described here.

When the depth information is available for every point in one or more images, 3D warping techniques [73] can be used to render nearly all viewpoints. An image can be rendered from any nearby point of view by projecting the pixels of the original image to their proper 3D locations and re-projecting them onto the new picture. To deal with the disocclusion artifacts in 3D warping, Shade et al. [104] proposed LDI to store not only what is visible in the input image, but also what is behind the visible surface.

To obtain these visual effects of a reconstructed architectural environment, Debevec et al. [24] used view-dependent texture mapping to render new views by warping and compositing several input images of an environment. Buehler et al. [13] applied a more principled way of blending textures based on relative angular position, resolution, and field-of-view. Vedura et al. [120] proposed an algorithm for creating novel views of a non-rigidly varying dynamic event by combining images captured from different positions, at different times. The algorithm operates by combining images captured across space and time to compute voxel models of the scene shape at each time instant, and dense 3D scene flow between the voxel models.

More recently, Saito et al. [96] proposed a method to view interpolation approach that is termed "Appearance-Based Virtual-View Generation". A 3D model of a scene is reconstructed from multiple images by using Multiple Baseline Stereo [80] and Shape from Silhouette [17, 88]. Dense and precise correspondences between the two images are obtained using this constructed 3D model, and then used to generate virtual views at arbitrary viewpoints without losing pixels even in partially occluded regions. Yaguchi and Saito [128] extended this method to uncalibrated case. They introduce a framework of projective grid space (PGS), which has a special 3D coordinate system established by epipolar geometry of camera. PGS is used for the reconstruction of 3D

models without camera calibration.

Kitahara and Ohta [56] proposed a method to reconstruct a 3D object shape effectively with a set of planes. The LOD (level of detail) of the 3D representation is controlled by adjusting the orientation, interval, and resolution of planes. This enabled to generate a 3D model that has spatial resolution just as fine for the purpose of producing an output 3D video.

Matsuyama et al. [69] proposed a parallel pipeline processing method for reconstructing a dynamic 3D object shape from multi view video images. They introduced the plane-based volume intersection method, its acceleration algorithm, the parallel pipeline implementation and additionally a high-fidelity texture mapping method. This allows obtaining a temporal series of full 3D voxel representations of the object behavior in real time.

Most of techniques in this category require camera calibration [115] to deal with 3D information. Considering image acquisition in a large-scale space such as a stadium, it is difficult to calibrate multiple cameras precisely. Even if the calibration can be performed, reconstruction of time-varying and accurate 3D model for the entire stadium has much effort. Model-based approach is not practical for dynamic events in a large space.

2.1.3 Rendering with Implicit Geometry

The techniques in this category (i.e. transfer-based approach) rely on positional correspondences across a small number of images to render new views. Without an explicit 3D model, new views are computed based on direct manipulation of these positional correspondences.

Chen and Williams [18] proposed view interpolation, which uses dense optical flow to generate intermediate views directly. Seitz and Dyer [102] additionally proposed a view interpolation technique termed view morphing. An intermediate view that is geometrically correct is synthesized between a pair of images for a static scene. Avidan and Shashua [1] employed a trifocal tensor for image transfer.

In these methods, dense correspondence between the original images is required to generate intermediate views. The correspondence is often generated manually or by optical flow; hence, almost all the targets are static images or slightly varying

images such as facial expressions. This means that if the correspondence is obtained automatically in each frame, this type of technique is useful for dynamic scenes. We introduce the way to obtain dense correspondence efficiently using projective geometry between cameras for virtual view synthesis.

More recently, view interpolation has been applied to images captured from different positions, at different times. Manning and Dyer [67] extended view morphing [102] to rigid objects with translation, which is termed dynamic view morphing. Wexler and Shashua [124] proposed another technique to morph a dynamic view with a moving object along a straight line path from three viewpoints. While the above two methods have only dealt with translation, Xiao et al. [127] extended the view morphing technique to a rotation case and applied it to non rigid objects with complicated motion.

All these methods calculate motion parameters of the objects in order to interpolate the appearance of the moving objects. If many moving objects are included in a scene, calculation of motion parameters of all objects needs much computation. It is difficult to capture complicated movements of players in distant view of sporting match. The interpolation on temporal axis is not necessary in case that the images are captured at the same time at different positions. Therefore we make use of simple view interpolation technique instead of dynamic morphing. The interpolation is performed between different views in every frame instead of dealing with the interpolation on temporal axis so that unstable process of motion estimation can be avoided.

Matusik et al. [70] proposed Image-based Visual Hull (IBVH) that is another virtual view synthesis method from multiple cameras. In IBVH, the hull shape of the object is represented by the intersection of silhouettes on the epipolar lines of one base camera. They also proposed a real-time method for creating and rendering visual hulls [71]. Unlike voxel or sampled approaches, it computes an exact polyhedral representation for the visual hull directly from the silhouettes. View-independent representation can be computed quickly. In order to reconstruct precise visual hull, extraction of accurate silhouette is important issue, and surrounding cameras should be located around the object. Although their methods are very effective, the silhouette extraction and the camera configuration may be critical problem in sporting scene. The camera placement has limitations in a stadium. Silhouettes of the dynamic objects are not always extracted accurately in complicated scenes.

2.2 Free Viewpoint Video for Virtual or Mixed Reality Environment

Inserting live-action of human movement in virtual environment or mixed reality environment has begun to attract increasing attention from researchers. Some methods already proposed are described here.

Virtualized Reality project [119] has shown the ability to integrate multiple digitized events with virtual contents. For example, a virtual baseball was added to the scene to create the illusion that the batter actually was swinging at the object. Another example is the integration of the reconstructed human models with a CAD model of a virtual basketball court. This gives us impression that the persons are playing basketball on a virtual court.

Gross et al. [41] presented "blue-c", a system combining simultaneous acquisition of video streams with 3D projection technology in a CAVE-like environment; this creates the impression of total immersion. Multiple live video streams acquired from many cameras are used to compute a 3D video representation of a user in real time. The resulting video inlays are integrated into a virtual environment. Although the impression of the total immersion is provided, blue-c does not allow tangible ways to manipulate 3D videos captured. There are few interactions described between these 3D human avatars and other virtual objects.

Another capture system was presented in [44]. The authors demonstrated a complete system architecture allowing the real-time acquisition and full body reconstruction of one or several actors, which could be integrated in a virtual environment. Images captured from four cameras are processed to obtain a volumetric model of the moving actors, which is used to interact with other objects in the virtual world. However, the resulting 3D models are generated without texture, leading to some limitations in applying their system.

Grau et al. [40] proposed a new virtual studio system for the production of 3D content. The system combines the ability to capture dynamic scenes, based on a multi camera system in a chroma-key environment, with a view-dependent projection for actor feedback. The system allows the generation and rendering of 3D models in pre-

view quality for on-set visualization in real time and in high quality for postproduction applications in an offline phase.

Prince et al. [89, 77] introduced a real-time system for capturing humans in 3D and placing them into a mixed reality environment. The subject is captured by nine cameras surrounding him/her. Looking through a head-mounted-display with a camera in front pointing at a marker, the user can see the 3D image of this subject overlaid onto a mixed reality scene. The 3D images of the subject viewed from this viewpoint are constructed using a robust and fast shape-from-silhouette algorithm.

Although the conventional work enabled insertion of human action in virtual environment or mixed reality environment, the target is still limited a couple of human movements. This thesis proposes the method for superimposing dynamic events held in a large space on the real world. We describe the differences between the proposed method and Prince's method that is supposed to be the most related work. View synthesis and presentation are performed based on the projective geometry between multiple cameras (i.e. transfer-based approach) in our method, while 3D models are reconstructed (i.e. model-based approach) in Prince's method. Their method requires strong calibration of multiple cameras, which is difficult to obtain in a large space. The projective geometry between cameras can be easily obtained from just images. Image-based registration technique enables the mixed reality presentation of a soccer match without the 3D information.

2.3 Sports Video Processing

The development of high-speed digital cameras and video processing has attracted people's attention in sports video analysis [123, 130]. Much work has been done on sports video including highlight detection and summarization [95, 62, 26], indexing [4], tracking of ball/players [84, 52], and 3D reconstruction [66]. In particular, we review the methods for view synthesis of sporting scene.

One approach for arbitrary view generation is reconstruction of a sporting match using CG animation [68, 8, 84, 25, 23, 92]. Data of the actions and positions of players are extracted from video images, and then the data derive CG model players. The viewer can watch an animation of the event from favorite viewpoint. In this approach, it is easy to render the object scene from various angles. Furthermore, quality of images does neither depend on the number of cameras nor the quality of original video images. However, reality of the rendered video is not always sufficient.

Another approach is arbitrary view-synthesis by computer vision-based technologies. Arbitrary view of the event can be generated from 3D structure of the scene that is reconstructed via images captured with video cameras. In order to achieve such realistic reconstruction, the methods for calibration of moving TV cameras and the modeling of the moving athlete have been proposed by Malerczyk et al. [66]. Carranza et al. [15] introduced a system that uses multi-view synchronized video footage of an actor's performance to estimate motion parameters and to interactively re-render the actor's appearance from any viewpoint. Koyama et al. [59] proposed a method for arbitrary view presentation for a soccer match. Each player is represented with a simplified 3D model, which is reconstructed from multiple videos. Observers can watch the motion of the players with a computer-generated virtual stadium from arbitrary viewpoint positions. Yan et al. [129] proposed a similar method, where the soccer field and the ball are graphically reconstructed, and the segmented players are rendered from the corresponding frames of the original video. The system can reconstruct not only the goalmouth scene but also the midfield scene as well. The reconstructed video is enriched by music and illustrations of the video contents.

While the conventional methods replay the scenes in part graphically, we propose a method where entire scene at virtual viewpoint is synthesized from real images. This

achieves more realistic visualization of the object scene. In addition, we introduce a new type of system that embeds a sporting event into the viewer's environment.

Chapter 3:

Virtual View Synthesis

3.1 View Interpolation for Dynamic Events in a Large Space

View interpolation technique [19] generates images at arbitrary viewpoints by transfer of correspondence among real cameras. In this transfer based approach; it is usually essential to obtain dense corresponding points among reference camera images. As a synthesis image relies on the correspondence information, the problem comes down to how to generate dense correspondence correctly.

We firstly describe the method of view interpolation used in our approach in the case of two views and three views. Subsequently, we give a description of how to obtain correspondence among real cameras in dynamic scene at large-scale space.

Suppose there are two real cameras at different positions to take the same object scene. An image of the virtual camera located in-between two cameras is generated as follows. Given pixel-by-pixel correspondence between two real cameras, the position and the value of the pixels on virtual camera are transferred by image morphing [9] as described by the following equations:

$$\acute{\boldsymbol{p}} = (1-\alpha)\boldsymbol{p}_1 + \alpha\boldsymbol{p}_2 \tag{3.1}$$

and

$$I(\acute{p}) = (1-\alpha)I(p_1) + \alpha I(p_2)\ , \tag{3.2}$$

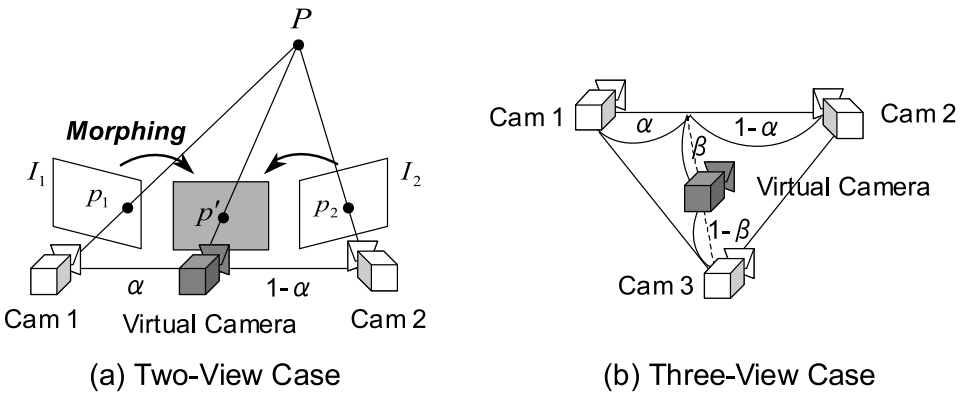


Figure 3.1: View interpolation.

where $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are the coordinates of the corresponding points p_1 and p_2 in images I_1 and I_2, respectively, and $I(p_1)$ and $I(p_2)$ are the pixel values of p_1 and p_2, respectively. $\acute{\boldsymbol{p}}$ represents the interpolated coordinates, and $I(\acute{p})$ represents the interpolated pixel value. α defines the interpolating weight assigned to the respective actual viewpoints as shown in Figure 3.1 (a). These functions generate warped image from real cameras. After two warped images are generated using two directed correspondences: from camera 1 to camera 2 and from camera 2 to camera 1, they are blended into a single image, which is a target image at an intermediate viewpoint. In blending two images, if the color of a pixel differs between these images, the corresponding pixel in the virtual view is rendered with the average of the colors; otherwise, the rendered color is taken from either of the actual images.

In case of view interpolation among three views, the viewpoint position is determined by weight α and weight β as shown in Figure 3.1 (b). The virtual view image is synthesized from three real camera images by morphing as in the case of two views. The following equations are used instead of the Equations (3.1) and (3.2):

$$\acute{\boldsymbol{p}} = (1-\alpha)(1-\beta)\boldsymbol{p}_1 + \alpha(1-\beta)\boldsymbol{p}_2 + \beta\boldsymbol{p}_3 \ , \tag{3.3}$$

and

$$I(\acute{p}) = (1-\alpha)(1-\beta)I(p_1) + \alpha(1-\beta)I(p_2) + \beta I(p_3) \ , \tag{3.4}$$

where $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, and $\boldsymbol{p}_3$ are the coordinates of the corresponding points p_1, p_2 and p_3 in images I_1, I_2, and I_3, respectively, and $I(p_1)$, $I(p_2)$, and $I(p_3)$ are the pixel values of p_1, p_2 and p_3, respectively. When the number of reference camera is three, blending the color of the reference images for all pixels may blur the virtual view image. Therefore the pixel value of the nearest camera is used for the edge pixels.

Our target is a dynamic event in a large space like a soccer match. The objective is to render the entire scene including stadium and players from a novel viewpoint. According to the assumption such that the most of the scene can be approximated with some static planar areas, and only players are dynamic small areas, virtual view images can be synthesized with view interpolation technique as shown in Figure 3.2. As the object scene contains many objects, the entire scene is segmented into several regions, and the appropriate projective transform is utilized in each region for view interpolation. The projective geometry between cameras such as fundamental

matrices and homographies is estimated in advance from captured images.

The soccer field is considered as static except change of lighting condition. The virtual view images of ground, goal, and spectators' seats are synthesized once in an image sequence, and are updated when the lighting condition is changed. On the other hand, view interpolation is performed in every frame for dynamic regions such as player and ball because their motions change over time.

Considering structure of soccer scenes, the static regions can be segmented into several plane regions. One is the background region, which can be approximated as one plane located far from cameras. The others are field regions such as the ground and the goal, which can be approximated as sets of planes. This segmentation is manually operated only once because the cameras are fixed. In the field regions, the correspondence for view interpolation is obtained through homography transforms of the each plane region. The pixel values of synthetic image are updated according to the light condition. In the background region, image mosaicing technique is used instead of view interpolation.

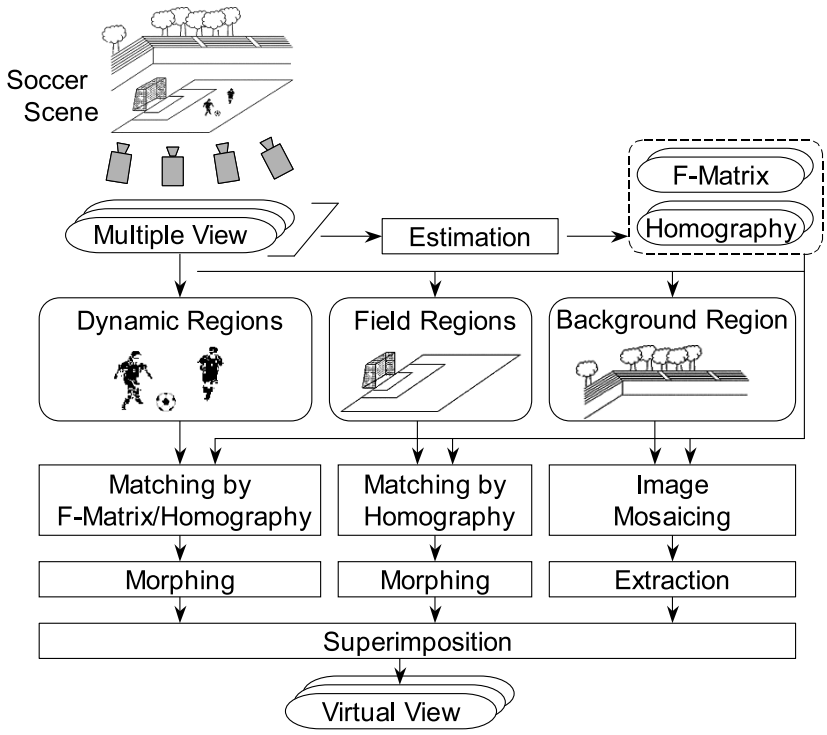


Figure 3.2: View synthesis for soccer scenes.

The dynamic regions can be extracted from entire scene by background subtraction. Once every player region is segmented and labeled automatically, the labeled regions of the same player are corresponded in the neighboring views through homography transforms of the ground plane among the views. Dense correspondence inside every region of the players and ball is obtained by applying fundamental matrices. The shadow of player/ball is projected on the ground, so that the correspondence is computed by homography of the ground plane.

Finally, the appearance of the entire scene from a novel viewpoint is given by superimposing the virtual view images in the order of background region, field regions, shadow regions, and dynamic regions.

3.2 Estimation of Projective Geometry

3.2.1 Fundamental Matrix

The epipolar geometry [30, 43] between two cameras is represented by the fundamental matrix (F-matrix) $\boldsymbol{F}$, which is a 3×3 matrix. If a point X in 3D space is projected onto image x_1 in the first view and x_2 in the second, the corresponding image points satisfy the following equation:

$$\tilde{\boldsymbol{x}_2}^{\top} \boldsymbol{F} \tilde{\boldsymbol{x}_1} = 0 \ , \tag{3.5}$$

where $\tilde{\boldsymbol{x}_1}$ and $\tilde{\boldsymbol{x}_2}$ are the homogeneous coordinates of x_1 and x_2, respectively. $\boldsymbol{F}$ is a rank 2 homogeneous matrix with 7 degrees of freedom; hence, it can be computed nonlinearly by at least seven correspondences in the two views.

Considering the search for corresponding points in stereo matching, the search area can be reduced by this geometry. Assuming that a point p is known in the first view, the corresponding point in the second must lie on the epipolar line l obtained by

$$\tilde{\boldsymbol{l}} = \boldsymbol{F} \tilde{\boldsymbol{p}} \ , \tag{3.6}$$

where $\tilde{\boldsymbol{l}}$ and $\tilde{\boldsymbol{p}}$ are the homogeneous coordinates of l and p, respectively. Therefore, the search does not need to cover the entire image plane and can be restricted to the epipolar line (see Figure 3.3). In the proposed method, the fundamental matrix is employed for obtaining a dense correspondence for the dynamic regions.

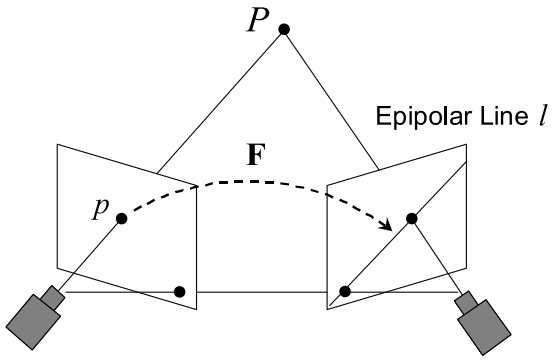


Figure 3.3: Fundamental matrix.

3.2.2 Homography

Image points on a plane in the first view are related to their corresponding image points in the second view using a homography $\boldsymbol{H}$, induced by a world plane, as

$$\tilde{\boldsymbol{p}_2} \cong \boldsymbol{H}\tilde{\boldsymbol{p}_1} \ , \tag{3.7}$$

where $\tilde{\boldsymbol{p}_1}$ and $\tilde{\boldsymbol{p}_2}$ are the homogeneous coordinates of the corresponding image points p_1 and p_2 (see Figure 3.4), respectively, and $\boldsymbol{H}$ is a 3×3 matrix with 8 degrees of freedom; hence it can be computed by at least four correspondences in the two views. Through a homography transform, a point in one view determines a point in the other. The proposed method employs homography transform for obtaining dense correspondences in the static regions.

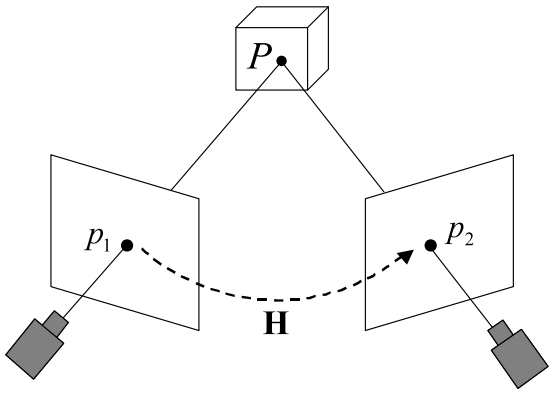


Figure 3.4: Homography.

3.3 Background Image Generation

In order to perform view interpolation in the static regions and the dynamic region separately, the background image where foreground objects, such as players and ball, do not exist is required for each real camera. If such background images can be captured before/after a soccer match, they are used for view synthesis for the static regions and background subtraction to extract the dynamic regions. Otherwise the background images are synthesized from an image sequence. We describe the method how to generate the background image for the soccer scene with reviewing already proposed methods.

Background subtraction is a widely used approach for detecting moving objects in videos from fixed cameras. In video surveillance, the object detection should be robust to variation in illumination conditions caused by weather, time of day. Several methods for performing background modeling or background subtraction have been proposed. These methods try to estimate the background model from the temporal sequence of the frames in several ways.

Wren et al. [125] have proposed to model the background independently at each pixel. The model is based on ideally fitting a Gaussian probability density function (pdf) on the last n pixel's values. However, the scene background is not completely static in outdoor environments with moving trees and bushes. For example, one pixel can be the image of the sky in one frame, a tree leaf in another frame, a tree branch in a third frame, and some mixture subsequently. Since the pixel has a different intensity (color) in such situation, a single Gaussian is not an adequate model. Instead, a mixture of Gaussians has been used to model such variations [109].

Elgammal et al. [27, 28] proposed to model the background distribution by a non-parametric model based on Kernel Density Estimation (KDE) on the buffer of the last n background values. Their method can deal with drawbacks such that the histogram might provide poor modeling of the background pdf. KDE guarantees a smoothed, continuous version of the histogram.

On the other hand, mean-shift vector techniques have recently been employed for various pattern recognition problems such as image segmentation and tracking [20, 21]. The mean-shift vector is an effective gradient-ascent technique that enables to detect

the main modes of the true pdf directly from the sample data with a minimum set of assumptions. However, it has a very high computational cost. Piccardi and Jan [83] proposed some computational optimizations promising to mitigate the drawback. In the method proposed by Han et al. [42], the mean-shift vector is used only for an off-line model initialization.

As a simple method, Lo and Velastin [63] and Cucchiara et al. [22] proposed to use the median value of the image sequence as the background, which performs better than use of temporal average. The methods mentioned above have different performance in speed, memory requirements and accuracy. The selection depends on the application requirement.

Considering extracting players in soccer scenes, modeling of the ground region is the most important for background subtraction. We select a simpler method that has less computational cost. The ground can be considered to be static except illumination changes. The mode value in the image sequence is used for the modeling. As players move fast on the ground in soccer scenes, the mode value at each pixel should be the color of the background. To deal with illumination changes, the background model is updated every predefined temporal sequence (we define this as 150 frames).

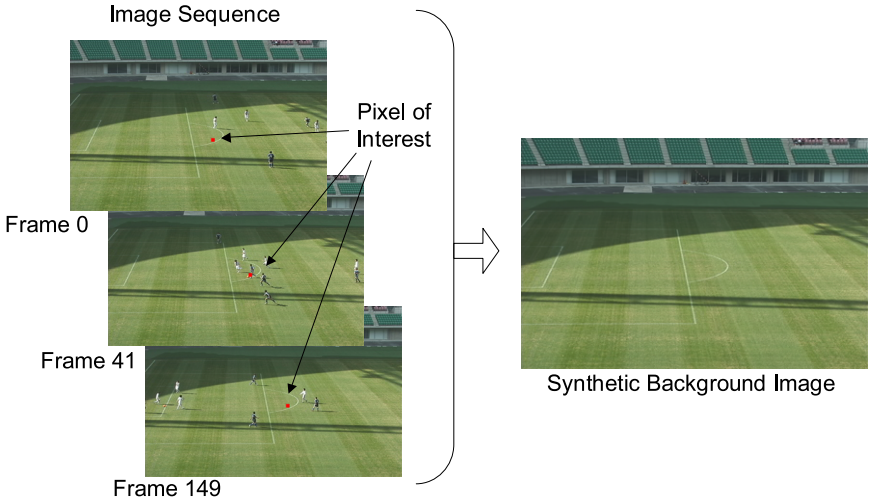
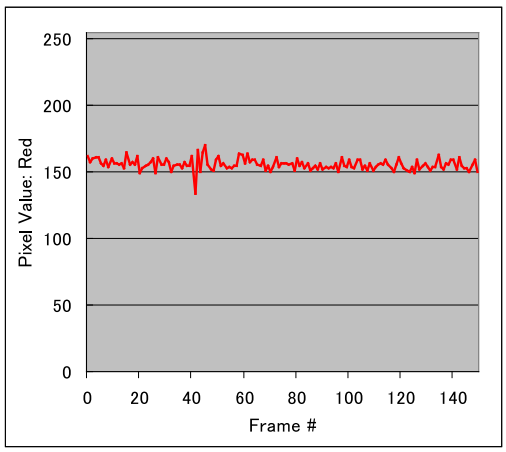
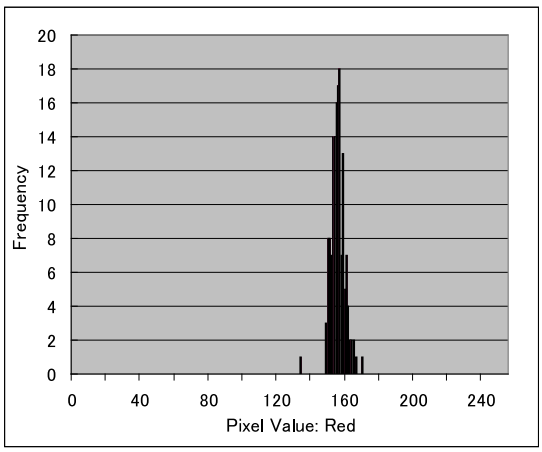


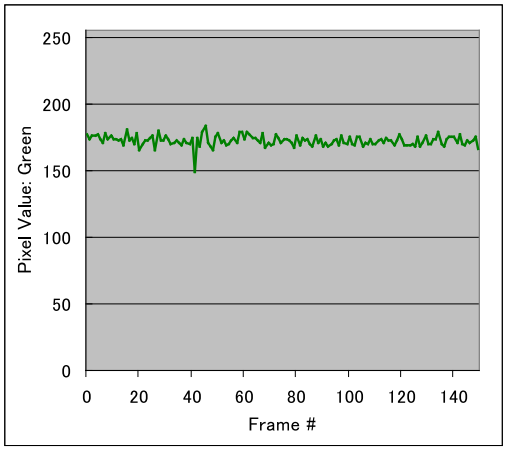
Figure 3.5: Background generation from the image sequence.



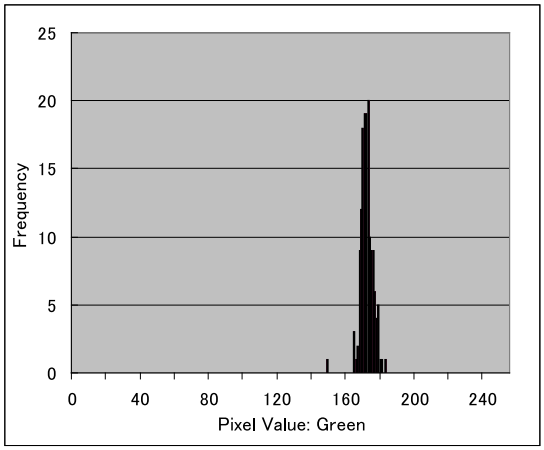
(a) Variation in pixel value (red)



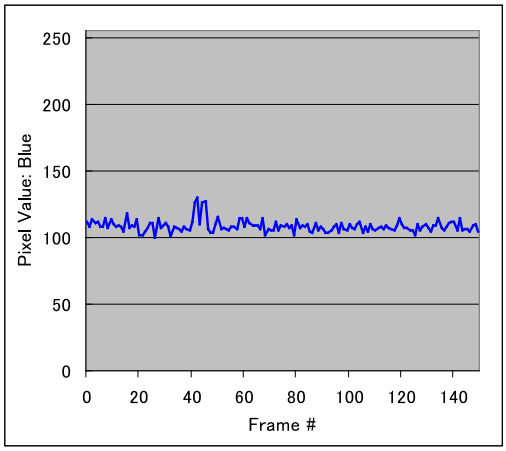
(b) Histogram of pixel values (red)



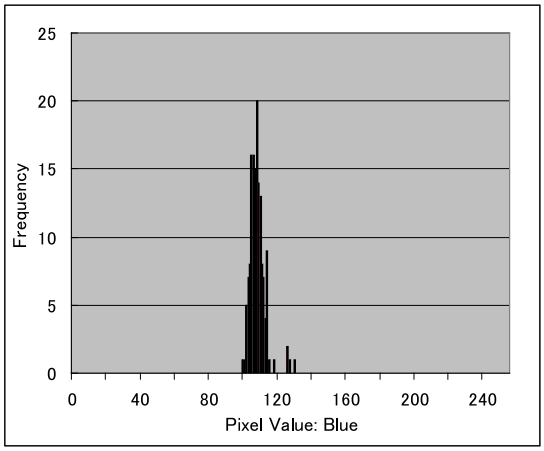
(c) Variation in pixel value (green)



(d) Histogram of pixel values (green)



(e) Variation in pixel value (blue)



(b) Histogram of pixel values (blue)

Figure 3.6: Variation in pixel value in the image sequence and the histogram.

Figure 3.5 presents the background image generated from the image sequence. The background model is well synthesized by utilizing the mode value of the image sequence for each pixel. Figure 3.6 shows variation in pixel value (red, green, and blue) in the image sequence and the histogram for one pixel whose location is indicated as a red point in Figure 3.5. A large change can be seen in the pixel value only when a player covers the pixel. The figure indicates that the use of mode value is appropriate for the sporting scene where the object movement is fast.

The generated background model is used for background subtraction. Additionally being used for view synthesis for the static regions, the background image is manually segmented into several plane regions that form the ground, goal, and spectator's seats.

3.4 View Synthesis for Static Regions

3.4.1 Field Regions

The ground and the goal can be considered as a single plane and a set of planes, respectively. We apply homography to the planes to obtain the correspondences required for the view interpolation. Equation (3.7) yields the pixel-wise correspondence for two views of a plane region. The homographies of the planes that represent the ground and the goal provide the dense correspondence within these regions. Virtual view images are synthesized based on the correspondence using view interpolation technique described in Section 3.1.

Figure 3.7 presents examples of virtual view images for the field regions. Figure 3.7 (a) and (d) show the real camera images, and (b) and (c) show the interpolated images from (a) and (d). The interpolating weights of the virtual view to the real views are 4 to 6 in (b) and 6 to 4 in (c), respectively.

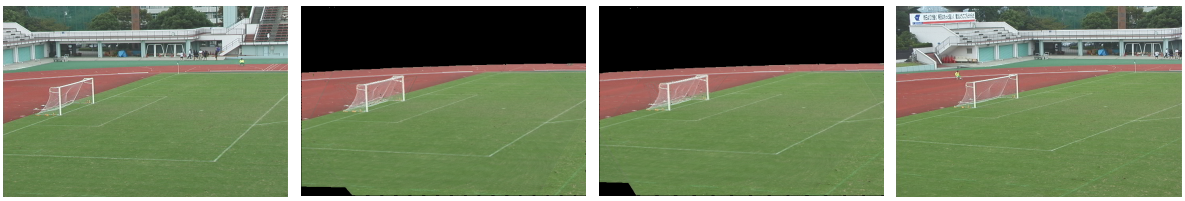


Figure 3.7: Examples of virtual view images for the field regions.

3.4.2 Background Region

The background is located far from the viewpoint positions of the cameras such that it can be considered as a single plane. It is not useful to apply morphing technique to this region because overlapping area is small. Instead, the technique of image mosaicing is employed for view synthesis.

Each of the two real camera images is composed in order to generate mosaic, which is the respective panoramic image of the background. Virtual view images are extracted from these panoramic images. Here, we assume that the backgrounds of the neighboring viewpoints have an overlapping region.

The composition starts with integrating the coordinate systems of the two views through the homography $\boldsymbol{H}_b$, which represents transformation from the first view to the second view, for the background. Next, the pixel values of the overlapping area are blended so that the pixel colors at the junction areas can smoothly connect the two backgrounds. The pixel value in the mosaic image is given by the following equation:

$$\acute{v} = \begin{cases} v_1 & (x < x_1) \\ (1-\beta)v_1 + \beta v_2 & (x_1 \leq x \leq x_2) \\ v_2 & (x > x_2) \end{cases}, \tag{3.8}$$

where

$$\beta = \frac{x - x_1}{x_2 - x_1} ,$$

v_1 and v_2 are the pixel values of I_1 and I_2, and x_1 and x_2 are the x-coordinates of the left hand side and the right hand side of the overlapping area, respectively (as

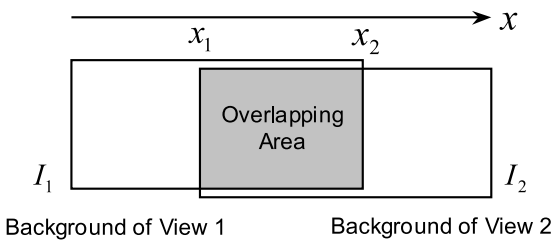


Figure 3.8: Image mosaicing.

shown in Figure 3.8). The partial area that is necessary for each virtual view is then extracted from the panoramic image. The following homography $\acute{\boldsymbol{H}}_b$ is used in the transformation of coordinates to complete the view synthesis.

$$\acute{\boldsymbol{H}}_b = (1-\alpha)\boldsymbol{E} + \alpha \boldsymbol{H}_b \ , \tag{3.9}$$

where α is the interpolating weight, and $\boldsymbol{E}$ is a 3×3 unit matrix. Figure 3.9 (a) and (b) illustrate the examples of background regions in real camera images, and (c) shows a mosaic image composed of (a) and (b). Figure 3.9 (d) and (e) present virtual view images for the background region, whose interpolating weights are 4 to 6 in (d) and 6 to 4 in (e), respectively.

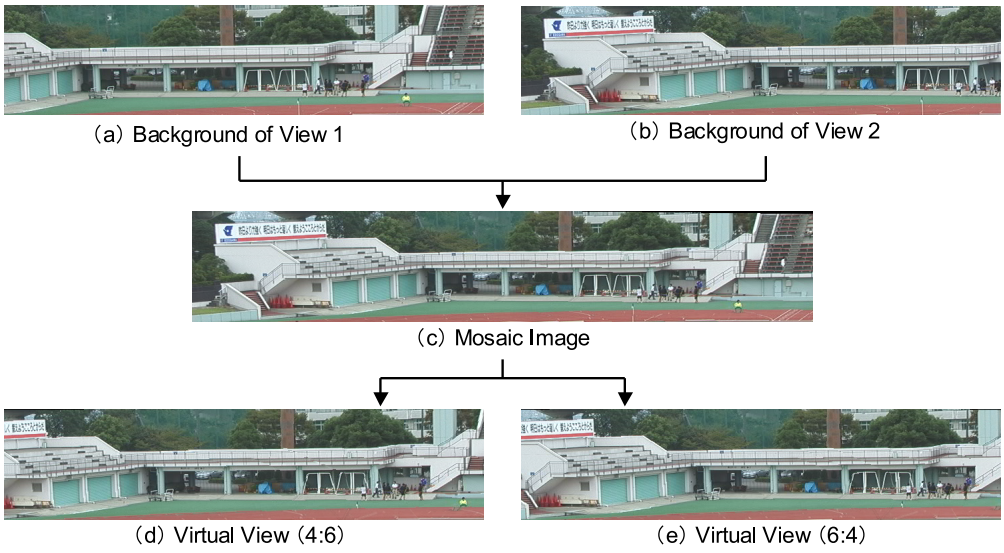


Figure 3.9: Examples of virtual view images for the background.

3.5 View Synthesis for Dynamic Regions

3.5.1 Overview

Figure 3.10 shows the flow of the process of view synthesis for the dynamic regions. The process in the case of view interpolation between two views is described. View interpolation is implemented for each frame because the shapes or the positions change over time in the dynamic regions. In every frame, all dynamic regions in two real cameras are extracted by subtracting the background from the original image. The extracted dynamic regions are segmented into player/ball regions and shadow regions using geometry and color information. View interpolation technique is applied to each region. Finally synthesizing virtual view image of each region completes view synthesis for the dynamic regions.

3.5.2 Extraction of Dynamic Regions

All the dynamic regions are extracted by subtracting the background from the original image. The image where neither the players nor the ball exists is used as the background of each camera. The segmentation of dynamic regions and static regions is sometimes difficult. Therefore, we extract dynamic regions by background subtraction using not only intensity but also color vector, containing 3 components: red, green, and blue. They are considered to be identical in pixels assigned to static

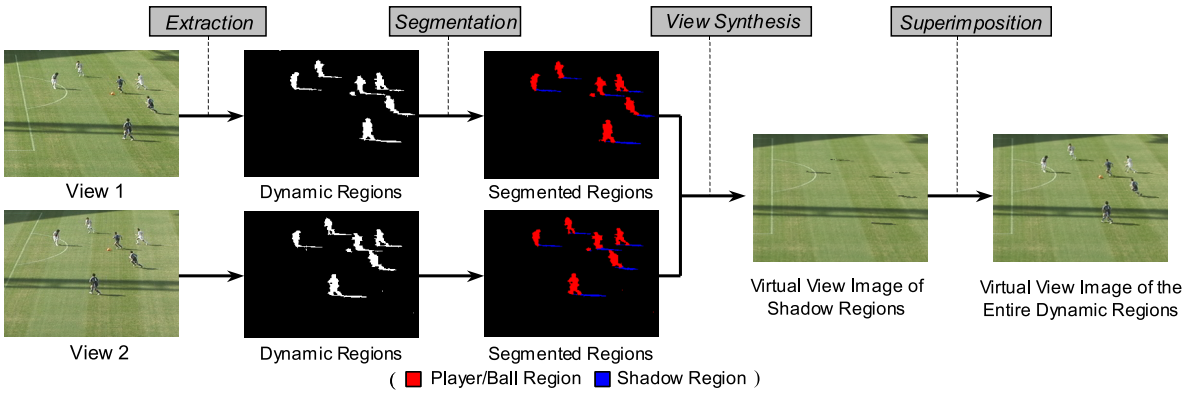


Figure 3.10: Process flow of the view synthesis for the dynamic regions.

regions between current frame image and background image while they vary in pixels of dynamic regions. Each pixel value in the silhouette image is determined as follows:

$$I_{sil} = \begin{cases} 1 & (\text{if } d_{val} > th_{val} \text{ and } d_{col} < th_{col}) \\ 0 & (\text{otherwise}) \end{cases} \tag{3.10}$$

where,

$$d_{val} = 0.299\ I_r + 0.587\ I_g + 0.114\ I_b\ ,$$

$$d_{col} = \frac{B_r I_r + B_g I_g + B_b I_b}{\sqrt{{B_r}^2 + {B_g}^2 + {B_b}^2} \cdot \sqrt{{I_r}^2 + {I_g}^2 + {I_b}^2}}\ ,$$

I_r, I_g, and I_b are the values of red, green and blue component in original image, respectively, and B_r, B_g, and B_b are the values of red, green and blue component in background image, respectively. th_{val} and th_{col} represent thresholds for determining if the pixel belongs the dynamic regions or not. Figure 3.11 shows the result of background subtraction. The dynamic regions are correctly extracted by the above method.

If view interpolation is applied to the sequence that has variations in lighting, we select a background with the same light condition. In this case, the extracted regions by background subtraction may contain shadows as well as the players/ball. View interpolation is performed additionally for the shadow regions. Using the conventional method, it is possible to synthesize shadows in another view by estimating the light sources in an environment; however, this is performed at a high cost of calculation.

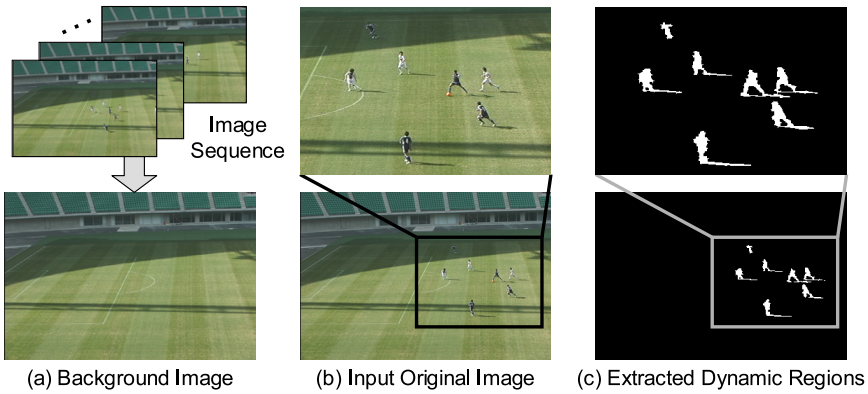


Figure 3.11: Extraction of the dynamic regions.

Alternatively, by applying the proposed method, we can project shadows on the virtual view image by transferring the shadow regions from the reference images using projective geometry between cameras.

3.5.3 Segmentation of Dynamic Regions

As a single scene usually contains a ball, several players, and possibly shadows, we deal with these dynamic objects separately. If shadows are included in the object scene, we first segment the extracted regions into the shadow regions and the player/ball regions. Both the geometric information and the color information are used for this segmentation. It is assumed that the shadow is usually projected on the ground in a soccer scene. Candidates for shadow regions are detected by applying homography of the ground plane to all the extracted dynamic regions in neighboring two view images. This detection based on the homography often includes a part of player's foot. Therefore, the pixel color is also used for shadow extraction. HSI (Hue, Saturation, Intensity) transform is applied to the candidates in each view image. The hue of the pixel is almost identical in the shadow regions between the current frame image and the background image, while it is different in the player/ball regions.

Figure 3.12 exhibits the segmentation results, where the combined method of geometric transform (homography transform) and color transform (HSI transform) is compared with the method using only homography transform or HSI transform. It is evident that the combined method is better than the independent methods at segmenting the dynamic regions into shadows and players/ball.

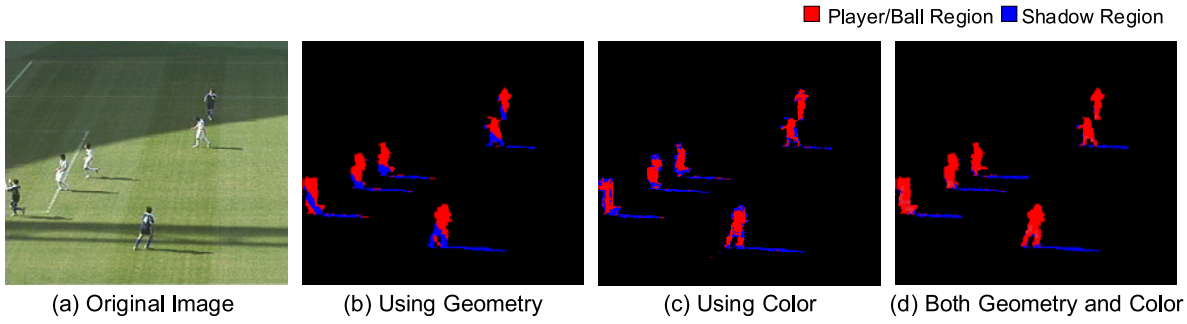


Figure 3.12: Segmentation of the dynamic regions.

3.5.4 Shadow Region

After segmentation, view interpolation is applied to the shadow and the player/ball regions, respectively. Since the shadows are considered to be projected on the ground, homography transform is applied to the shadow regions as well as the field regions. The virtual view images for shadow regions are synthesized using the homography of the ground plane as explained in Section 3.4.1.

3.5.5 Player/Ball Region

The method for generating virtual view image for the player/ball regions is described below. After the silhouettes of all the players and the ball are extracted, the labeling process segments each player and the ball. Subsequently, the corresponding silhouettes are obtained using the homography of the ground plane as shown in Figure 3.13. This is based on the assumption that one foot of a player is always in contact with the ground. Even if a player jumps, the error caused by the jump is sufficiently small; therefore, the homography of the plane that represents the ground can still locate the corresponding silhouettes. Some players, however, may not have one to one correspondence due to occlusion. In such a case, the segmented silhouettes in the previous frame are used for the segmentation of the players. As shown in Figure 3.14, the foot position of the occluded player is calculated by the homography of the

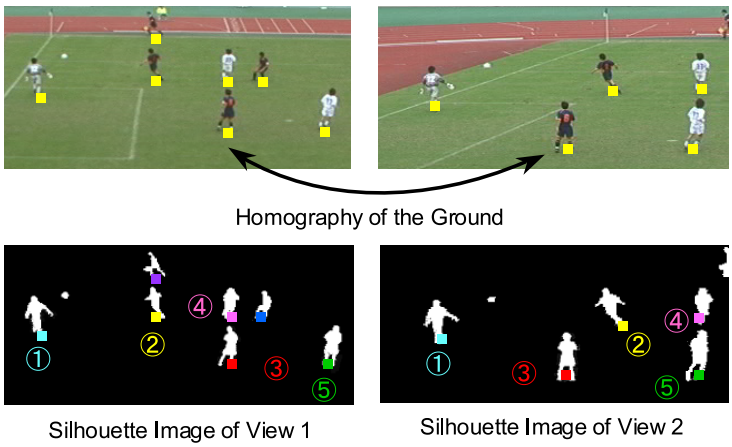


Figure 3.13: Silhouette correspondence.

ground plane from the neighboring view. The bounding box (rectangle surrounding the segmented player) is then projected onto the current frame from the previous frame. Thus, the occluded player can also have a correct correspondence. If the occlusion is detected in both views, the players are treated as one large object.

The pixel correspondence within the silhouettes is obtained by drawing epipolar lines in two different views, view 1 and view 2, using the fundamental matrix. On each epipolar line, the correspondences of intersections with boundaries, such as a_1 and a_2, and b_1 and b_2 in Figure 3.15, are obtained first. The correspondences between the pixels within the silhouette are obtained by linear interpolation of the intersection points. After the dense correspondence within the silhouette is obtained, the pixel positions and values are transferred from the source images of view 1 and view 2 to the target image by image morphing in the same way as in the field regions. However, view interpolation only generates virtual view images, where the zoom ratio is identical to that of real cameras. In order to provide zooming effects in free-viewpoint observation, it is necessary to control the 3D position of the virtual camera or its focal length. As the proposed method, which is based on view interpolation, cannot directly deal with the extrinsic and intrinsic parameters, we deal with a zooming feature by expanding or contracting images. View interpolation is modified as given by the following equation,

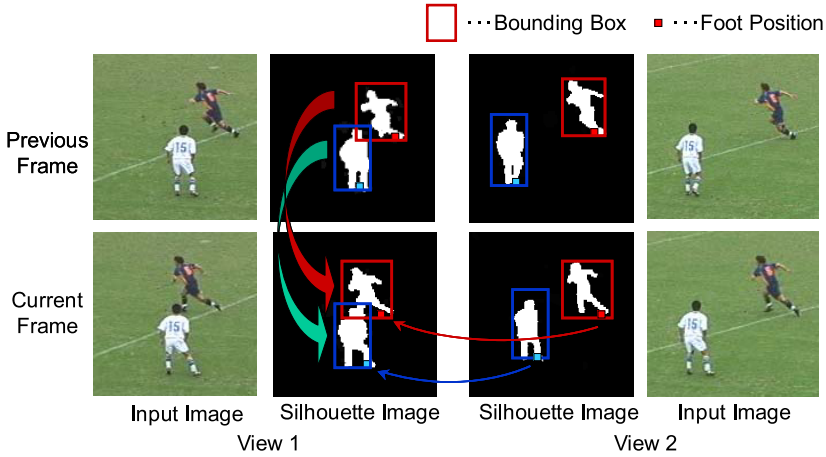


Figure 3.14: Silhouette correspondence in the case of occlusion.

instead of Equation (3.1).

$$\acute{\boldsymbol{p}} = (1-\alpha)\left[(\boldsymbol{p}_1 - \boldsymbol{c}_1)\frac{\acute{f}}{f_1} + \boldsymbol{c}_1\right] + \alpha\left[(\boldsymbol{p}_2 - \boldsymbol{c}_2)\frac{\acute{f}}{f_2} + \boldsymbol{c}_2\right] \ , \tag{3.11}$$

where $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are the coordinates of the principal points in images I_1 and I_2, respectively, and f_1 and f_2 are the focal lengths of cameras 1 and 2, respectively. $\acute{f}$ represents the focal length of the virtual camera. This equation enables zooming in or out approximately by expansion and contraction using the ratio of the focal length of the real camera to the focal length of the virtual camera. The pixel value is transferred using Equation (3.2). Virtual views are generated by blending the two warped images. The above algorithm is applied to every pair of silhouettes. After synthesizing them in order of distance from the viewpoint, all player/ball regions are overlaid onto the shadow regions. This concludes view interpolation for dynamic regions.

Finally, superimposition of the images, in the order of background region, field regions, and dynamic regions, completes the virtual view image of the entire scene. Figure 3.16 presents the reconstruction of the player from different angles. Not only the global appearance of the entire scene but also the local appearance of the player can be represented to a great extent.

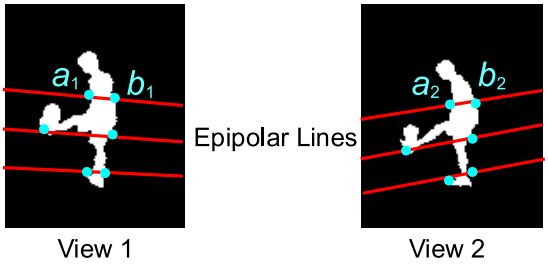


Figure 3.15: Pixel correspondence.

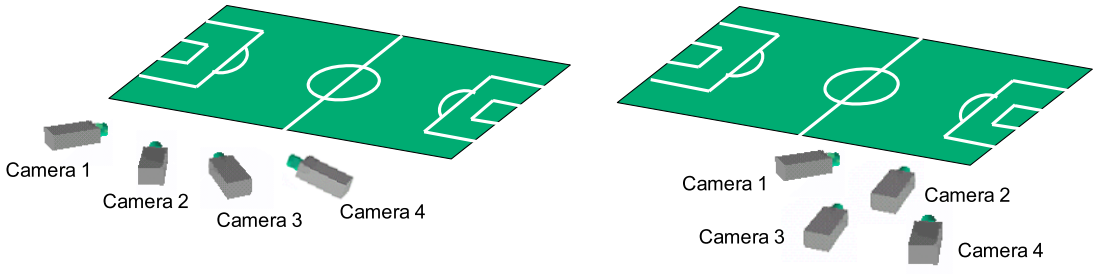
Figure 3.16: Reconstruction of the player from different angles.

3.6 Experimental Results

3.6.1 Image Acquisition

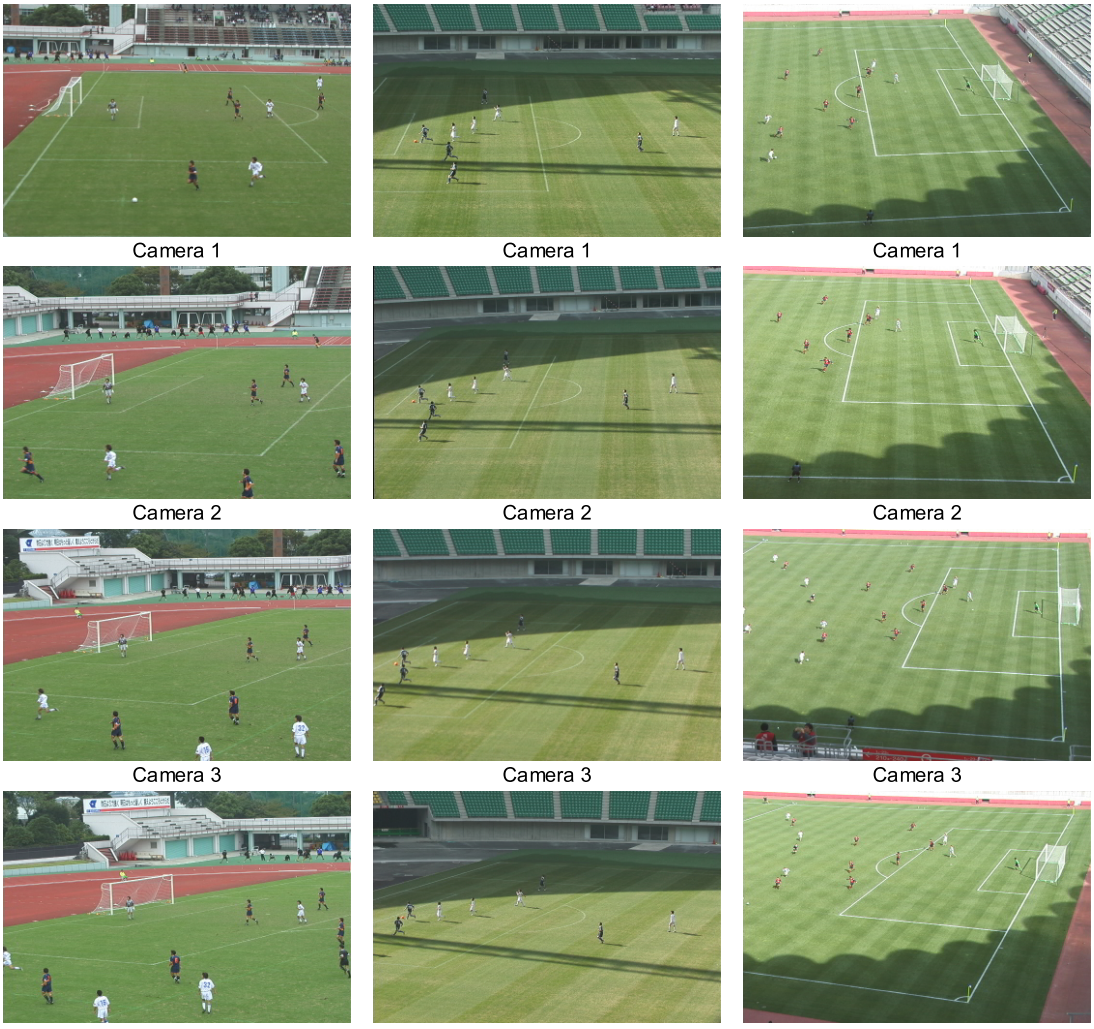
We have applied the proposed method to several image sequences of actual soccer matches captured using multiple video cameras at three kinds of soccer stadiums: the Edogawa Athletics Stadium in Tokyo, the Kashima Stadium in Chiba, and the Oita Stadium in Oita, Japan. As shown in Figure 3.17, a set of 4 fixed cameras was placed on one side of the soccer field in all three stadiums in order to capture the penalty area. Neighboring cameras had an overlapping region of the background for image mosaicing. The captured videos were converted to BMP format image sequences, composed of 720×480 pixels, 24-bit RGB color (8-bit per color) images, and then used for virtual view synthesis. Figure 3.18 describes example of multiple view images captured in each stadium.

The fundamental matrices between the viewpoints of the cameras and the homographies between the planes in the neighboring views were computed using the corresponding points. In this experiment, we manually selected about 50 corresponding points, whose 3D positions varied in the object space, for fundamental matrices and 20 points on each plane for homographies in the image sequence between neighboring views.



(a) Edogawa Athletic Stadium, (b) Oita Stadium (c) Kashima Stadium

Figure 3.17: Camera configuration in the stadiums.



(a) Edogawa Athletic Stadium (b) Oita Stadium (c) Kashima Stadium

Figure 3.18: Examples of multiple view images captured in the stadiums.

3.6.2 Results on Real Images

Figure 3.19 presents some results of the synthesized virtual view images for the soccer scene captured in the Edogawa Athletic Stadium. Figure 3.19 (a), (f), (k) and (p) present images captured using real cameras and the others present virtual view images generated by the proposed method. Figure 3.20 shows the close-up view of Figure 3.19. The position of players and the location of the background gradually change depending on the angle of the virtual viewpoint, which is determined by the interpolating weights between two real camera viewpoints. For example, the virtual viewpoint of (b) is located at a position whose relative weight is 2 to 8 between cameras 1 and 2. Although our method involves the rendering of separated regions, the synthesized images appear very realistic because the boundaries between the regions are not visible.

Figure 3.21 compares the virtual and real camera images for one frame in soccer scenes captured in the Edogawa Athletic Stadium. Figure 3.21 (c) and (e) show the virtual view image generated from the real camera images (a) and (b). Figure 3.21 (d) and (f) show the real camera image whose position is close to but does not coincide with the position of the virtual camera. By comparing the virtual and real views, realistic images at virtual viewpoint can be obtained without distortion or holes. The player regions and the field regions captured by the two real cameras have been correctly reconstructed in the virtual view image. Slight differences in the position of the players arise from the difference in the viewpoint position.

Next, we have applied the proposed method to three view images captured in the Kashima Stadium. Figure 3.22 presents the results of view interpolation among three cameras. The soccer scene including the shadows is well represented from the virtual viewpoints.

We have also obtained results for other scenes including shadows captured in the Oita Stadium, where view interpolation is performed between two views (see Figure 3.23). Figure 3.23 (c) shows the resultant image when view interpolation is applied to shadow and player/ball regions separately after segmentation, while (d) shows the result without segmentation. In the case without segmentation, the player/ball region and the shadow region are dealt with one dynamic object in process of the

view interpolation. Although Figure 3.23 (d) lacks a part of or the entire shadows of the players, all shadows are projected correctly in (c). This comparison shows that we successfully represented scenes including shadows in another viewpoint by applying view interpolation to shadow region and player/ball region separately.

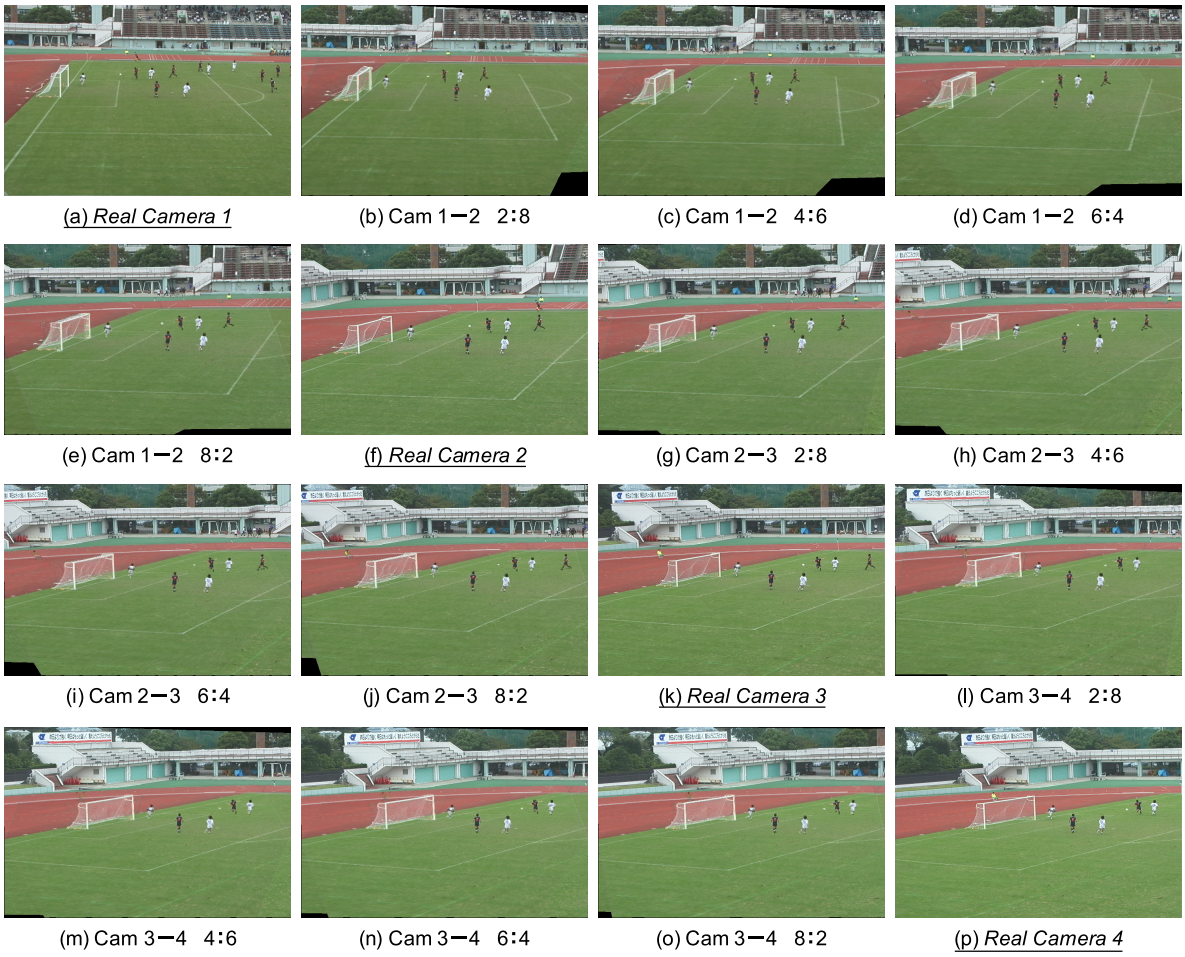


Figure 3.19: Synthesized virtual view images at one frame for the entire soccer scene from real camera images in the Edogawa Athletic Stadium.

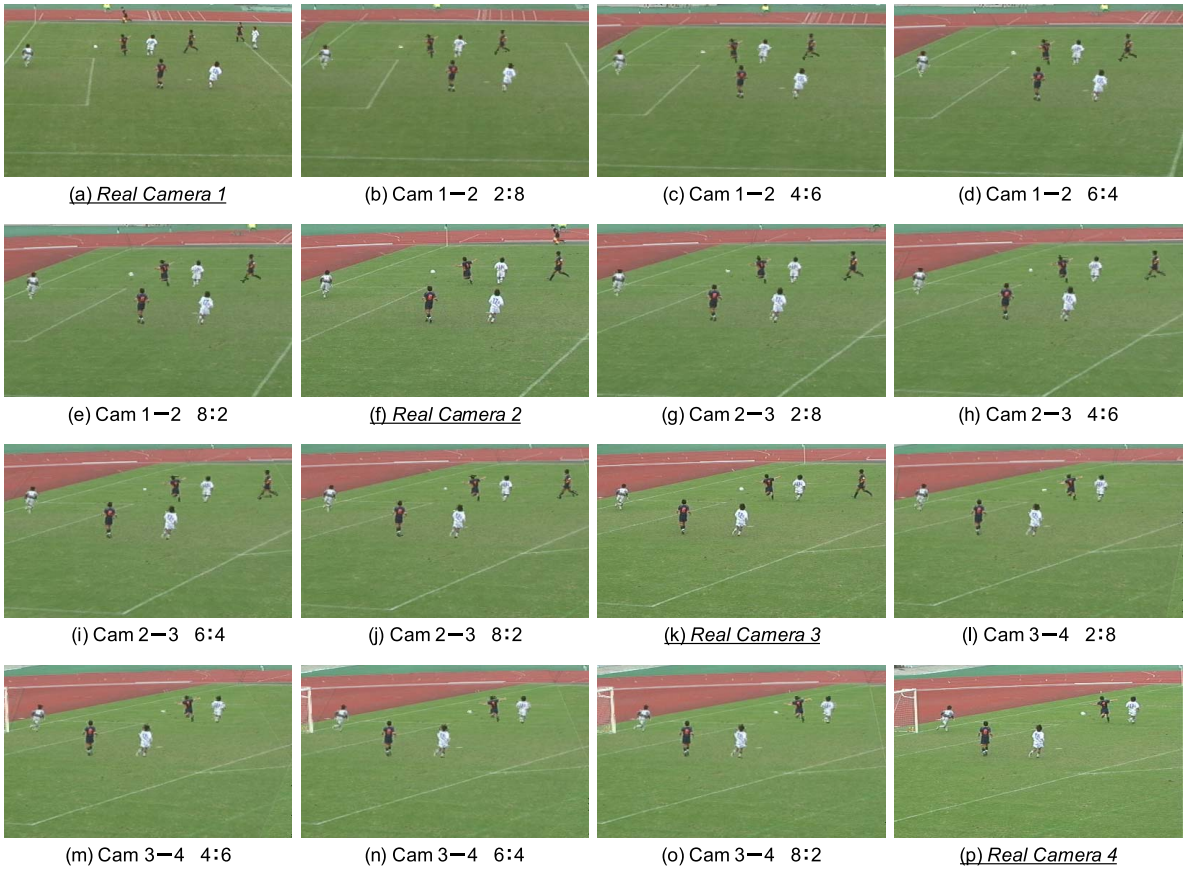


Figure 3.20: Close-up views of the previous figure.

(a) Real Camera 2 (Reference Camera A)

(b) Real Camera 4 (Reference Camera B)

(c) Virtual Camera Image (Camera 2-4 5:5)

(d) Real Camera Image (Camera 3)

(e) Close-up View of (c)

(f) Close-up View of (d)

Figure 3.21: Comparison between the virtual and real camera images on the real soccer scene.

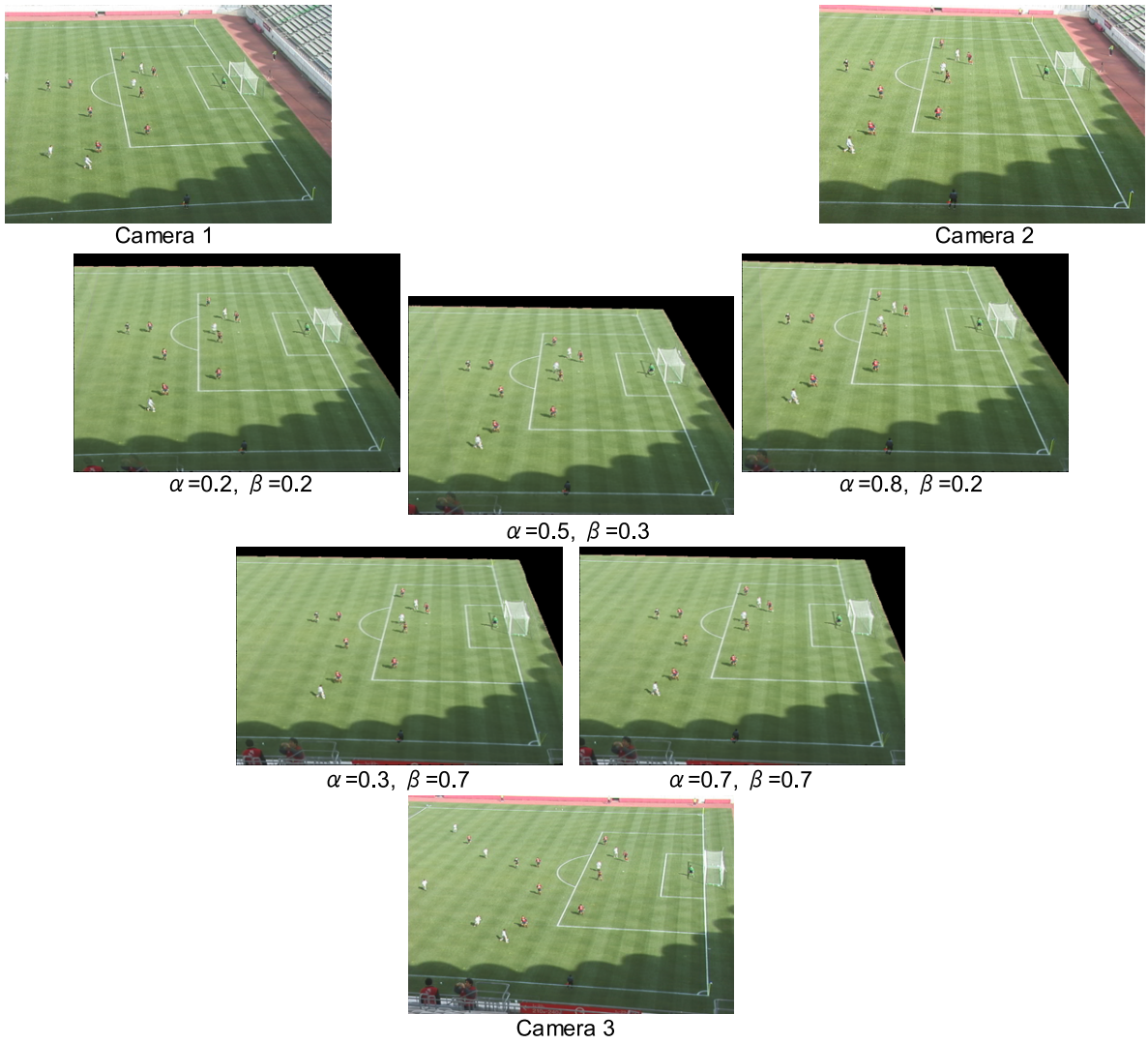


Figure 3.22: View interpolation among three views for the soccer scene captured in the Kashima Stadium.

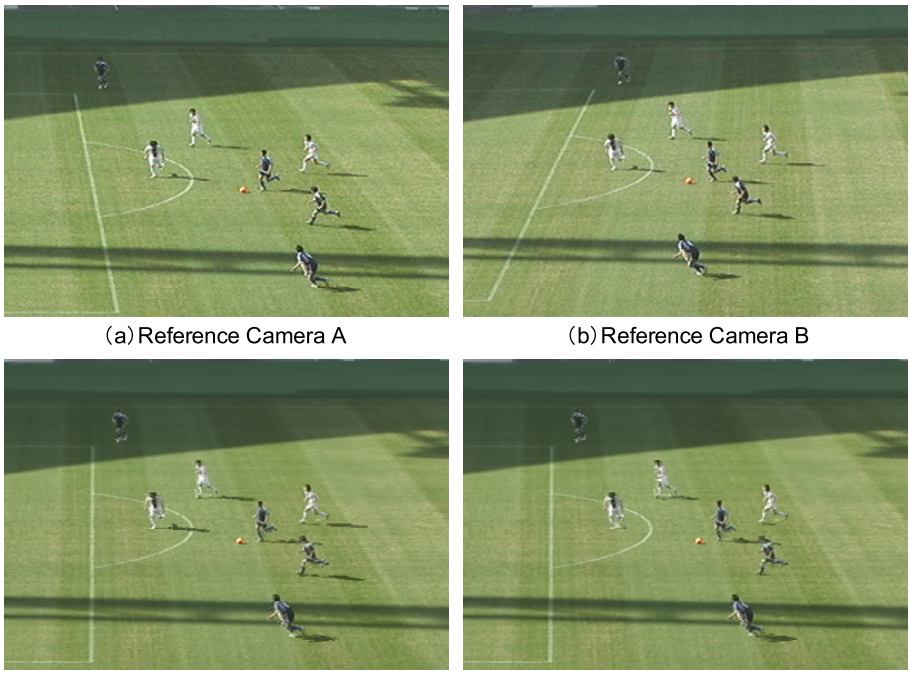


Figure 3.23: Comparison of the synthesized virtual view images with/without segmentation of player regions and shadow regions in view interpolation.

3.6.3 Results on Synthetic Images

We also have applied the proposed method to a synthetic scene for evaluation. As seen in Figure 3.24, the proposed method is applied to two view computer-generated images drawn by OpenGL, where four cuboids are placed on one plane. Figure 3.24 (c) shows the synthesized image generated by the proposed method from (a) and (b) with an interpolating weight value of 0.5. This result is synthesized by superimposing the virtual view image for the cubical region on the virtual view image for the plane region. Figure 3.24 (d) shows the image drawn by OpenGL from the same viewpoint as (c). The color difference between (c) and (d) is presented in (e). Although errors can be seen on the edges of the objects, most of the areas in the synthesized image are almost identical in appearance. This result indicates that the proposed method represents the objects at the correct positions in the virtual view image with certain color differences. The pixel correspondence error is responsible for a significant part of the color differences.

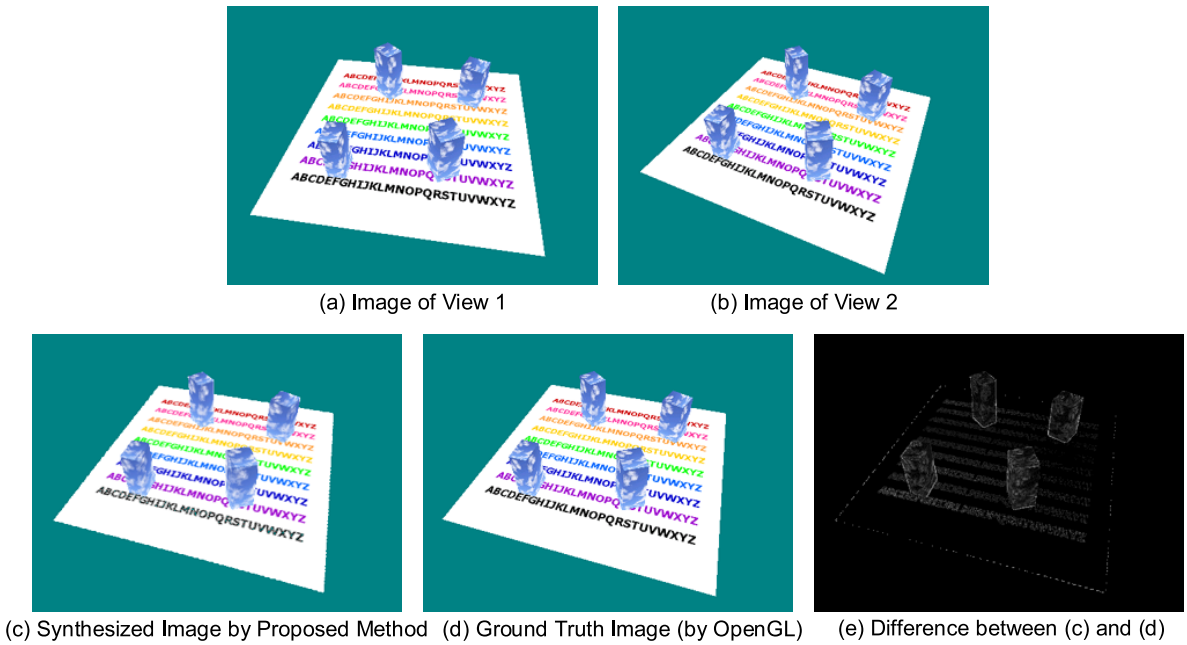


Figure 3.24: Comparison between the virtual view image and ground truth image on the synthetic scene.

3.6.4 Discussion

Firstly, geometric consistency in virtual view images is discussed here. In the proposed method, virtual view images are synthesized by transfer of pixel-by-pixel correspondence using linear interpolation between neighboring cameras. It is already known that shape distortion can be seen in linearly interpolated views except when the optical axes of two real cameras are parallel and angles of rotation around the optical axis are identical [102]. In our experiment, the cameras were placed at the upper deck in the stadium for direction of penalty area. The optical axes of the cameras were almost parallel, and the difference of the angles of rotation around the optical axis was very small. This is demonstrated by epipolar lines drawn in neighboring views. Figure 3.25 illustrates epipolar lines between neighboring cameras. These epipolar lines are close to parallel. This means that the optical axes of these cameras are close to parallel as well. Therefore natural-looking virtual view images without distortion were successfully synthesized. When applying view morphing method that enables to generate virtual view image without distortion, there is no big difference between the result by the proposed method and the result by view morphing. It turned out that the geometric consistency is preserved in virtual view images synthesized by the proposed method.

Next, change of the appearance in view interpolation is considered. In virtual view synthesis for the dynamic regions, epipolar geometry and the silhouette information are used for obtaining pixel correspondence. On each epipolar line drawn in neighboring view images, the edge points of the silhouette are corresponded firstly, and then

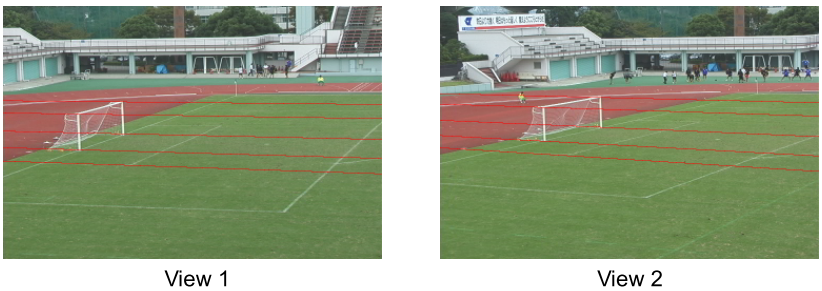


Figure 3.25: Epipolar lines between neighboring cameras.

the points within the silhouette are corresponded by linear interpolation of the edge points. In the case that the appearances of player are different between two views, the different points in 3D space are sometimes corresponded between two images through the above process. In the proposed method, it is assumed that players are located far from the cameras and the appearance of players does not vary greatly in neighboring views. Furthermore, the target of reconstruction in the proposed method is entire stadium. We focus on global appearance rather than detail texture of players. The effect of incorrect correspondence is considered to be sufficiently small for the global appearance. Hence the realistic images at virtual viewpoints were successfully synthesized. To improve the detail appearance, block matching using color information or edge matching using dynamic programming [79, 86] is supposed to be useful for obtaining the correct correspondence. Once the dense correspondence is correctly obtained, virtual view images that have improved texture can be generated by linear interpolation.

Subsequently, camera configuration is taken up. It is assumed that all the cameras capture the same target area, and that variations in lighting and scale across cameras are negligibly small. We manually adjusted the brightness and the focus of the multiple video cameras in the experiment so that the size of players and the overall colors in the captured scene can be almost identical across the cameras. In the neighboring camera images, an overlapping region of the background is required for image mosaicing. At the camera configuration shown in Figure 3.17 (a), four cameras were set at a distance of about 10 meters. This set up appears to be adequate for covering the penalty area. If more cameras are used, the quality of the synthesized image may be improved. It is difficult to formulate the theory that shows how many cameras is required or how long the distance between cameras that is suitable for reconstructing the scene is. It depends on the complexity of the object scene. This remains to be solved in the future.

Finally we clear up the limitations of the color of the object scene and manual work in the proposed method. The only restricting condition is that the colors of the uniform and the ball should differ from that of the ground. There is no particular limitation in the color of the players' uniform except the above condition. As soccer matches usually satisfy this condition, the proposed method can be applied to other soccer matches in other stadiums.

Some manual work is required in the current method. One of them is to give corresponding points for estimating projective geometry between cameras, that is fundamental matrices and homographies. It can be easily implemented by just clicking feature points on GUI. The other manual work is to specify the background region and the field region on the captures image in each camera. This process can be easily performed by generating mask images. The manual work mentioned above is required only once because the cameras are fixed.

Chapter 4:

Free Viewpoint Video Generation

4.1 Free Viewpoint Video

In the previous chapter, we have explained virtual view synthesis technique for dynamic events in a large space. We describe how to generate free viewpoint videos using the view synthesis method in this chapter. Free viewpoint video is synthesized by selecting two real cameras from multiple cameras as reference cameras, interpolating weight and zoom ratio in each frame for the image sequence.

Figure 4.1 shows flow of the process in every frame for free viewpoint video generation. As the static regions are considered to undergo little or no changes over time, view interpolation is implemented only once for the image, where neither players nor the ball is present. The virtual view images of the static regions are synthesized for every possible viewpoint beforehand. However, if the captured scenes have variations in lighting, the background image needs to be generated for every lighting condition in the sequence. In such a case, we synthesize the background image every 150 frames for the image sequence in advance.

On the other hand, view interpolation is implemented in every frame for the dynamic regions. After virtual view images of the players, ball, and shadows are generated by the method described in the previous chapter, they are superimposed onto virtual view image of the background. When this process is repeated every frame, free viewpoint video is produced. For example, a viewer can focus on a specific player in close-up view or may track a ball movement using a zoom-out virtual camera by selecting viewpoint appropriately.

Figure 4.2 presents an example of free viewpoint video that gives viewers the impression of fly-through over the soccer field or playing together in the soccer match by changing positions of the viewpoint with the ball movement. Another example is a video that produces a 3D effect of walking around an action scene as the movie “ The Matrix. ” Figure 4.3 shows the part of the image sequence of freeze-and-rotate virtual camera motion. We have created two videos to compare the proposed system and the “ Eye Vision ” system. This comparison indicates that rotating the virtual camera by interpolating intermediate viewpoints makes the video much more effective than just switching real cameras.

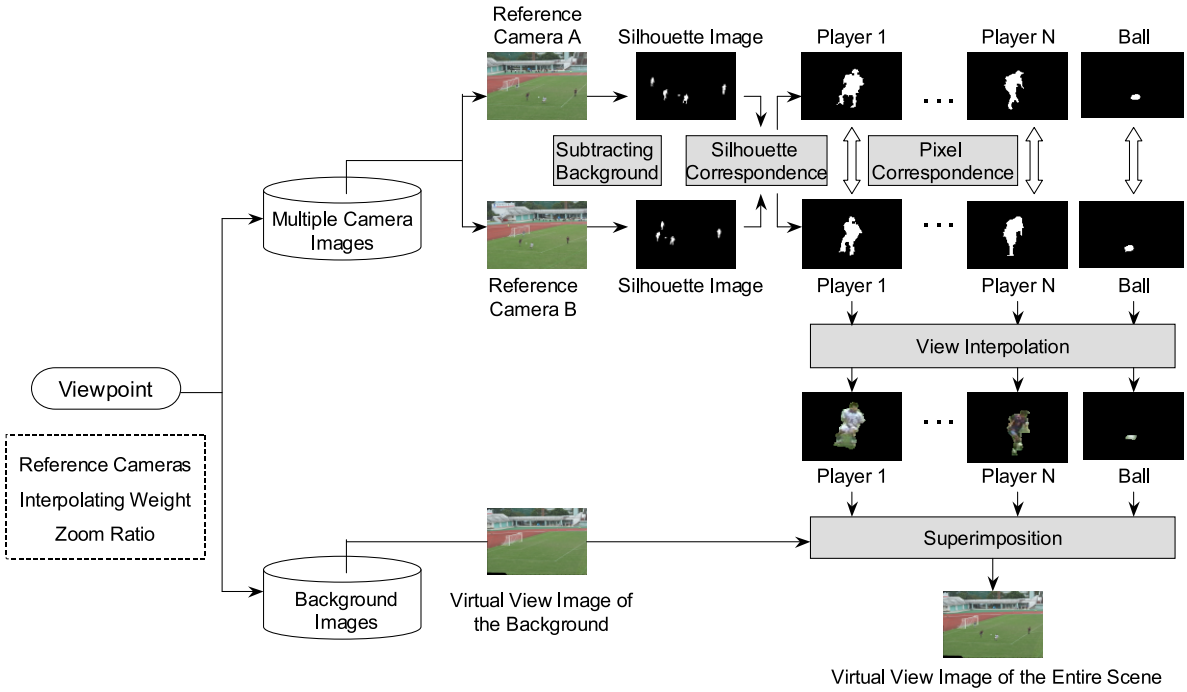


Figure 4.1: Process flow of the free viewpoint video generation.

Figure 4.2: Fly-through view image sequence.

Figure 4.3: Visual effect of the freeze-and-rotate camera motion.

4.2 Viewpoint on Demand System

4.2.1 System Overview

Existing television broadcasts only deliver pre-produced content wherein producers manually select video cameras for relaying sporting events; this is essentially a one-way communication. On the other hand, the Internet facilitates interactive communication between the broadcasting station and the viewers, in which the content can be interactively modified according to the viewers' demands. If the viewers can select preferred viewpoints, they will derive great enjoyment from watching the exciting scenes in these events. We introduce a system termed "Viewpoint on Demand System" as an example of such interactive communication media. The system, which allows viewer to select his/her favorite viewpoint during replay, consists of offline and online processes for rendering dynamic scenes effectively as shown in Figure 4.4.

4.2.2 Offline Process

A soccer match is captured using uncalibrated multiple cameras in a stadium, and the video images are stored in advance. The projective geometry used for view synthesis is estimated between neighboring cameras. The proposed system employs fundamental matrices between the cameras and homographies between the planes in neighboring views, which form the ground, goal, and background. The virtual view images of static regions are synthesized for all virtual viewpoints in each lighting condition. For every frame in the image sequence, dynamic regions are extracted from the captured image by subtracting the background. The extracted regions are segmented into players/ball regions and shadow regions for the view synthesis. In the player regions, the position of each player and the correspondence map of the players between neighboring views are obtained. Both the labeled images and the silhouette correspondence are stored at every two or three neighboring viewpoints for the online process.

4.2.3 Online Process

The online process is performed according to the interaction with a viewer. When the viewer selects the scene and the viewpoint, the stored information such as labeled images and silhouette correspondence regarding the two or three reference cameras near the selected viewpoint is loaded. The pixel correspondence within the silhouettes is obtained by drawing epipolar lines. The pixel correspondence in shadow regions is obtained by applying homography of the ground plane. The view interpolation is applied in the player, ball, and shadow regions for synthesizing virtual views. Finally the virtual view images in each region are superimposed on the virtual view images of the stadium at the corresponding viewpoint. The entire scene from the selected viewpoint is presented to the viewer.

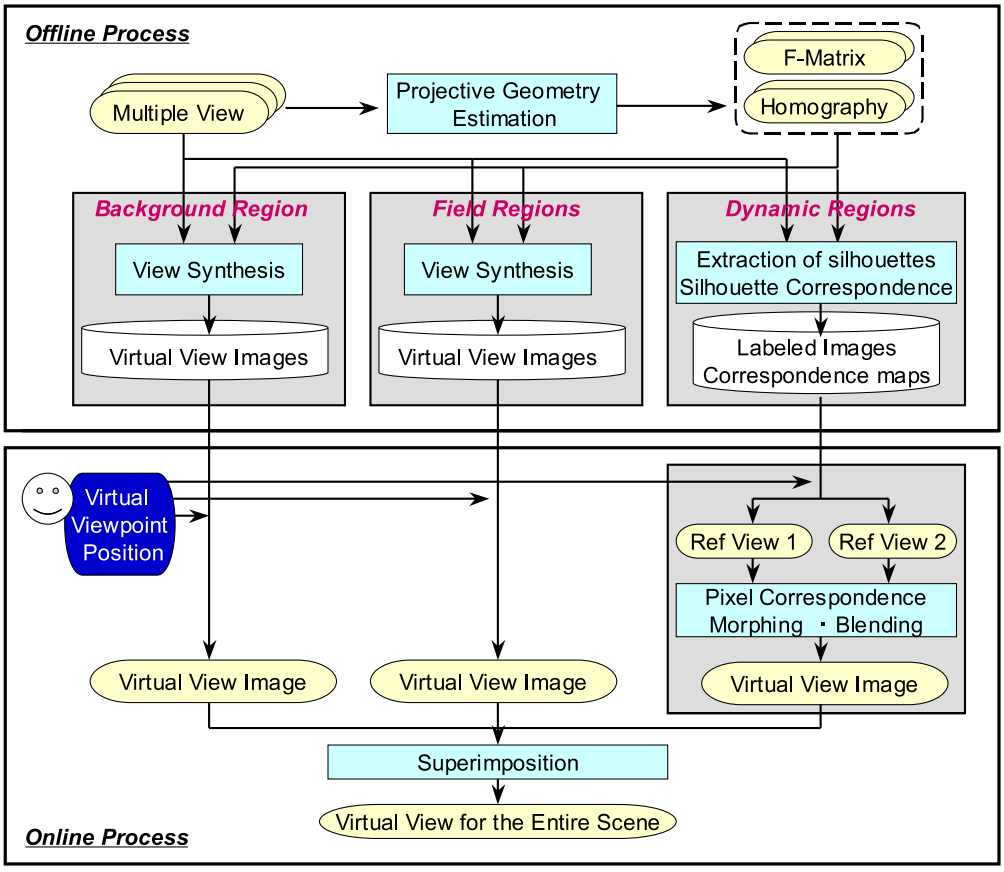


Figure 4.4: Offline and online processes in the Viewpoint on Demand System.

4.2.4 User Interface

Figure 4.5 presents the interface of the system. Two slider bars are provided for viewpoint selection. The horizontal slide bar at the bottom of the window determines the position of the virtual viewpoint, which is represented by the reference cameras and the interpolating weight α in Equation (3.11). The vertical slide bar on the right of the window determines the zoom ratio of the virtual camera to the real camera, which is represented by $\acute{f}/f_1$ and $\acute{f}/f_2$ in Equation (3.11). Once the viewer selects favorite scenes, rendering of the soccer scene starts with the position and zoom ratio that have been initially defined. The generated virtual view images are displayed in the center of the window, according to the viewpoint. While watching the video, the viewer can change the viewpoint at any time. He/She can move the viewpoint from right to left with the horizontal slide bar and zoom in/out with the vertical slide bar.

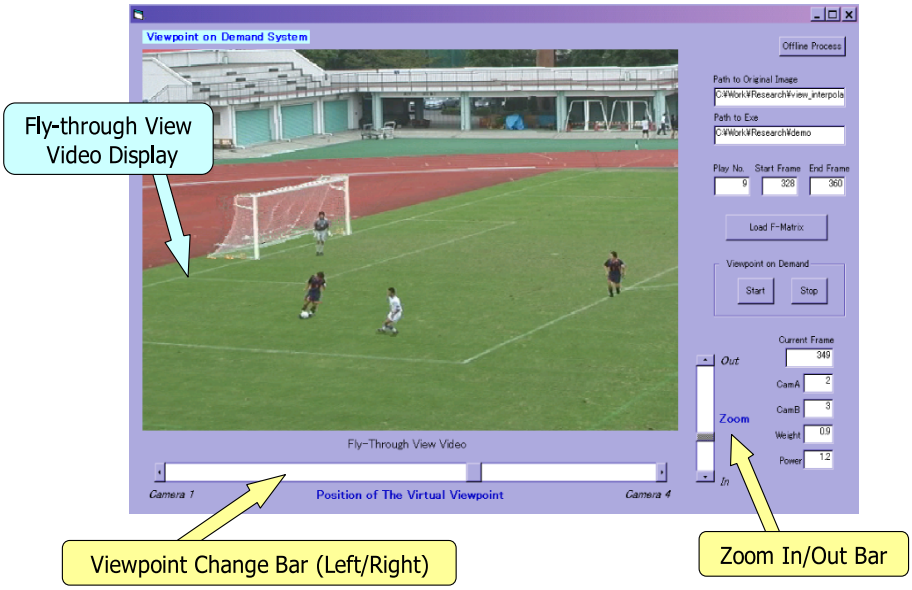


Figure 4.5: The interface of the Viewpoint on Demand System.

4.2.5 Experimental Results

Figure 4.6 presents examples of the images shown on the window of the proposed system, We virtually moved the camera from right to left by zooming in. For example, Figure 4.6 (a) shows the scene of frame number 322 where the virtual viewpoint is placed at the interpolating weight 4 to 6 between camera 3 and camera 4, and the zoom ratio of the virtual camera to the real camera is 0.8.

Figure 4.7 presents another example of the image sequence for free viewpoint replay. Frame 1462 and frame 1468 contain some occlusions, but the occluded players are constantly tracked, and their appearance is well synthesized. This application offers a new framework for presenting a soccer match on demand.

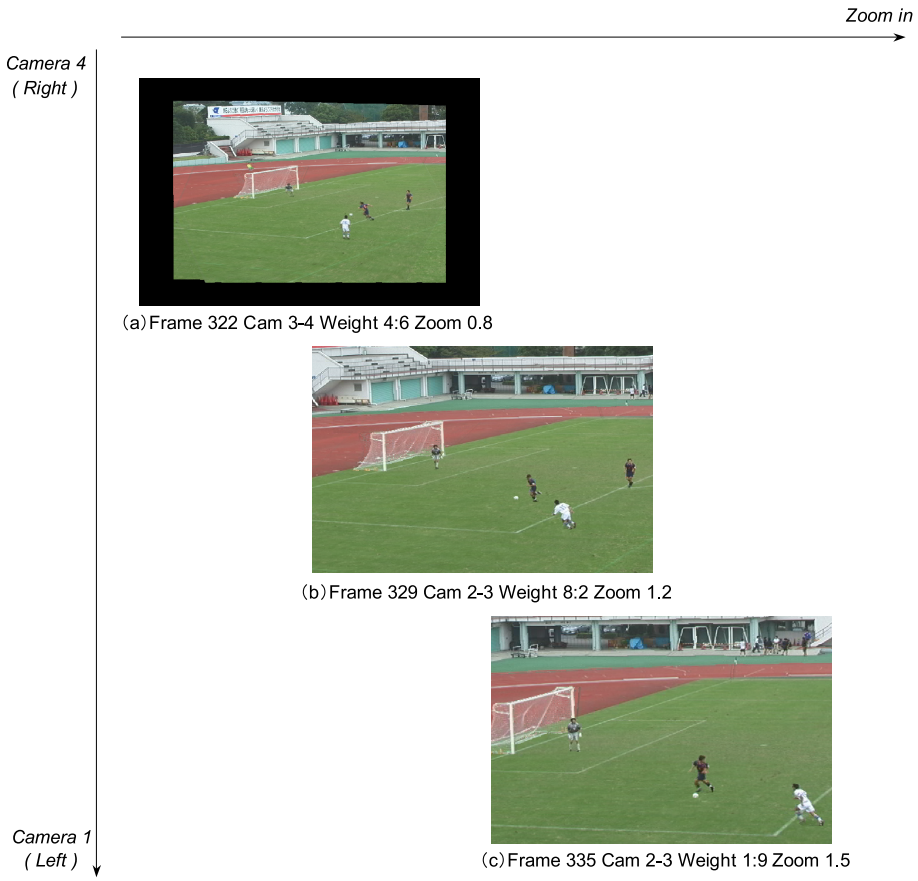


Figure 4.6: Examples of the image window of the Viewpoint on Demand System.

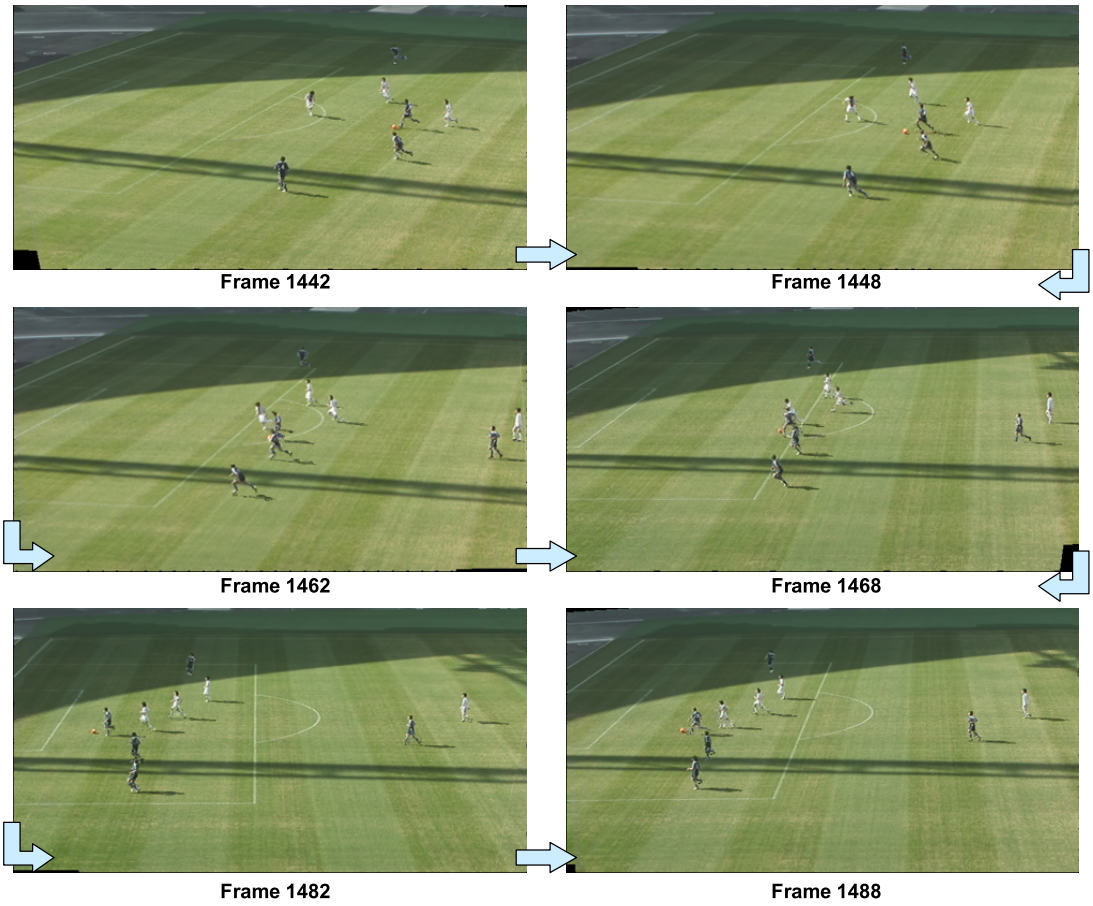


Figure 4.7: Example of free viewpoint replay for the sequence including shadows in the Viewpoint on Demand System.

4.2.6 Discussion

The performance of the Viewpoint on Demand System is examined below. The processing time was measured by using the desktop PC and the laptop PC which have the following spec.

- Desktop PC
  CPU: Pentium 4 3.2 GHz, Memory: 2 GB, Graphic Card: ATI Radeon 9800
- Laptop PC
  CPU: Pentium M 1.4 GHz, Memory: 768 MB, Graphic Card: ATI Mobility Radeon

The system runs at 3.7 fps with the desktop PC on an average while it runs at 1.8 fps with the laptop PC. The processing time depends on the number of dynamic objects in the output image.

Figure 4.8 shows the correlation between the processing time and the number of the objects. The horizontal axis represents the number of the dynamic objects in the image. The vertical axis represents the processing time. It turns out to be linear in the number of dynamic objects in the output image. This is because the

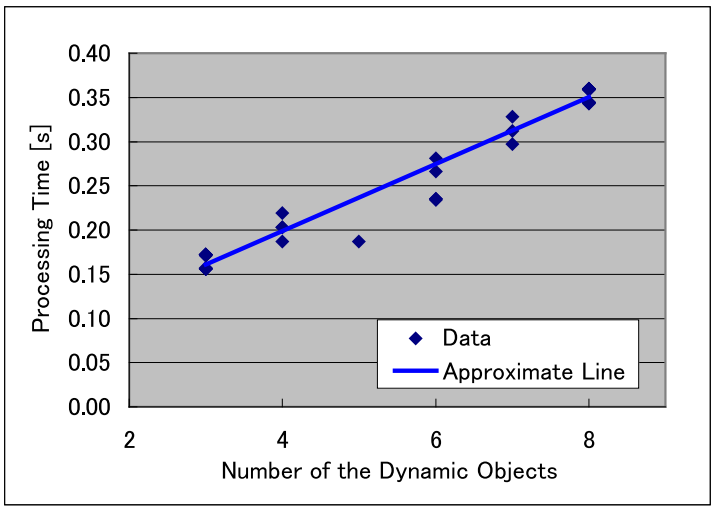


Figure 4.8: Correlation between the processing time and number of the objects in the Viewpoint on Demand System.

process for virtual view synthesis is performed sequentially for every player and ball. The correlation diagram indicates that applying parallel processing can reduce the processing time.

The limitation in free viewpoint video generation is considered here. Although the proposed method allows the viewer to watch a soccer match from his/her favorite viewpoint, there is a limitation in the range of viewpoint movement. The viewpoint on demand system gives the viewer to 2 DOF: pan and zoom. In the case of view interpolation between two cameras, virtual viewpoint movement is limited on the line connecting two cameras. In the case of three cameras, the viewpoint can move within the triangle formed by three cameras. If there are many cameras, which form triangles surrounding an object, the movable area of virtual viewpoint movement becomes large. It is important to arrange the camera configuration according to the desired system.

The proposed system cannot display close-up view such as facial expression of player because the objective of the system is to present global appearance of entire soccer scene from a novel viewpoint. In order to remove this limitation, some other cameras are required for capturing players. The players should be taken with close-up cameras and their virtual view images should be synthesized by applying an appropriate method for athletes. The resolution of the synthetic image can be improved by combing the virtual view synthesis using far cameras and close-up cameras.

The realized system sometimes causes failure. The segmentation/correspondence of the players fails when more than 4 or 5 players overlap; hence, a set play may be difficult situation for the view synthesis. It is essential to improve the system for such cases. One method for solving this problem is to use player information obtained from more than two cameras. If the proposed method is combined with a tracking method using multiple cameras or some sensors, for example the method proposed in [52], the accuracy of the segmentation and correspondence can be improved.

Chapter 5:

Mixed Reality Presentation

5.1 Expansion toward Mixed Reality

5.1.1 Overview of Mixed Reality

Mixed Reality (MR) is a technology that allows mixing the virtual synthesized world and the real physical world [2, 3]. In the broad meaning, it is not limited to the sense of sight. MR can potentially apply to all senses, including hearing, touch, and smell. This thesis, however, focuses on a visualization technology for mixture of virtual and real worlds.

There are two categories in MR: Augmented Reality (AR) and Augmented Virtuality (AV). In AR, computer-generated virtual objects are inserted in the real environment while real objects are added to the virtual environment in AV. Recently AR has drawn a lot of attentions as a tool for enhancement of the real world. For example, it is used to annotate objects and environments with public or private information. In medical application, doctors can use the augmentation technology as a visualization and training aid for surgery.

In MR environment, objects in the real and virtual worlds must be properly aligned with respect to each other, or the illusion that the two worlds coexist will be compromised. Many researchers have been working on these problems in order to generate the natural appearance of virtual-real mixed world.

In this thesis, the method of free viewpoint video presentation is expanded to the field of MR. Free viewpoint video is not displayed with the image of original stadium but overlaid on a desktop stadium model in the real environment. A novel approach

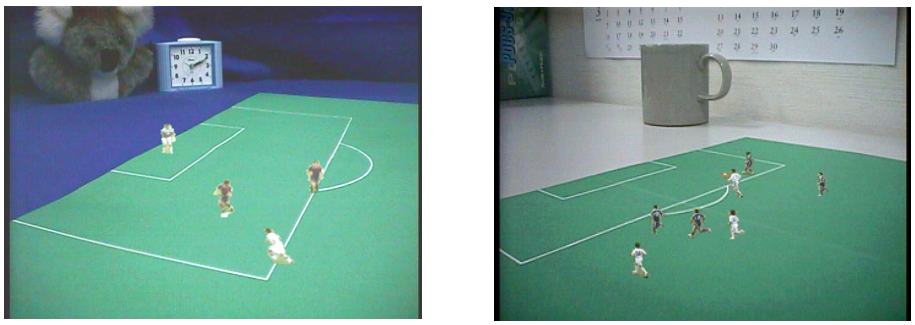


Figure 5.1: Example images of mixed reality presentation of a soccer match.

for inserting a virtual soccer match into the real environment is introduced. Figure 5.1 presents an example of MR presentation of a soccer match. Virtual soccer players and ball are overlaid on a small stadium in the real world. Virtual soccer scene is synthesized from real camera images using image-based rendering technique as shown in Chapter 3 and Chapter 4. This enables a viewer to watch realistic soccer scene in front of him/her in the real world.

5.1.2 Instruments

In MR environment, displays are required for viewing the merged virtual and real environments. The displays used to build a MR system can be classified in the following types.

- Retinal display
- Head-Mounted display
- Handheld display
- Head-Mounted projector
- Spatial display

Figure 5.2 shows the different types of displays where the displays are located with respect to the viewer and the real object.

Retinal displays [58, 91] utilize low-power semiconductor lasers to scan modulated light directly onto the retina of the human eye, instead of providing screens in front of the eyes. This produces a much brighter and higher resolution image with a potentially wider field of view than a screen-based display. However, only monochrome (red) images are presented since cheap low-power blue and green lasers do not yet exist.

Head-mounted displays (HMDs) are the common devices that are used to build an MR environment. HMD lets a user see the real world, with virtual objects superimposed by optical or video technologies (see Figure 5.3). Video see-through HMD uses video captured from cameras mounted to the display as a background for overlaying

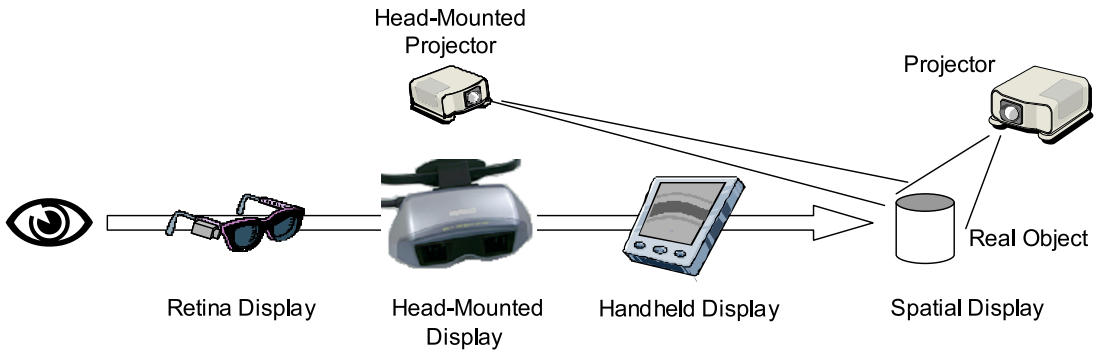


Figure 5.2: Different types of displays in MR environment.

virtual objects. Both information from the taken image of the real scene and a head tracker can be used for obtaining the user's viewpoint location. Optical see-through HMDs provide overlay through a half-silvered mirror or a transparent display. When users mount this type of display on their heads, imagery is provided in front of their eyes. The user-dependent calibration and precise head tracking are required for a correct graphical overlay. The head-attached displays achieve the visualization for just one user.

On the other hand, some MR applications use handheld displays. The flat-panel display presents virtual objects overlaid onto the image of the real world captured with an attached camera in the same way as a video-see through HMD. Tablet PCs, personal digital assistants (PDAs) [35, 37, 36, 82, 121] or cell phones [75] are typically used as displays. All of these examples combine processor, memory, display, and interaction technology in one single device. This enables augmentation of outdoor scenes.

Another approach is that augmentation is achieved by projecting virtual information directly on physical objects. One of the examples is use of head-mounted projectors [46, 45], whose images are projected along the viewer's line of sight at objects in the real world. The target objects are coated with a retroreflective material that reflects light back along the angle of incidence. Multiple users can see different images on the same target projected by their own head-mounted projectors.

In contrast to body-attached displays (head-attached or handheld), spatial displays detach the instruments from the user and integrate it into the environment. Three

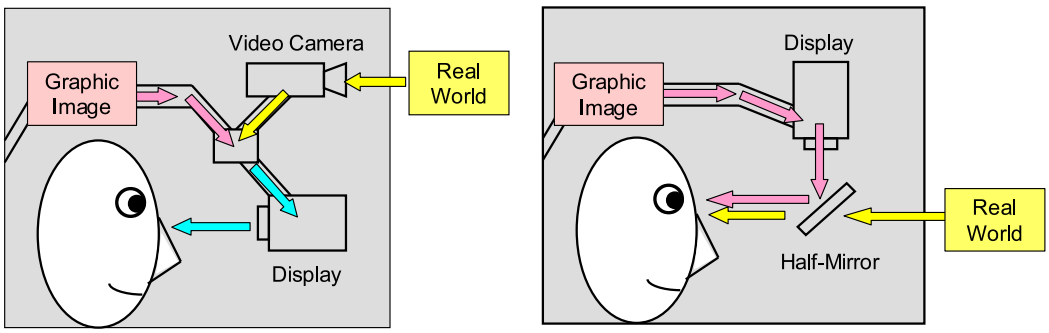


Figure 5.3: Video see-through HMD and optical see-through HMD.

different approaches exist which mainly differ in the way they augment the environment: either using video see-through, optical see-through or direct augmentation by projection.

Spatial video see-through displays make use of superimposition of virtual objects onto the video and display the merged images on a regular monitor. Spatial optical see-through displays generate images that are aligned within the physical environment. Spatial optical combiners, such as planar or curved mirror beam splitters [10], transparent screens [78], or optical holograms [11] are essential components of such displays. Recently the use of projector-based spatial displays is getting increased. This approach applies front-projection to seamlessly project images directly on physical objects' surfaces instead of displaying them on an image plane (or surface) somewhere within the viewer's visual field. Single projector [117] and multiple projectors [93] are applied to increase the potential display area.

The several types of instruments are used in MR systems. The choice of displays depends on the application requirements. We evaluate the appropriateness of each type of display for presenting a dynamic event such as a soccer match in mixed reality environment.

Current retinal displays can present just monochrome images. They are not useful for presentation of sporting events where full color images should be displayed. Projection displays are also inappropriate for dynamic scenes where the shape of the object changes over time because physical objects to be augmented are required for visualization. The possibility remains in head-mounted displays, handheld displays or spatial see-through displays. The optical see-through HMD requires user-dependent calibration and precise head tracking. The spatial optical see-through display requires a large system configuration. Therefore video see-through types of displays are considered to be most suitable for the dynamic events. These types have advantage that images of the real world captured by a camera attached to the display can be used for registration. The proposed systems in this thesis utilize a video see-through HMD or a handheld display. Spatial video see-through displays are not considered to be appropriate for the case that user controls the camera. It is difficult to control the camera while staring at the monitor.

We describe two kinds of system configuration to be used for presenting a soccer match here. One system utilizes a video see-through HMD, and the other system

consists of a web camera and a handheld display. Figure 5.4 and Figure 5.5 illustrate each system configuration. In the first system, a Canon video see-through HMD "VH-2002", which has been developed by Mixed Reality Systems Laboratory Inc. (MR Lab) is used. The HMD offers integrated stereoscopic camera whose optical axes coincide with the display axes, stereoscopic display, simultaneous capture, other features, including;

- wide field of view: 51 degree in horizontal direction, 37 degree in vertical direction
- high resolution: VGA (640×480)
- light weight: 286[g]
- standard camera I/O: NTSC

A viewer sees a desktop stadium through the HMD while virtual view images of soccer scenes are overlaid. The camera attached to the display captures the real environment, and the image is used for determining the viewpoint position. Once the virtual view images of players and ball are synthesized according to the viewpoint, they are superimposed on the captured image and then presented via the HMD. The

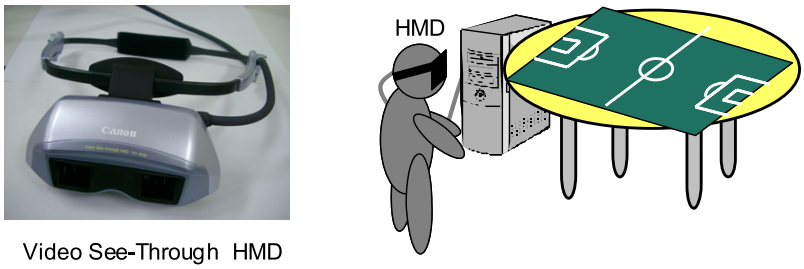


Figure 5.4: System configuration using an HMD.

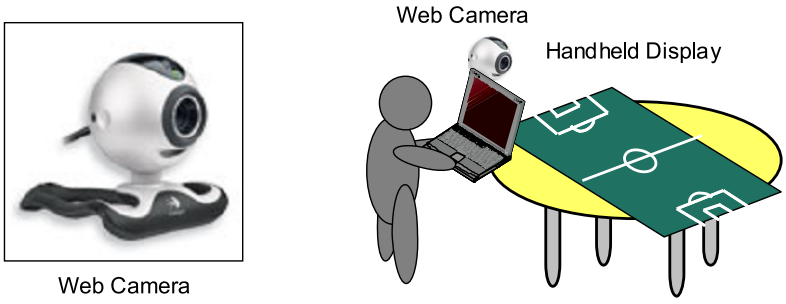


Figure 5.5: System configuration using a web camera and a handheld display.

viewer changes the viewpoint by moving his/her head.

The second system consists of a web camera produced by Logicool: "QuickCam Pro 4000" which has 640×480 video resolution and a handheld display. The web camera captures the real environment including the stadium model, and the image is used for viewpoint determination. After the virtual view images of players and ball are synthesized according to the viewpoint, they are superimposed on the captured image and then presented through the handheld display. The viewpoint position is controlled by moving the web camera.

The advantages and disadvantages of both systems are discussed. An HMD has intuitive interface because the camera is located close to the user's eyes. The user can change the viewpoint seeing the screen in front of his/her eyes. This allows the user to offer immersive impression into the merged virtual and real world. An HMD, however, is heavy and so inappropriate for long-term use. It may make the user uncomfortable. Another disadvantage is the cost. It depends on the ability, but it is typically more expensive than the system that consists of standard camera and display. By contrast, the second system with a web camera is easy to build. This system can be used not only in indoor scene but also in outdoor scenes because processor, memory, display, and interaction technology are combined in one single device. A handheld display, however, cannot produce immersive impression as well as an HMD. A user may have difficulties in controlling the viewpoint because the location of the camera is distant from the user's eyes.

5.1.3 Registration Techniques

In order to generate natural views of virtual objects superimposed in a real scene, there are several issues to be addressed such as geometry, illumination, and time consistency. The virtual objects have to be overlaid at the desired position. Shading of the virtual object has to match to that of the other objects in the real scene. Motions of the virtual objects and the real objects have to be coordinated. The detail explanations about the consistency of geometry and illumination are described below.

1) Geometric Registration

As regards geometric registration, many kinds of methods have been studied, such as the methods using positioning sensors [5], vision-based methods using images captured by cameras [60, 103, 107, 32], and combining methods using both of them [6, 108, 76]. In typical MR applications, virtual objects, whose 3D shape and positions are known, are inserted into the 3D space of the real world. The camera calibration and the acquisition of Euclidean 3D measurements of the environment are generally required. However, in the proposed method, real images of a sporting scene are overlaid onto a stadium model. The superimposed virtual objects, which are virtual viewpoint images synthesized from uncalibrated multiple cameras, have no 3D positions. This indicates that the conventional method is not useful for the geometric registration for virtual soccer match presentation.

Figure 5.6 illustrates comparison of the registration technique in the proposed method with that in the conventional method. The top figure describes the conventional registration method while the bottom one shows the proposed method. In the top figure, the appropriate view of the virtual object is overlaid on the image of the real world using projection matrix of the camera after the alignment of the virtual object and the real world coordinate systems. The registration is performed based on the 3D relationship. By contrast, in the bottom figure, such 3D-based registration cannot be utilized because the virtual objects to be overlaid have no 3D information. The registration between the real world and the virtual objects should be achieved using only 2D image information.

Another issue is generation of the appropriate view of the overlaid virtual objects. The conventional method can easily synthesize arbitrary viewpoint image since the virtual objects are usually generated by computer graphics. On the other hand, the virtual view synthesis is not easy in the proposed system because the soccer scene is taken by uncalibrated camera in a real stadium. Even if the 3D position and pose of the camera can be obtained, they cannot be directly used for the view synthesis. These two technical issues are solved as follows.

- virtual view synthesis
  utilization of the proposed method described in Chapter 3 and Chapter 4
- geometric registration
  proposal of new image-based registration technique using planar structure

To generate appropriate view of the virtual objects which correspond to players and ball, we utilize the proposed method in Chapter 3 and Chapter 4. This means the view synthesis problem comes down to viewpoint selection problem. As explained in the previous chapter, virtual viewpoint position is determined in the proposed method by three parameters: reference cameras, interpolating weight, and zoom ratio. The key point is a technique how to obtain these viewpoint parameters for mixed reality

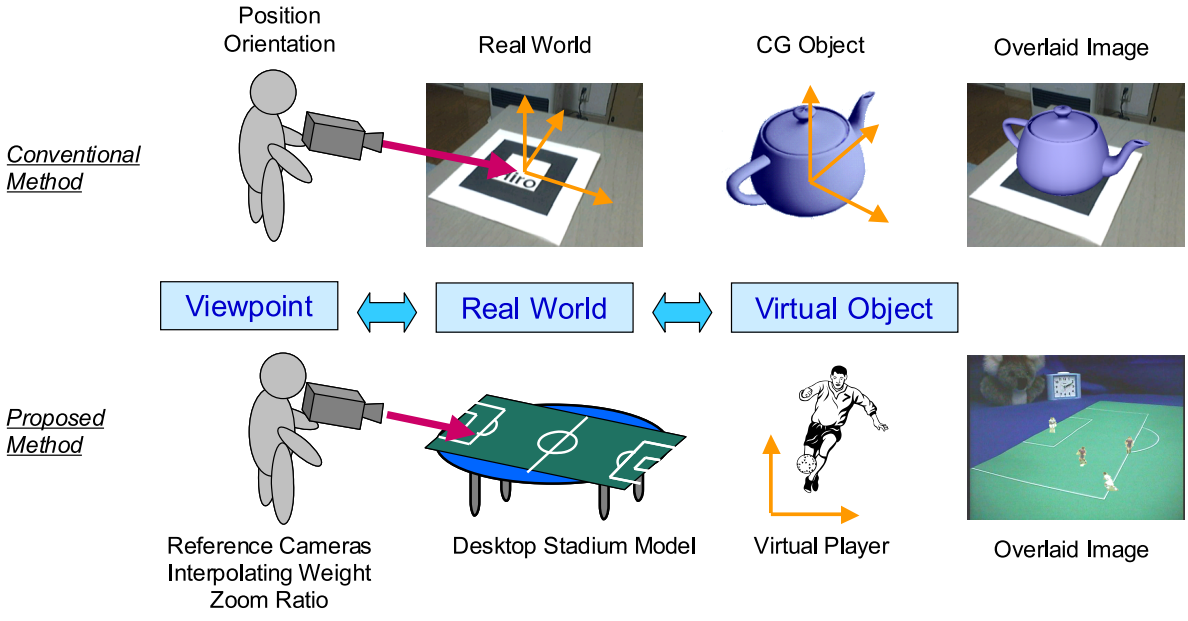


Figure 5.6: Comparison of the registration techniques.

presentation of a soccer match.

The new image-based registration technique exploits a characteristic in structure of sporting scenes. Players in many sports scenes move around on a planar area such as tennis court, soccer ground and baseball field. The projective geometry of the plane region can be useful under the limited condition that 3D information is not available. It is easily obtained using just images through corresponding points. We proposed the new registration method as one of the vision-based methods.

2) Photometric Registration

As regards photometric registration, most work has focused on the measurement of the real lighting environment and material properties. Sato et al. [97, 98] proposed a method to acquire a radiance distribution from image or an illumination distribution from shadow. Kanbara and Yokoya [54] proposed a method to acquire both illumination distribution and camera position by using a special marker that has a spherical mirror inside it. On the other hand, another research shows that it is no necessary to match lighting condition of the real environment in drawing the shadows of the virtual object. Sugano et al. [111] proposed the method for shadow representation of the virtual object in order to provide a stronger connection between the real world and virtual objects and to increase virtual object presence. Loscos et al. [65] presented a system for interactively remodeling and relighting real scenes.

The objective of the proposed system is replay of real sporting events. Hence, the shading of the virtual objects should be match to the original environment of the stadium rather than the real environment where a viewer is present. Shadows of players and ball are projected according to the original lighting condition in a stadium. If the captured scene has shadow regions, the shadow is overlaid on the desktop stadium using IBR technique; otherwise, it is not overlaid. The appearance of shadow on the desktop stadium enables more realistic visualization of the scene in MR environment.

5.1.4 Vision-Based Tracking

Tracking user's viewing orientation and position is crucial for MR systems. In vision-based methods, the camera tracking [94] is key issue for achieving the accurate geometric registration. This section describes several tracking methods for constructing MR systems.

The easiest way to do it is modifying the environment with fiducial markers placed in the environment at known locations. The position and orientation of a camera are estimated according to the appearance of the fiducial markers. A software library termed "ARToolkit" [55] which utilizes marker-based tracking allows building MR systems easily. The ARToolKit video tracking libraries calculate the camera position and orientation relative to markers in real-time. Marker based methods have greater robustness and lesser computational requirements. The placement of markers, however, is not always possible.

Tracking in unprepared environments relies heavily on tracking visible natural features. The markerless tracking is a difficult computer vision problem. Offline camera tracking from an image sequence [33, 87] has advanced to the point where commercial solutions have become available. 2D3 [116] offers industry-standard camera and object tracking solution. The application "boujou" has achieved automatic tracking in footage such as films, TV series, commercials, industrial and architectural visualizations, scientific and forensic reconstructions, and simulations. The 3D structure of the scene is given by automatically finding hundreds of features in every frame. This allows integration of CG with many types of footage. These algorithms achieve high accuracy even without a priori knowledge. They take advantage of time-consuming but effective batch techniques such as global bundle adjustment. It is best for special effect and postproduction because it does not run in real-time currently.

Whereas the markerless tracking from just natural features in video sequence is not suitable for online system, some approaches overcome this problem by using additional information. The common approach for dealing with unstructured environments is to impose some structure whose 3D positions are known. For each video frame, camera pose minimizing the reprojection error is estimated using a set of 3D points and corresponding 2D points. Wuest et al. [126] and Bleser et al. [12] have proposed the

methods utilizing a CAD model of the target object to be tracked. The camera pose is estimated in real-time by tracking edges or points of the object. These systems are robust to strong changes of the lighting conditions and partial occlusions. Vacchetti et al. [118] combine offline information in key frames with online information deduced from a traditional frame-to-frame tracking approach to ensure robustness. To reduce the number of unknowns, they use a 3D model of the scene and the knowledge that all image points must lie on the surface of the 3D model.

Another approach for markeless tracking is to learn the 3D structure of target scenes instead of imposing 3D objects in the environment. Genc et al. [38] and Subbaraoyz at al. [110] utilize a learning-based method for camera tracking. The system first learns the scene structure by employing a marker-based tracker. Once the implicit 3D model of the scene is constructed, the system computes the pose of the camera in real-time by tracking the learned feature in the scene where markers do not exist.

There is the opposite approach of placing fiducial markers in the environment. The marker is attached to the user for tracking, and a bird's-eye view camera observes the user from a fixed third-person viewpoint. This approach has advantage such that both user's view camera (inside-out tracker), and bird's-eye view camera (outside-in tracker) can be used for tracking. Klein and Drummond [57] demonstrate an AR system on a tablet PC without the use of markers placed in a scene. They achieved robust and accurate registration by combing edge-based tracking in a camera attached to the tablet PC and LED tracking in a outside camera that captures the tablet PC with LEDs. Satoh et al. [99, 100] proposed similar methods where markers are placed to both user and scene.

Although these methods enable to build an online MR system in unprepared environment, real-time registration tends to be less reliable since it cannot rely on batch computations such as bundle adjustment. There are many ongoing researches for minimizing registration error and latency, reducing calibration requirements.

This thesis introduces two approaches to tracking in order to build MR presentation systems for dynamic events. In the first approach, geometric registration is performed with tracking of natural features. Considering that the proposed system is applied not only a desktop stadium but also a toy of soccer stadium or a real empty stadium, it is desirable to achieve MR presentation in unprepared environment. Just feature lines, such as goal line and penalty area lines on the soccer ground are used

for tracking. Camera calibration is not required. As explained above, it is difficult to tracking scene features accurately in real-time. This feature-based MR presentation system emphasizes accuracy of tracking rather than processing time in order to confirm the effectiveness of the novel concept such that dynamic event is replayed in MR environment. The system determines the appearance and positions of virtual players to be overlaid according to the camera pose obtained by feature tracker. This feature-based system is described in detail in Section 5.2.

The second approach demonstrates an online MR system where marker-based tracking is used. Interactive visualization is important for soccer match observation. A fiducial marker is placed on the stadium model in the real environment for less computational requirements. The position and orientation of the camera are obtained by detecting the marker. This marker-based MR presentation system emphasizes processing time rather than accuracy of tracking. A viewer can watch virtual soccer match online in the real environment while changing the viewpoint. The details of the online marker-based system is explained in Section 5.3.

5.2 Feature-Based MR Presentation System

5.2.1 System Overview

Figure 5.7 shows the flow process of the MR presentation system based on tracking natural features. A camera captures a desktop stadium model placed in the real environment. This camera is termed MR camera to be distinguished from stadium cameras that are used to take a soccer match in a stadium. Virtual view images of players, ball, and shadows in soccer scenes are overlaid onto the MR camera image. A viewer can observe moving soccer players on the desktop stadium model by watching mixed image of the real environment and virtual view images.

The multiple soccer videos captured in real stadiums as explained in Section 3.6 are used for MR presentation. Virtual view images of the players, ball, and their shadows are generated from the stadium camera images and overlaid on the desktop stadium model according to the MR camera pose. The viewpoint position can be changed by moving the MR camera. In this system, MR camera does not require strong calibration. All processes including view synthesis and registration are performed using just captured images.

The MR presentation process consists of three stages: (1) calculation of the viewpoint position, (2) virtual view synthesis of the dynamic regions in a soccer scene, and (3) overlay of the synthesized images on the stadium model. At the first stage,

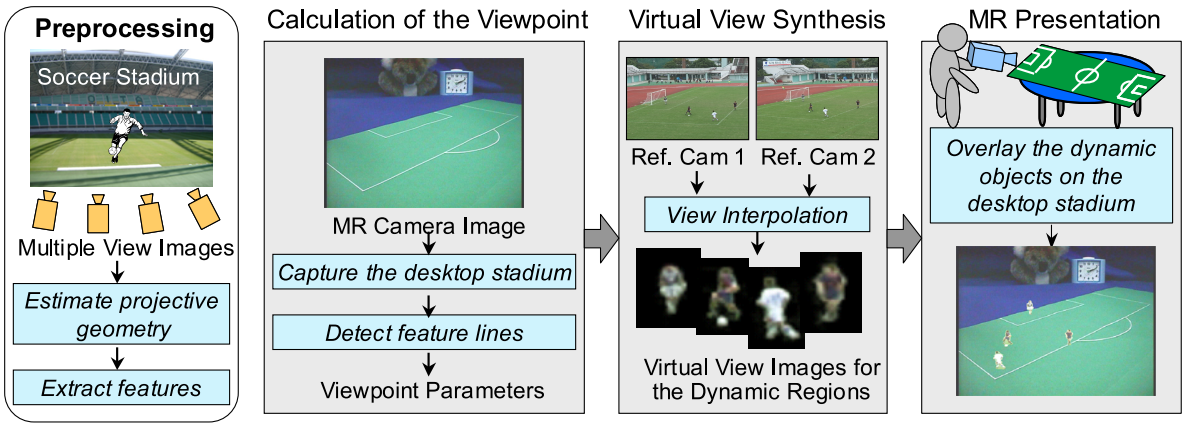


Figure 5.7: Process flow of the feature-based MR presentation system.

the viewpoint position that corresponds to the MR camera position is determined by tracking feature lines in MR camera images. At the second stage, the virtual view images of the players, ball, and shadows are synthesized from neighboring cameras near the viewpoint position, which are reference cameras, using view interpolation technique. At the final stage, the synthesized images of the dynamic objects are overlaid on the desktop stadium model. The detail explanation of each process starts in the next section.

5.2.2 Calculation of Viewpoint

Our view synthesis algorithm is based on view interpolation between neighboring two or three cameras near the virtual viewpoint. The explanation for the case of two cameras is described here. The viewpoint position is specified by three parameters, which are (a) neighboring two reference cameras, (b) interpolating weight value between the reference cameras, and (c) zoom ratio between the real and the virtual cameras. In order to generate a soccer scene from the viewpoint of the MR camera using three parameters, we have the following assumptions. The viewpoint movement in the direction of the sideline can be controlled by selecting reference cameras and the interpolating weight. The movement in the direction of the goal line can be controlled by changing the zoom ratio between the reference cameras and the MR camera. The viewpoint position is approximated with these assumptions to synthesize virtual view images from uncalibrated camera images.

1) Detection of the Feature Lines

To calculating the viewpoint parameters, natural feature lines are tracked in MR camera image sequences. The stadium model contains feature lines, which are easy to track, such as lines of the penalty area or the goal area. These feature lines are used instead of using any fiducial markers. The efforts for locating markers can be reduced. The Canny operator [14] is firstly applied for edge detection, and all the edge points are mapped into a Hough space. The strong peaks that form the lines of the penalty area and the goal area are then found in the Hough space. The examples of result of the line detection are shown with the edge images in Figure 5.8. All parameters for

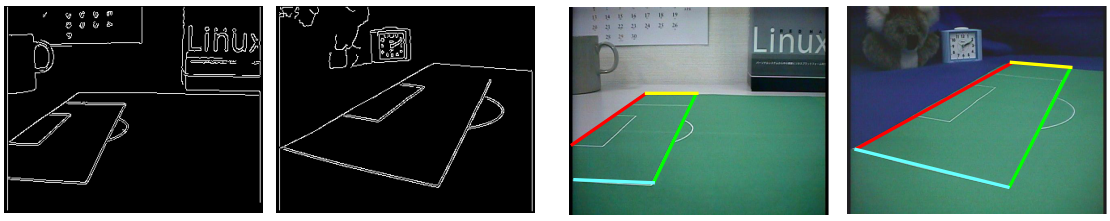


Figure 5.8: Examples of edge images (left) and detected feature lines (right).

specifying the virtual viewpoint position are determined based on these feature lines. In examples of Figure 5.8, four lines are tracked and used for the determination of the viewpoint.

2) Determination of the Reference Cameras and the Interpolating Weight

A vanishing point is utilized for selection of the reference cameras and determination of the interpolating weight between the cameras. A vanishing point is the point to which the extensions of parallel lines appear to converge in a perspective projection. The position of the vanishing point in image decides the orientation of the camera to the parallel lines. In the proposed system, the orientations of the MR and stadium cameras to the direction of the goal lines are estimated using the vanishing point. Two cameras whose orientations are the closest to those of the MR camera are selected as the reference cameras.

At preprocessing stage, the locations of the vanishing points of all stadium cameras are obtained by extending lines of the goal area and the penalty area. During MR presentation, feature lines are detected in every MR camera image, and the location of the vanishing point is obtained in the same way (see Figure 5.9). A horizontal component of the location of the vanishing point is compared for calculating the viewpoint position because we assume that the viewpoint is moved almost horizontally from side to side, and that all stadium cameras are placed at the almost same height. According to such assumptions, two cameras in which the locations of the vanishing points are closest to the vanishing point in the MR camera are selected as reference

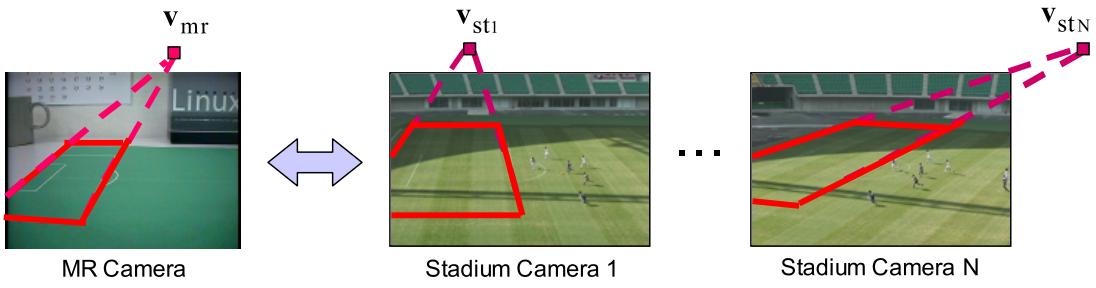


Figure 5.9: Comparison of the location of vanishing points between in the MR camera image and in the stadium camera images for feature-based presentation.

cameras. Then the relative distance between the vanishing point of the reference camera and that of the MR camera determines the interpolating weight α by the following equation:

$$\alpha = \frac{x_{mr} - x_{stL}}{x_{stR} - x_{stL}} \ , \tag{5.1}$$

where $\boldsymbol{v}_{stL} = (x_{stL}, y_{stL})^\top$ and $\boldsymbol{v}_{stR} = (x_{stR}, y_{stR})^\top$ represent the vanishing points in left and right reference camera images, respectively, and also $\boldsymbol{v}_{mr} = (x_{mr}, y_{mr})^\top$ represents the vanishing point in the MR camera image.

3) Determination of the Zoom Ratio

Just selection of the reference cameras and determination of the interpolating weight may not generate the virtual view image from the same viewpoint of the MR camera. The zoom ratio between the MR camera and the reference cameras is controlled for precise registration. If extrinsic parameters of the MR camera and the reference cameras are calculated, the 3D relationships between these cameras are obtained. However, all parameters of the cameras are unknown since the proposed system uses uncalibrated cameras. The virtual view images of the soccer scene are adjusted by expansion or contraction of the images using the zoom ratios to be overlaid on the stadium model. The focal lengths of the reference stadium camera f_{st} and the MR camera f_{mr} determine the zoom ratio as

$$z = \frac{f_{mr}}{f_{st}} \ . \tag{5.2}$$

The focal lengths of the MR camera and stadium cameras are actually fixed, but the zoom ratio can be calculated by changing the focal length of the MR camera virtually when we consider zooming as the change of the focal length. As the intrinsic parameters of the cameras are unknown, the focal length is computed with two vanishing points $\boldsymbol{v}_1 = (x_{v1}, y_{v1})^\top$ and $\boldsymbol{v}_2 = (x_{v2}, y_{v2})^\top$ by the following equation:

$$x_{v1}x_{v2} + y_{v1}y_{v2} + f^2 = 0 \ . \tag{5.3}$$

It is supposed that the skew of the camera is 0, aspect ratio is 1, and principal point is the center of the image. The viewpoint parameters obtained by the above method are used for virtual view synthesis for the dynamic objects.

5.2.3 Overlay of Dynamic Objects

The proposed system performs homography transform for the registration between the dynamic object and the stadium model. In order to generate natural views of the soccer match, the players, ball, and shadows need to be rendered correctly onto the desktop stadium in a MR camera. Their positions in the camera image are determined by the homography that represents a transformation between the ground plane in the original soccer scene and that of the stadium model in the MR camera. This homography is computed from more than 4 corner points of the goal area or the penalty area. The intersection points of the detected feature lines are used as the corner points of each area.

The position of player can be determined by transforming the foot position of the original soccer scene by the homography. However, the players do not always contact the ground, and detection of the foot position in soccer scenes is neither stable nor accurate; hence, it may give the appearance that players vibrate during the replay. In order to present the motion of the players stably, the centroid of the player region is utilized instead of the foot position. The distance of the centroid from the ground plane is considered for accurate registration. In addition, the position of the ball is

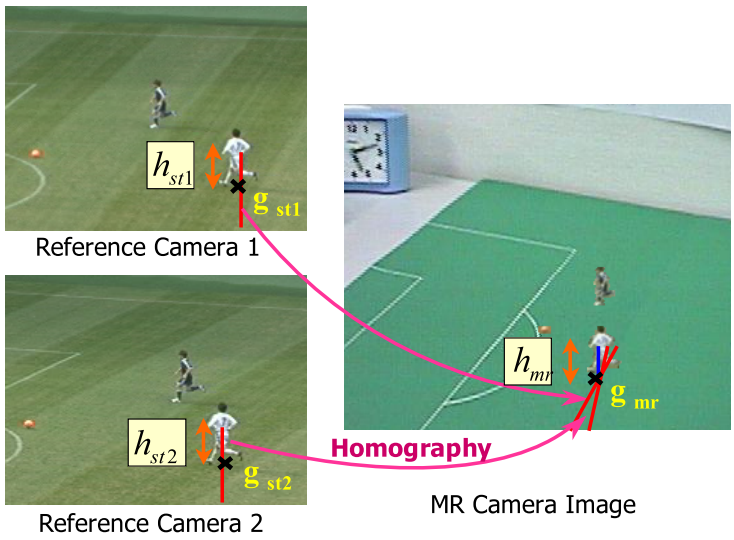


Figure 5.10: Determination of the player and ball positions in the MR camera image.

transformed by applying the same technique using the centroid because the ball in the air cannot be transformed correctly by the use of simple homography transform. The centroidal lines of each player and ball (described as the red lines for the player in Figure 5.10) are projected onto the MR camera image from two reference camera image using the following equation:

$$\tilde{\boldsymbol{p}}_{mr} \cong \boldsymbol{H}\tilde{\boldsymbol{p}}_{st} \ , \tag{5.4}$$

where $\boldsymbol{H}$ is the homography that represents the transformation between the ground planes, and $\tilde{\boldsymbol{p}}_{st}$ and $\tilde{\boldsymbol{p}}_{mr}$ are homogeneous coordinates of the position in the reference stadium camera image and in the MR camera image, respectively. The intersection point $\boldsymbol{g}_{mr}$ of the projected lines is the position of player/ball on the stadium model in the image. Backprojection of the intersection point to reference cameras gives their positions $\boldsymbol{g}_{st1}$ and $\boldsymbol{g}_{st2}$ on the ground plane in the reference images. The distances are then calculated between the centroid and the ground plane, that is h_{st1} and h_{st2}. The following equation gives the distance h_{mr} between the centroid and the ground plane in the MR camera:

$$h_{mr} = (1-\alpha)h_{st1}z_1 + \alpha h_{st2}z_2 \ , \tag{5.5}$$

where z_1 and z_2 are the zoom ratios, and α is the interpolating weight. Even when the players are jumping off the ground or the ball fly in the air, the positions of players and ball can be calculated stably.

If the captured scene in the stadium has shadow regions, the shadows of players and ball need to be overlaid on the stadium model. As shadows are projected on the ground plane, simple homography transform between the ground plane of the soccer stadium and the desktop stadium determines the shadow positions on the desktop stadium. We regard the color value of the shadow regions as to become half as much as that of the original stadium model. After the shadows are overlaid on the stadium model, the players and ball are overlaid additionally. Figure 5.11 shows an example of the represented shadows with the players and ball on the stadium model from two reference camera images.

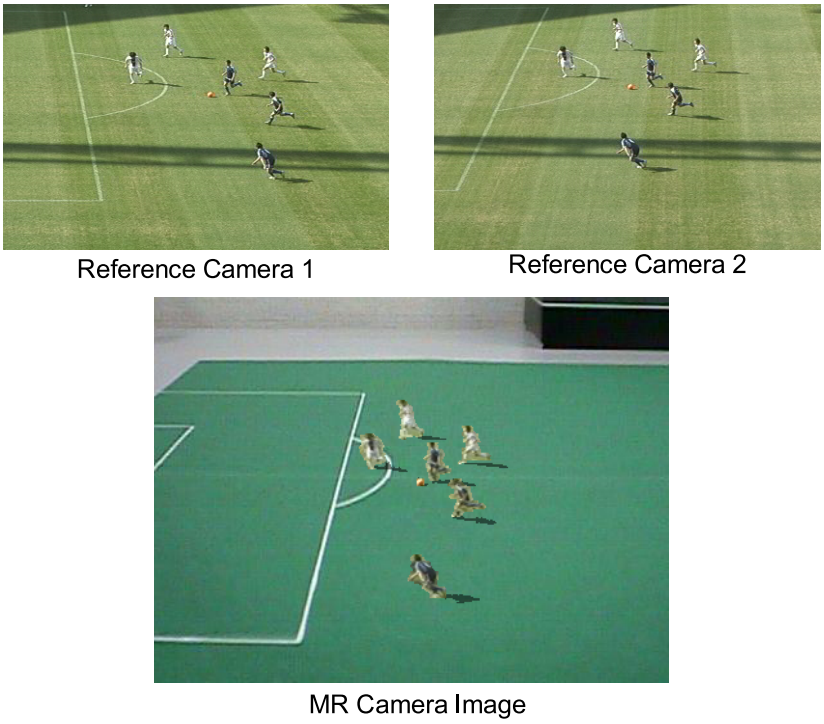


Figure 5.11: Representation of the shadows on the stadium model.

5.2.4 Experimental Results

We have implemented a feature-based MR presentation system for an actual soccer match. Figure 5.12 describes the experimental scene where a viewer sees a desktop stadium model on a table through an HMD. In this experiment, an HMD was used for capturing the real world and presenting virtual soccer scene overlaid on the stadium model. The use of HMD has the advantage that it allows the viewer to feel strong immersive impression. The system configuration using an HMD is illustrated in Figure 5.4. Note that other system configuration such as a system using a handheld display with a web camera attached to can be also acceptable. The proposed method was applied to the multiple soccer videos captured in the Oita Stadium in Oita city and Edogawa Athletics Stadium in Tokyo, Japan. These videos are the same as explained in Section 3.6.

The following preprocessing was performed first;

- Estimation of projective geometry between stadium cameras:
  fundamental matrices and homographies
- Calculation of the vanishing points and corner points in each stadium camera image
- Extraction of the dynamic regions
- Segmentation of the dynamic regions into player/ball and shadow regions
- Corresponding players in neighboring views

The fundamental matrices between the viewpoints of the cameras and the homogra-

Figure 5.12: Watching a soccer match in the viewer's environment using an HMD.

phies between the planes that form the ground in the neighboring views were computed using the corresponding points. We manually selected about 50 corresponding points, whose 3D positions varied in the object space, for fundamental matrices and 20 points on each plane for homographies in the image sequence. Then the positions of the vanishing points and corner points of penalty and goal area were calculated in each stadium camera image using the lines of the goal or the penalty area. In addition, the dynamic regions were extracted in each frame of the image sequences. In the scene including shadows, the extracted regions were segmented into the player/ball and the shadow regions. After every region was segmented and labeled, the regions of the same player in the neighboring views were corresponded by using the homography of the ground plane between the views. The above processes were implemented as preprocessing of the MR presentation, and the dataset was stored in a PC.

During the replay, the texture and 2D positions of all player and ball regions in two reference camera images, the correspondence map, and the positions of the shadow regions in reference images are loaded to the PC according to the viewer's viewpoint in each frame. MR presentation is performed in the following order;

1. Capturing the stadium model with the MR camera
2. Extraction of feature lines in the camera image
3. Calculation of the viewpoint parameters
4. Virtual view synthesis for the players, ball and shadows
5. Estimation of the homographies of the ground plane between the reference cameras and the MR camera
6. Calculating the rendering positions of the players, ball and shadows
7. Overlaying them on the stadium model

After observation starts, the lines indicating the goal area and penalty area of the stadium model are detected in the MR camera image that is the image captured with the HMD in every frame. Then virtual view images of the dynamic regions are synthesized according to the viewpoint position determined by the feature lines. Next, homographies of the ground plane between each of reference camera images and the MR camera image are calculated. The rendered positions of the dynamic objects are determined with the homographies. Finally the synthesized soccer scene is overlaid onto the desktop stadium model through the HMD. This process is iterated until the

replay of the soccer match ends.

Figure 5.13 presents the captured scenes in the stadiums and results of the overlaid soccer scenes on the desktop stadium model. The first and the second columns depict the reference camera images used for the virtual view synthesis, and the third column depicts the synthesized virtual view images with the original soccer scene. The fourth column shows the displayed soccer scenes on the HMD. The interpolating weight and the zoom ratio are indicated as w and z, respectively, at the bottom of each image. For example, the image on the top of the last column was generated based on the parameters that interpolating ratio is 0.31 between camera 1 and camera 2, and zoom ratio is 1.05. We see that the virtual players and ball can be inserted naturally in the real world. When comparing the overlaid scene with the original soccer scene, the players are located at almost correct positions on the stadium model. The appearance of players and ball looks different in the virtual view images and the overlaid soccer scenes. This is because the locations of the players and the ball are modified using homography transformation of the ground plane in the HMD camera images. Thus the overlaid soccer scene is comfortably fitted to the stadium model.

Figure 5.14 and Figure 5.15 show the free viewpoint video images of the soccer scenes without shadows and scenes including shadows, respectively. From the top on the left to the bottom on the right, the results indicating the motion of the players and ball are replayed smoothly. Shadows are also overlaid so naturally that viewer does not feel any discomfort. However, errors of the rendering positions of the shadows occur when segmentation process for the dynamic regions fails to separate player regions and shadow regions. Sometimes a gap between a foot of the player and the shadow was observed.

Figure 5.16 describes the examples of the close-up views of the replayed soccer match. It seems that small athletes play soccer on the table. In the right image, the reality of the soccer player is increased due to the appearance of the shadow regions.

Figure 5.17 compares the accuracy of registration. One player positions on MR camera image in the case of using the centroid and using the foot positions are presented on the top figure. As the difference is found in horizontal movement, just x-coordinates are shown in the figure. Although the player position vibrating in the case of using the foot position, the player is moving smoothly in the case of using the centroid. In Figure 5.17 (b), the top image sequence shows the result where

the centroid is used for determining the rendering positions of the players and ball. The bottom image sequence shows the result where the foot position is used. The sudden change in the distance between players can be seen in the result using the foot position because the player's positions are not obtained stably. The effectiveness of the registration method using the centroid is confirmed in this comparison.

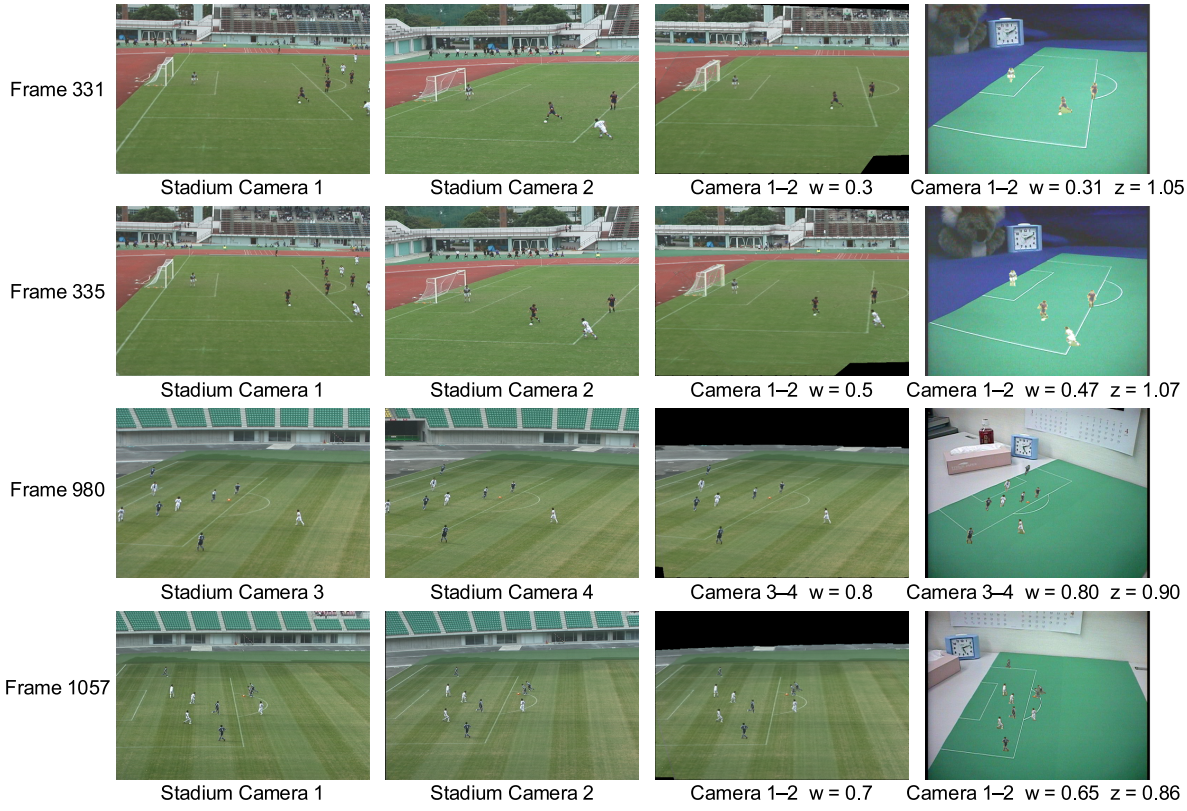


Figure 5.13: Result image of feature-based MR presentation and the original soccer scene.

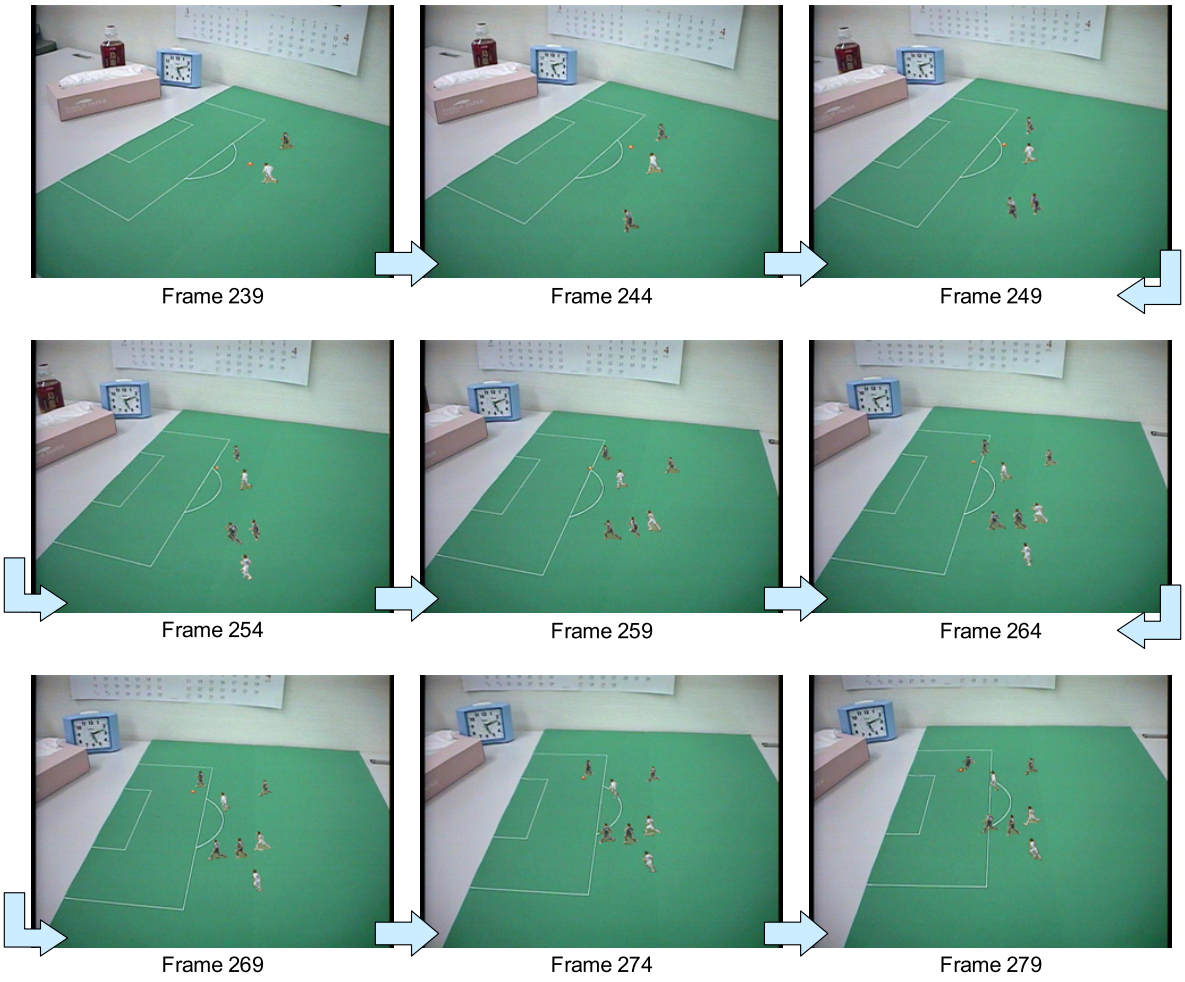


Figure 5.14: Result image sequence in the feature-based MR presentation system.

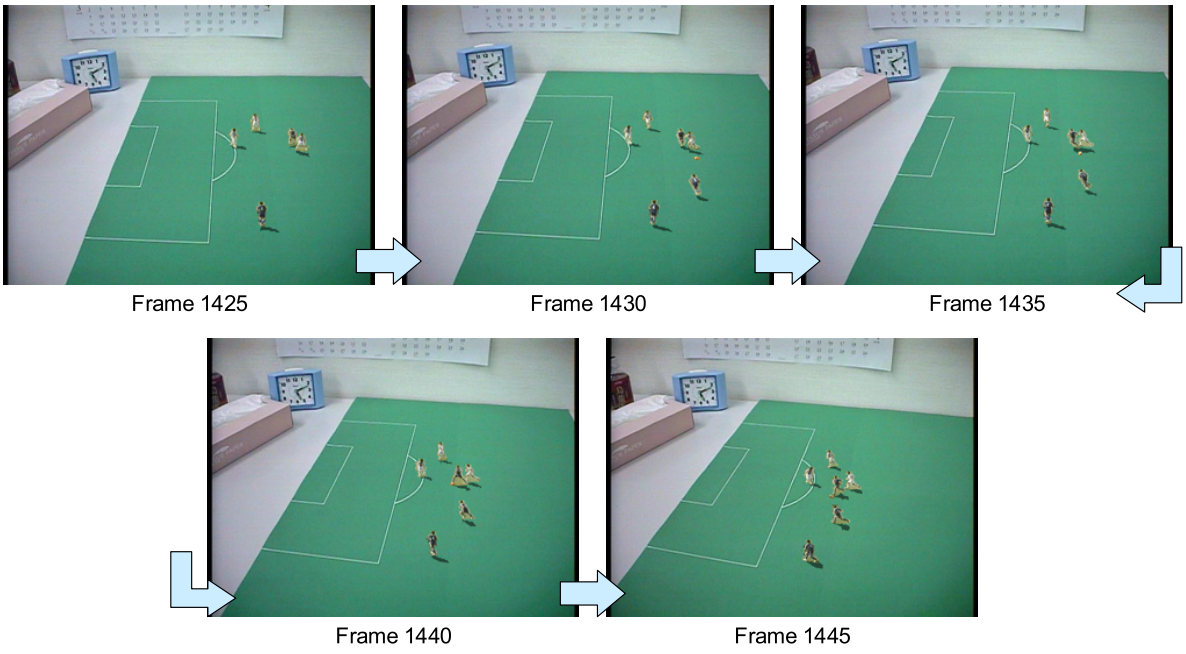
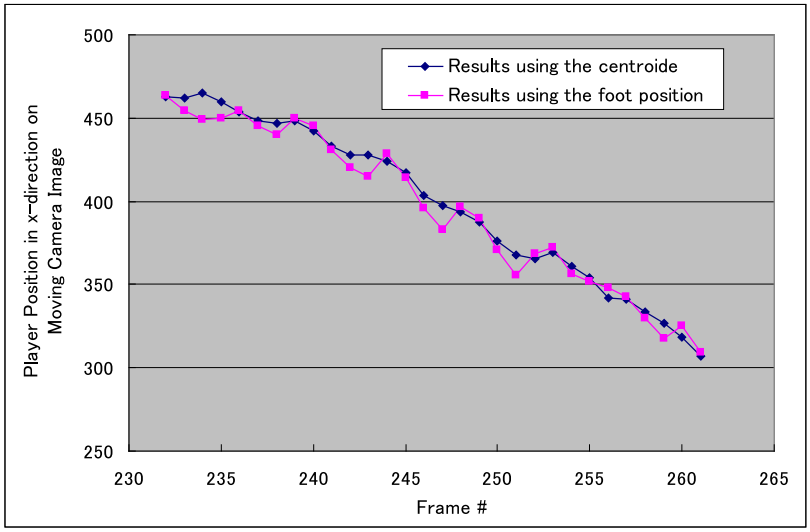


Figure 5.15: Result image sequence including shadows in the feature-based MR presentation system.

Figure 5.16: Close-up view of the feature-based MR presentation of a soccer match.



(a) The transition of the centroid of one player.

(a) Results using the Centroid

(b) Results using the Foot Position

(b) The overlaid players on the stadium model.

Figure 5.17: Comparison of the player positions between in the case of using the centroid and the foot position for the registration.

5.2.5 Discussion

The appearance consistency in MR presentation is discussed here. The virtual view images of players and ball are synthesized according to the viewpoint position which is determined approximately by two reference cameras, interpolating weight and zoom ratio. The viewpoint position represented by three parameters is not precisely identical to the position of the MR camera. Therefore the appearance of the dynamic objects presented to the viewer is not always geometrically accurate. However, this problem cannot be solved under the condition that virtual view image is synthesized by utilizing view interpolation technique from uncalibrated camera images. Minimizing the error is key issue. In the proposed system, the viewpoint position is selected appropriately by the use of vanishing point so that the appearance of overlaid dynamic objects can be natural. The effective visualization has been achieved in a sense that we focus on the global appearance of the dynamic events in the real world.

The current system does not run in real-time because real-time feature tracking is not accurate enough for registration. However, the feature-based system has expandability. It is unnecessary to place fiducial marker in the environment. If the tracking runs in real time, the virtual soccer match can be replayed not only desktop stadium but also other miniature stadium or real empty stadium.

5.3 Marker-Based MR Presentation System

5.3.1 System Overview

Figure 5.18 shows the overview of the second MR presentation system that is a marker-based MR presentation system. An MR camera takes images of the real world including a desktop stadium with a marker. As this system utilizes marker-based tracking for registration, a 2D square marker is attached to the stadium model. According to the camera position and orientation, virtual players, ball, and their shadows are overlaid on the stadium model. A viewer can observe moving soccer players on the desktop stadium model from his/her favorite viewpoint by controlling the MR camera.

The rough flow of the process is identical to that of feature-based system. The difference can be seen in the processes at the first stage and the last stage. At the first stage, instead of detecting feature lines of the stadium model, the marker is detected, and the position and orientation of the camera are obtained using a software library termed "ARToolKit" [55]. The detail explanation of this software is included in Section 5.3.3. At the last stage, the rendering positions for overlaying dynamic objects are not directly determined from the positions in reference camera images. First, the 3D positions of players and ball on the stadium model are obtained and then projected to the 2D image of the stadium model using the projection matrix of

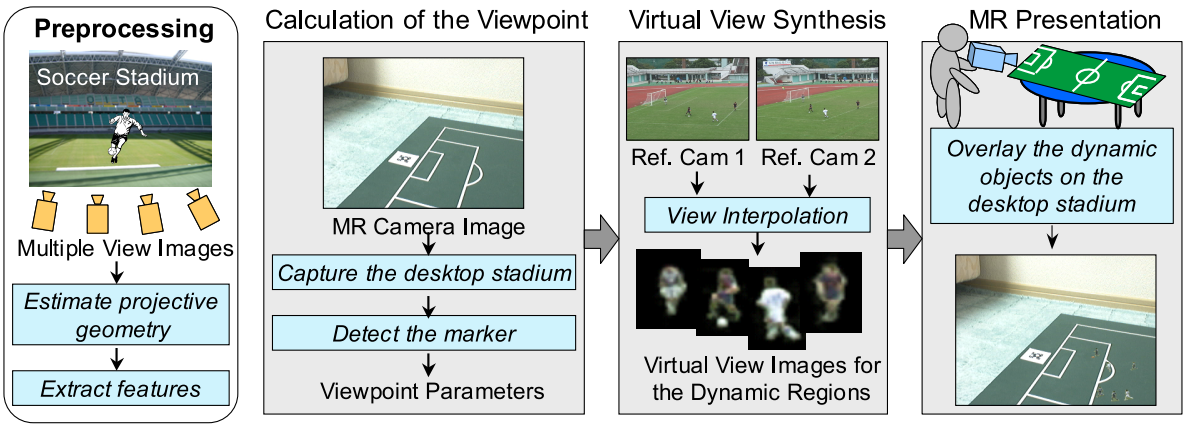


Figure 5.18: Process flow of the marker-based MR presentation system.

the camera. It is assumed that the intrinsic parameters of the MR camera are given preliminarily. Although the camera does not always require calibration process, the intrinsic parameters of the camera are estimated using a checker pattern for achieving precise registration in the proposed system. In the next section, the method how to obtain the intrinsic parameters of the camera is described.

5.3.2 Camera Calibration

When calculating rendering position of players, accurate registration can be performed if the intrinsic parameters of the MR camera are known. In the ARToolKit, default camera properties are contained in the camera parameter file that is read in each time an application is started. The parameters should be sufficient for a wide range of different cameras. However, using a relatively simple camera calibration technique, it is possible to estimate parameters for the specific cameras that are being used. If the camera parameters are obtained, the video image can be warped to remove camera distortions.

In the proposed system, the MR camera to be used is calibrated in preprocessing. The technique proposed by Zhang [132] is utilized to calibrate the camera easily. This technique only requires the camera to observe a planar pattern shown at a few (at least two) different orientations. Figure 5.19 shows the example images of checker pattern captured for the camera calibration. Either the camera or the planar pattern can be freely moved. The motion does not need to be known. The procedure consists of a closed-form solution, followed by a nonlinear refinement based on maximum likelihood criterion. Radial lens distortion is taken into account. The intrinsic parameters obtained by this method are used for determining the positions of players and ball to be overlaid on the stadium model.

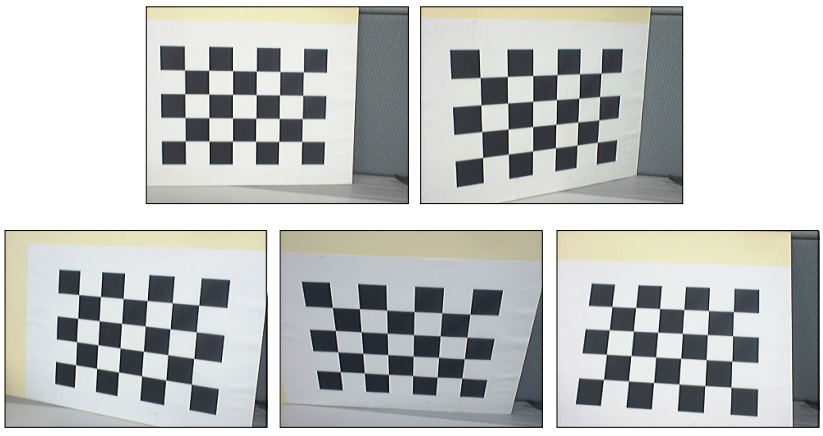


Figure 5.19: The checker pattern used for camera calibration.

5.3.3 Calculation of Viewpoint

As described in Section 5.2.2, the viewpoint position is specified by three parameters: reference cameras, interpolating weight, and zoom ratio. These parameters are determined by using the position and orientation of the MR camera in this system. The viewpoint of the camera is easily obtained using 2D square marker whose shape and positions are known. The "ARToolKit" allows us to calculate the camera position and orientation relative to physical markers in real time. This software enables the easy development of a wide range of AR applications. There are several steps in camera tracking. The algorithm is described briefly.

1) Detection of the marker and estimation of the camera position and orientation

Figure 5.20 illustrates the process flow of the marker detection using ARToolkit. The images included in this figure are cited in [55]. First the live video image is turned into a binary image based on a lighting threshold value. All the square regions are then searched in this binary image; this includes squares that are not the tracking markers. For each square, the contour is detected, and the corner is obtained in

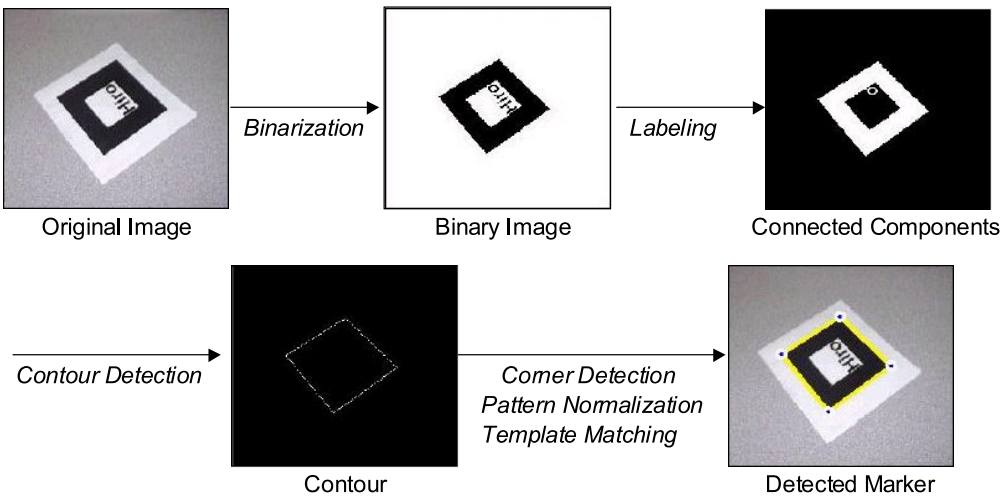


Figure 5.20: Process flow of the marker detection in ARToolkit.

sub-pixel order. The pattern inside each square is then captured and compared with some pre-trained pattern templates. If there is a match, the square is recognized as a tracking marker. Using the known square size and pattern orientation, the position of the camera relative to the physical marker is calculated.

In the proposed system, the world coordinate system is defined with a model plane at $Z = 0$ on the desktop stadium as shown in Figure 5.21. The X-axis and Y-axis are taken along with the direction of sideline and the direction of goal line, respectively. ARToolKit provides us the position and orientation of the MR camera relative to the marker. If the marker location is known, the relationship between the 3D space on the model plane and the 2D image in the MR camera can be obtained. A point in 3D space $\boldsymbol{X} = [X, Y, Z]^\top$, whose homogeneous coordinate is represented by $\tilde{\boldsymbol{X}} = [X_1, X_2, X_3, X_4]^\top$, is projected to a point on 2D image $\boldsymbol{x} = [x, y]^\top$ whose homogeneous coordinate is represented by $\tilde{\boldsymbol{x}} = [x_1, x_2, x_3]^\top$ using the following equation:

$$\tilde{\boldsymbol{x}} \cong \boldsymbol{P}\tilde{\boldsymbol{X}} \ , \tag{5.6}$$

where

$$\boldsymbol{P} = \boldsymbol{A}\ [\boldsymbol{R} \mid \boldsymbol{t}] \ , \tag{5.7}$$

$\boldsymbol{P}$ is the projection matrix, $\boldsymbol{A}$ is the matrix that represents intrinsic parameters, $\boldsymbol{R}$ is the rotation matrix, and $\boldsymbol{t}$ is the translation vector of the camera. We use $\boldsymbol{P}$ and $\boldsymbol{t}$ given by ARToolkit, and use $\boldsymbol{A}$ given by calibration process explained in the previous section.

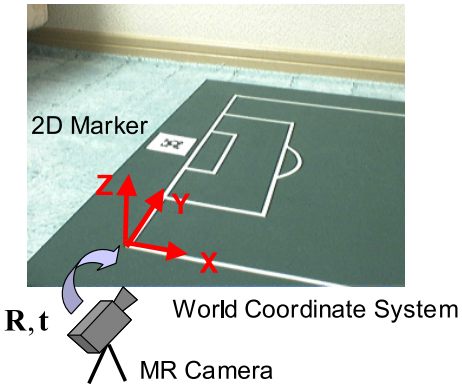


Figure 5.21: The world coordinate system defined in the proposed method.

2) Determination of the Reference Cameras and the Interpolating Weight

A vanishing point is applied for selection of the reference cameras and determination of the interpolating weight between the cameras in the same way as the feature-based system. The method how to obtain the vanishing point is different from the previous system. In each MR camera frame, 2D positions of feature lines such as lines of the goal area and the penalty area are calculated from the corresponding 3D positions on the stadium model using the projection matrix. Although the size of the soccer field depends on the stadium, it is assumed to be obtained in advance. The other parameters for the field such as the sizes of the penalty area and goal area can be known from the official standards. Figure 5.22 shows the examples of the projected lines using the projection matrix. The feature lines are correctly projected at almost same positions as the features lines on the desktop stadium model. This means that the position and orientation of the camera are calculated correctly. The vanishing point is obtained by extending the projected lines in the image, and then two reference camera and interpolating weight are determined in the same way as described in Figure 5.23.

3) Determination of the Zoom Ratio

The zoom ratio between the MR camera and the reference cameras is controlled for the precise registration. As the rotation matrix and the translation vector of the MR camera are known, the zoom ratio is determined by utilizing the camera position.

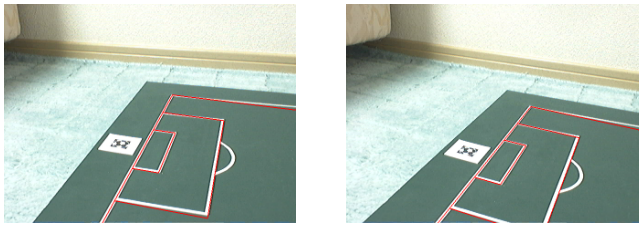


Figure 5.22: Examples of the projected feature lines using estimated camera position and orientation.

As the positions of the reference cameras need to be calculated, the rotation matrix and the translation vector for all stadium cameras are obtained in preprocessing. Because the stadium cameras are not calibrated, each camera position is estimated via homography between the ground plane of the stadium in 3D space and its image. The method for obtaining the rotation matrix $\boldsymbol{R} = (\boldsymbol{r}_1\ \boldsymbol{r}_2\ \boldsymbol{r}_3)$ and the translation vector $\boldsymbol{t}$ of the camera using homography $\boldsymbol{H}$ is explained here.

A point on the model plane in 3D space $\boldsymbol{X} = (X, Y, Z)^\top$ is projected to a point in 2D image $\boldsymbol{x}$ by Equation (5.6). This equation can be written using elements as follows:

$$\tilde{\boldsymbol{x}} \cong \boldsymbol{A}\ [\ \boldsymbol{r}_1\ \boldsymbol{r}_2\ \boldsymbol{r}_3 \mid \boldsymbol{t}\ ] \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} . \tag{5.8}$$

On the other hand, the point $\boldsymbol{X}$ is transformed by homography $\boldsymbol{H}$ which represents the transformation points on the model plane to the camera image:

$$\tilde{\boldsymbol{x}} \cong \boldsymbol{H} \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} . \tag{5.9}$$

Equation (5.8) and Equation (5.9) derive the following equation:

$$\boldsymbol{H} \cong \boldsymbol{A}\ [\ \boldsymbol{r}_1\ \boldsymbol{r}_2 \mid \boldsymbol{t}\ ] . \tag{5.10}$$

When the matrix $\boldsymbol{A}$ including intrinsic parameters is known, $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ and $\boldsymbol{t}$ are easily

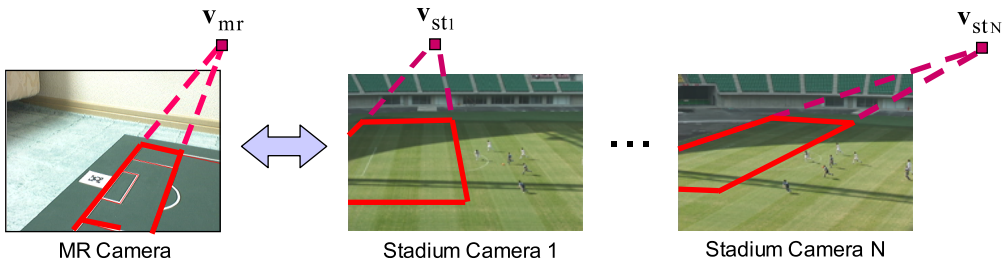


Figure 5.23: Comparison of the location of vanishing points between in the MR camera image and in the stadium camera images for marker-based presentation.

obtained from $\boldsymbol{A}^{-1}\,\boldsymbol{H}$. As $\boldsymbol{R}$ is a rotation matrix, its columns should be orthonormal. Therefore $\boldsymbol{r}_3$ is given by the cross product $\boldsymbol{r}_1 \times \boldsymbol{r}_2$. The above process is performed for all stadium cameras in preprocessing.

To calculate the zoom ratio for MR presentation, the position of MR camera that is obtained by using ARToolkit is compared with the position of the reference cameras obtained the above process. Finally all viewpoint parameters are determined so that virtual view images of players and ball can be synthesized according to the viewpoint.

5.3.4 Overlay of Dynamic Objects

The registration between the dynamic objects and the desktop stadium is performed by applying homography transform. Unlike feature-based system, the positions of players and ball in a MR camera image are not directly calculated. Their 3D positions on the desktop stadium are obtained first and then projected to the MR camera image using projection matrix obtained in each frame.

1) Projection to the Desktop Stadium Model

The method for calculating the 3D positions of players, ball, and shadows on the stadium model is explained. The homography transform is used for obtaining them.

In advance, the homography between the ground plane of the stadium in 3D space and its image in each stadium camera is obtained by corresponding feature points. In the homography transform, the centroid is used instead of the foot positions for the same reason as mentioned in Section 5.2.3. The centroidal lines of each player and ball (described as the red lines for the player in Figure 5.24) are projected onto the desktop stadium model from two reference camera image using the following equation:

$$\tilde{\boldsymbol{p}}_{model} \cong \boldsymbol{H}\tilde{\boldsymbol{p}}_{st} \ , \tag{5.11}$$

where $\boldsymbol{H}$ is the homography that represents the transformation between the ground plane and its image, and $\tilde{\boldsymbol{p}}_{st}$ and $\tilde{\boldsymbol{p}}_{model}$ are homogeneous coordinates of the position in the reference camera image and on the stadium model, respectively. The intersection point $\boldsymbol{g}_{model}$ of the projected lines is the position of player/ball on the stadium model. The intersection point is backprojected to reference cameras for obtaining its positions $\boldsymbol{g}_{st1}$ and $\boldsymbol{g}_{st2}$ on the ground plane in the reference images. The distances h_{st1} and h_{st2} are then calculated between the centroid and the ground plane.

2) Projection to the MR Camera Image

After the position on the ground plane of the desktop stadium is obtained, it is projected to MR camera image using projection matrix $\boldsymbol{P}$. The position in MR camera image $\boldsymbol{g_{mr}}$ is obtained as follows:

$$\tilde{\boldsymbol{g}}_{mr} \cong \boldsymbol{P}\, \tilde{\boldsymbol{G}}_{model} \ , \tag{5.12}$$

where $\boldsymbol{G}_{model}$ is 3D position of the intersection point $\boldsymbol{g}_{model}$ in the world coordinate system. The following equation then gives the distance h_{mr} between the centroid and the ground plane in the MR camera image,

$$h_{mr} = (1-\alpha)h_{st1}z_1 + \alpha h_{st2}z_2 \ , \tag{5.13}$$

where z_1 and z_2 are the zoom ratios and α is the interpolating weight. This gives the positions of the players and ball in the MR camera image finally. Even when players are jumping off the ground, the positions of players can be calculated stably.

If the captured scene in the stadium has shadow regions, the shadows of players and ball need to be overlaid on the stadium model. As shadows are projected on the

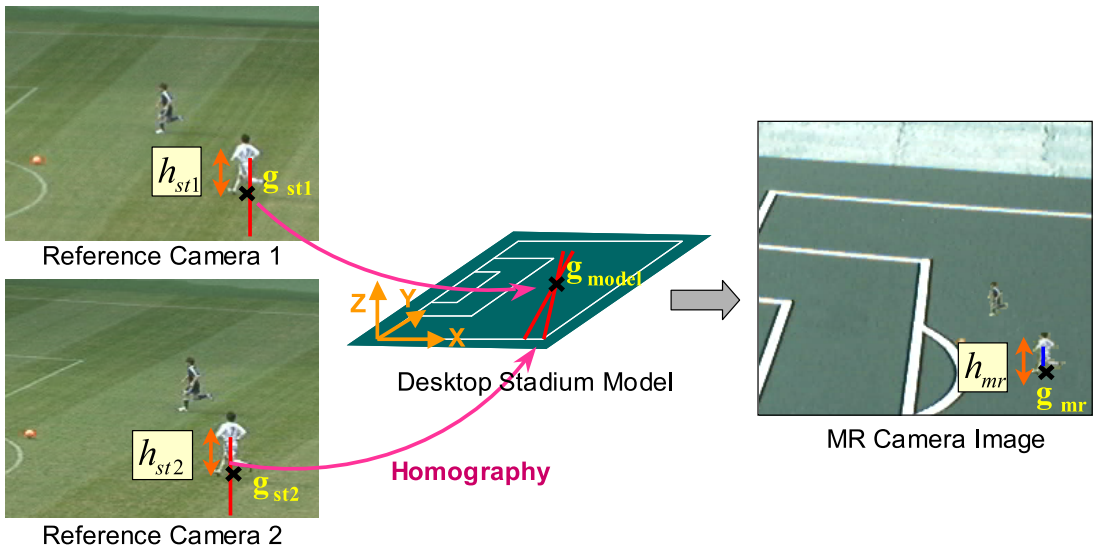


Figure 5.24: Determination of the player and ball positions in the MR camera image via the desktop stadium model.

ground plane, homography transform between the ground plane of the original soccer stadium and its image determines the shadow positions on the desktop stadium. Then projection to the MR camera image can overlay the shadow regions onto the desktop stadium in the camera image. We regard the color value of the shadow regions as to become half as much as that of the original stadium model. Thus the shadows are additionally drawn on the desktop stadium.

5.3.5 Experimental Results

We have constructed a marker-based MR presentation system as the second application in the MR environment. Figure 5.25 describes examples of a marker and a desktop stadium model. The 2D square marker is placed on the stadium model where virtual players, ball and shadows are overlaid. It is assumed that the location of the maker on the model plane and the size of the marker are known. In this experiment, we attached the marker whose size was 4cm at the center of outside of the goal area. A web camera and a monitor were used for capturing the real world and presenting virtual soccer scene overlaid on the stadium model. The system configuration using a web camera is illustrated in Figure 5.5. The system with an HMD can be used alternatively.

The multiple soccer videos captured in the Oita Stadium were used for the replay. The preprocessing is almost identical to that in the experiment with feature-based system. Three processes were added to the list of the preprocessing shown in Section 5.2.4.

- Estimation of the homographies between the ground plane in 3D space and its image in each stadium camera
- Estimation of the rotation matrix and translation vector of the stadium camera
- Calculation of the intrinsic parameters of the MR camera

These processes were implemented as the preprocessing of the MR presentation and the dataset was stored in a PC.

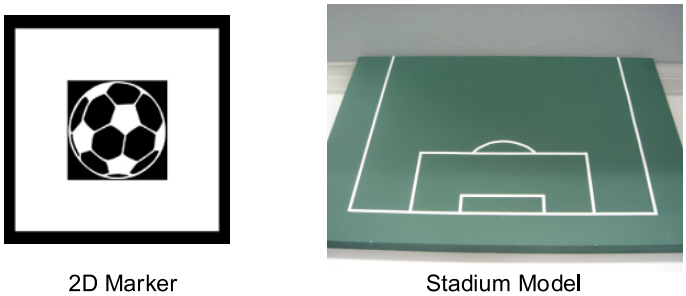


Figure 5.25: Examples of a 2D marker and a desktop stadium model.

During the replay, the stored data are loaded to the PC according to the viewer's viewpoint in each frame. MR presentation is performed in the following order;

1. Capturing the stadium model with the MR camera
2. Extraction of the marker in the camera image
3. Estimation of the rotation matrix and translation vector of the MR camera
4. Calculation of the viewpoint parameters
5. Virtual view synthesis for the players, ball and shadows
6. Calculating the rendering positions of the players, ball and shadows
7. Overlaying them on the stadium model

After replay starts, the marker is detected, and the rotation matrix and translation vector of the MR camera are estimated in each frame; this gives viewpoint parameters. The virtual view images of the dynamic objects are synthesized according to the viewpoint position. Finally the virtual players, ball, and shadow are overlaid at the determined rendering position.

When watching the soccer match from a remote location, the preprocessed data of the soccer scene are transferred via Ethernet from the PC. The projective geometry between stadium cameras, such as fundamental matrices and homographies, the vanishing points, the corner points of the penalty and the goal area, and the rotation matrices and translation vectors of each stadium camera are sent to the remote PC in advance. At each frame in the replay, the texture and 2D positions of all player and ball regions in two reference camera images, the correspondence map, and also the positions of the shadow regions in reference images are sent to the remote PC.

Figure 5.26 and Figure 5.27 present the captured scenes in the stadium and results of the overlaid soccer scenes on the desktop stadium model. The top two images are the reference camera images used for the virtual view synthesis, and the bottom image is the represented soccer scene. The interpolating weight and the zoom ratio are indicated, respectively, at the bottom of each image. We see that the virtual players and ball can be inserted naturally onto the real world. In comparing the overlaid scene with the original soccer scene, the players are located at almost correct positions on the stadium model. On the other hand, the appearance of players and ball is changed by view interpolation appropriately.

Figure 5.28 and Figure 5.29 show a part of the image sequence of free viewpoint MR

presentation. The marker is tracked, and the position and orientation of the camera are obtained successfully. According to the viewpoint, the soccer match is replayed naturally on the desktop stadium model.

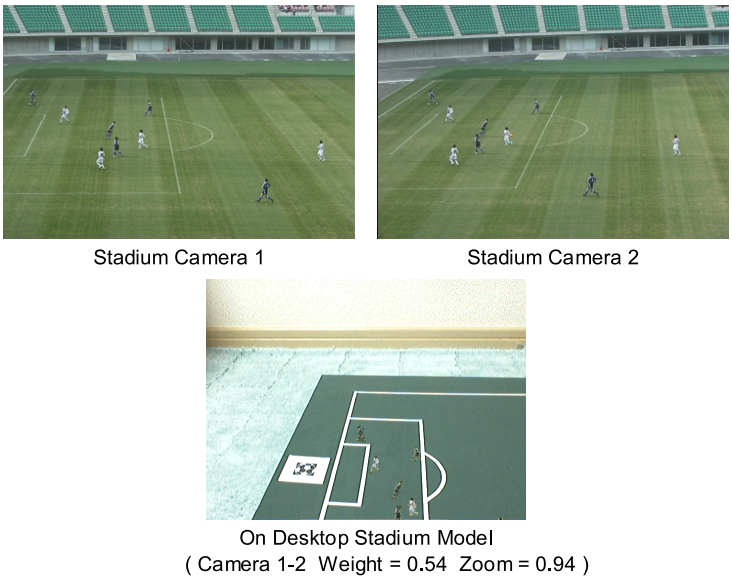


Figure 5.26: Result image #1 of marker-based MR presentation and the original soccer scene.

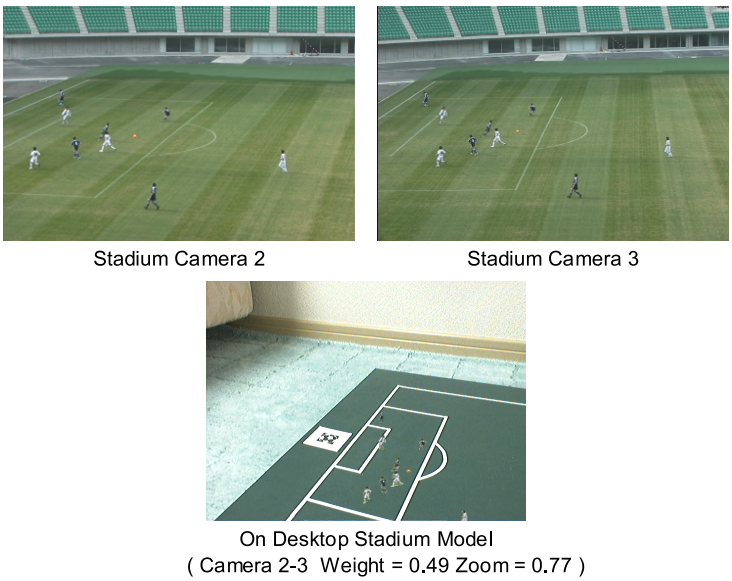


Figure 5.27: Result image #2 of marker-based MR presentation and the original soccer scene.

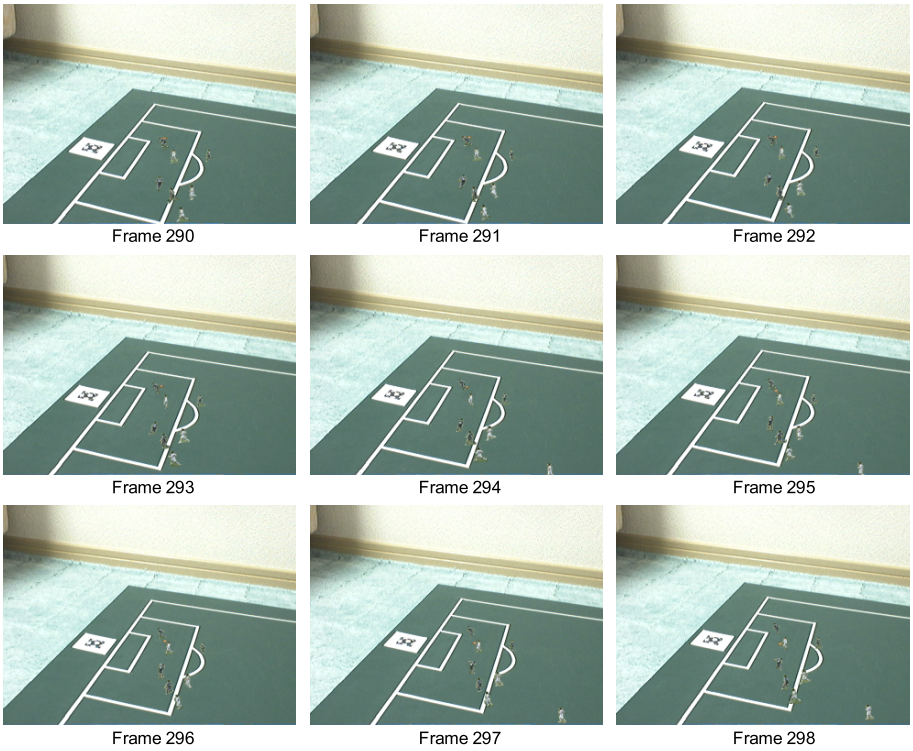


Figure 5.28: Result image sequence #1 in the marker-based MR presentation system.

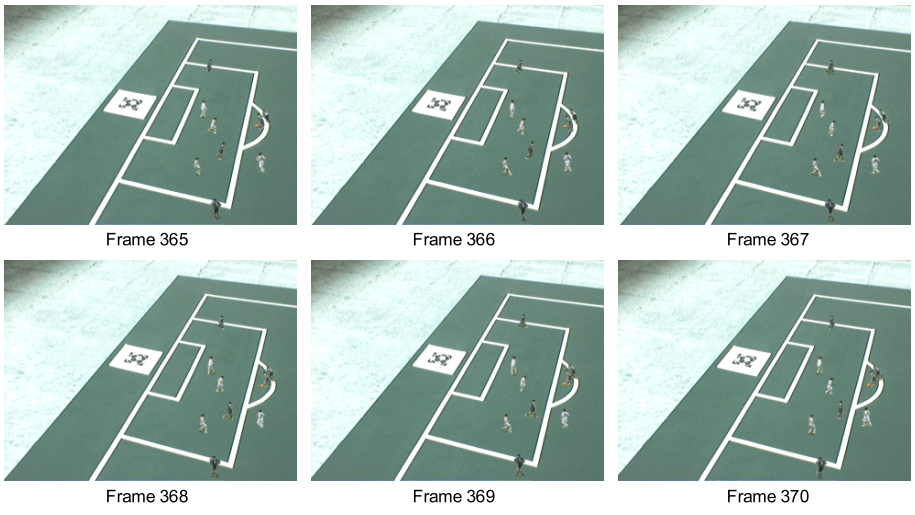


Figure 5.29: Result image sequence #2 in the marker-based MR presentation system.

5.3.6 Discussion

The performance of the marker-based MR presentation system is examined below. The processing time was measured by using the desktop PC (CPU: Pentium 4 3.2GHz, Memory: 2GB, Graphic Card: ATI Radeon 9800). The system runs at 1.7 fps with the desktop PC on an average. The bottleneck is virtual view synthesis process that takes about 0.5 second while the rendering process takes about 0.1 second. As described in the discussion for the Viewpoint on Demand System (Section 4.2.6), the process time can be reduced by the parallel processing. If this problem is solved, the propose system may run at video frame rate.

In the marker-based system, the marker should be visible in MR camera image in every frame. If the marker is not detected or other object is recognized as a marker, the virtual objects are not presented or are presented at incorrect positions. The system is very sensitive to visibility of the marker and this causes unstable appearance of the virtual objects. One of the solutions to deal with this problem is to use additional information obtained in captured images. For example, edges or corner points on the desktop stadium are useful in addition to the marker. If the edges or corner points are detected when the marker detection fails, the camera position (viewpoint position) can be calculated based on the edges or points information. Therefore miss detection can be recovered by combining marker detection and edge or corner extraction. The registration using both marker and edge/corner detection may improve the stability of the proposed system.

Another direction for the improvement is to increase the number of markers. In this case, even when one marker is not detected, detection of another marker can avoid incorrect registration. Use of multi-marker is effective for stable marker detection and precise registration [7]. The pose estimated from a single marker is sometimes not stable enough. The overall pose estimation can be improved by fusing information from several markers. Furthermore, multi-marker is useful for wide-area tracking [131]. In the experiment, registration error was observed in the region far from the marker. This error can be reduced by placement of several markers at different positions. In the case of using multi-marker, it generally needs to predefine the location, pattern and size of each marker in the same way as use of one marker.

Chapter 6:

Conclusions

6.1 Summary

This thesis has presented a novel method of view synthesis for dynamic events in a large space, which allows representations of realistic entire scenes of the event and representations of the dynamic objects in a mixed reality environment from novel viewpoints. With lower level domain knowledge such that the scene consists of some dynamic small areas and almost static planar areas, free viewpoint videos have been successfully synthesized from uncalibrated cameras by utilizing a view interpolation technique. We have presented results of the virtual view synthesis method by generating videos that show virtual camera motions of flying through in dynamic scenes and freeze-and-rotate motions around one object. The Viewpoint on Demand System has enabled a viewer to select his/her viewpoint freely through GUI. This is one of the prototype systems of interactive media for sports broadcasting.

In order to insert such dynamic events in a mixed reality environment, an image-based registration method using projective geometry between cameras has been proposed. This allows us to watch the events overlaid onto the real world. The visualization of a soccer match has been demonstrated on a small desktop stadium in the real world using an HMD or a handheld display with a web camera attached to. The feature-based presentation system has achieved accurate registration between virtual players and the desktop stadium using feature tracking. It has indicated a great effect of image-based registration with projective geometry between cameras. On the other hand, the marker-based system has enabled online mixed reality presentation of a soccer match in the real environment by taking advantage of real-time and robust marker tracking. Interactive visualization of dynamic events was demonstrated effectively in a mixed reality environment.

These systems have made a new way to enjoy sporting events. The proposed approach is based on the condition that objects move on a planar area. It does not require strong calibration of multiple cameras that capture sporting events. The only information that can be obtained in captured images is sufficient for free viewpoint video presentation. Therefore this approach can be easily extended to other sporting events and entertainments.

6.2 Contributions

The major contribution of this thesis is creating a new framework for free viewpoint video synthesis and presentation of dynamic events in a large space with uncalibrated cameras. Most work related free viewpoint video has focused on foreground objects in synthesizing novel views. Typically, human motions such as movements of players in sporting scenes are represented without a background or with computer generated scenes. On the other hand, the entire scene including players and stadium is reconstructed on virtual view images in the proposed method. This gives viewers to more immersive impression into dynamic events than the conventional work. The grate advantage of the proposed methods is unnecessity of camera calibration. The projective geometry between cameras that is used for view synthesis is easily obtained by corresponding feature points in captured images. The interactive visualization system: Viewpoint on Demand System demonstrates great potential for on-demand system of sports broadcasting.

In addition, this thesis has proposed a new approach to inserting a dynamic event in viewer's environment. Although many applications and products have been developed to the enhancement of sport coverage, the conventional visualization just inserts virtual lines/objects into live video, or replays the scenes into graphical animation. The augmented video is presented to viewers typically via a television screen or a computer screen. It does not give enough immersive impression or interactivity to the viewer because the viewer's environment is not taken into account. In this thesis, the systems that overlay a soccer match onto a small desktop stadium in the real world have been constructed. The methods for calculating viewer's position using only captured images have been developed for performing virtual view synthesis. An image-based registration method using homography transform and centroid of the dynamic object has been proposed for overlaying the sporting scene. Feature-based registration and marker-based registration have achieved mixed reality presentation of sporting events. This approach realizes the completely opposite idea of visualization of sporting scenes to the conventional presentation.

Under the above major contributions, there are many minor contributions in this work. To present the entire scene of a dynamic event, virtual view synthesis methods

appropriate for each object included in the target scene have been proposed.

The methods of background generation and subtracting the background have been developed for extracting the dynamic regions accurately in sporting scenes. The mode value of each pixel in an image sequence was selected as the color of the background, and then it was updated according to the lighting condition. Both the intensity and color vector were utilized for background subtraction. Thus the extraction of the dynamic regions can be robust to change of lighting conditions.

The following processes have enabled to obtain a dense correspondence for time-varying objects in a large space effectively. The segmentation method was improved to segment dynamic regions into shadow regions and player/ball regions. Both geometry and color information was utilized for precise segmentation. The silhouette correspondence was performed based on the foot positions of players and homography transform. The problem of establishing correspondence in the case of occlusion was solved by use of the silhouette information in the previous frame and the foot positions in the neighboring view. The pixel correspondence was obtained using silhouette and epipolar geometry.

The representation of shadows in virtual view images is another contribution. Homography transform enabled view synthesis of shadow regions. It is not necessary to estimate light sources or to reconstruct the 3D model of the object. This is a novel idea for synthesizing shadows in virtual view images with transfer-based approach.

For mixed reality presentation, in addition to registration methods, two system configurations have been provided for allowing viewers to watch sporting events in front of them in the real world. One is the system with an HMD in order to provide immersive impression and intuitive interface. The other system consists of a web camera and a handheld display, which is easy to build and can be applicable to outdoor scenes. These systems offer a new type of observation of sporting events. These minor contributions lead to major contributions.

6.3 Future Work

The approach presented in this thesis can be extended in several ways. The following describe these possibilities.

If the proposed systems are put to practical use, the processing speed needs to be improved. The process for virtual view synthesis makes up a large percentage of the time. As mentioned in Section 4.2.6, the processing time for virtual view synthesis has a correlation with the number of dynamic objects. Since corresponding pixels and warping images are implemented independently for every dynamic object, a parallel processing is possible for this part. If we parallelize the algorithm using a PC cluster, the processing speed can be reduced. This means that we can perform the process where the computational cost does not depend on the complexity of the scene.

As the proposed approach is based on the condition that players move on a planar area, the basic ideas of free viewpoint video generation and mixed reality presentation can be useful for the other dynamic events such as entertainments (e.g. rock concert, dance performance) as well as sporting events. When the proposed approach is applied to such events, appropriate scene segmentation is necessary for view synthesis according to the property of the object. In addition, the method of tracking humans (players) should be modified depending on the event. The tracking process involves extraction foreground objects and corresponding them among different cameras. Once the dense correspondence is obtained for the objects among real cameras, the proposed approach works for view synthesis and mixed reality presentation effectively.

A stereoscopic display is an interesting challenge. In order to generate stereoscopic view, it needs to produce disparity between the left image and the right image. If each dynamic object has appropriate disparity according to the distance from a viewer, stereoscopic display become possible with a device such as an HMD. As uncalibrated cameras are used in this work, the distance between the objects and the viewer cannot be obtained explicitly. The relative distance, however, may be useful for producing parallax. The stereoscopic vision will increase the viewer's experience.

6.4 Prospect for Applications

The proposed approach has potential for development of variety of applications. There is high possibility in a system producing digital video special effects. Many kinds of visual effects can be produced for entertainments depending on the requirement. Freeze-and-rotate camera motion is one example. Tracking specified player could be another example. This kind of effect is appropriate for live broadcast or postproduction. Replays of the exciting scenes of the first half of a match can be provided to audiences in the halftime or in rerun of the match on the next day. If the computational performance is improved, presentation of special effects becomes possible just after the play.

Along with the broadcast digitizing and convergence of communication and broadcasting, video on-demand systems are getting more attractive applications. A practical system for interactive visualization can be constructed based on the proposed Viewpoint on Demand System. It may change the current one-way broadcasting in near future.

As regard mixed reality presentation, one possibility is visualization of a soccer match in an empty stadium in the real world. We have introduced systems that allow overlaying soccer scenes onto the desktop stadium model. If the desktop stadium can be replaced by an empty stadium, audience can watch soccer match at real empty stadium. In the 2002 FIFA World Cup, when the game was held in Korea, many Japanese supporters met at the domestic stadium in order to watch the game together on the large screen in the stadium. All supporters stared at the screen during the match. The soccer field was empty at that time. If the soccer match had not been displayed on the screen but represented directly on the field, audience could have enjoyed the match a lot. This is one of the motivations of our work. One easiest way to realize this is to use an HMD for each audience. If the feature lines on the field can be detected, our method described in Section 5.2 can be applied for mixed reality presentation in real empty stadium. Even if the lines are not observed, our approach can be useful when head motion is obtained by other method such as use of positioning sensors.

Another possibility of mixed reality is a remote lecture system. Current systems

typically display just 2D video of a lecturer at remote location. These systems lack realistic high presence and immersion into learning environments. If attendees can look at the lecturer with real environments, they might have great learning experience. Suppose that multiple cameras capture the lecturer from different angles. The proposed method can synthesize images of the lecturer at virtual viewpoints using real camera images. In addition, it can redisplay him/her at remote location in the real world.

Appendices

A Projective Geometry between Cameras

A.1 Epipolar Geometry

Consider the images x and x' of a point X observed by two cameras with optical centers C and C'. These five points all belong to the epipolar plane defined by the two intersecting rays CX and $C'X$ (see Figure A.1). In particular, the point x' lies on the line l' where this epipolar plane and the image plane Π' of the second camera intersect. The line l' is the epipolar line associated with the point x, and it passes through the point e' where the baseline joining the optical centers C and C' intersects Π'. Likewise, the point x lies on the epipolar line l associated with the point x', and this line passes through the intersection e of the baseline with the plane Π. The points e and e' are termed the epipoles of the two cameras. The epipole e' is the (virtual) image of the optical center C of the first camera in the image observed by the second camera, and vice versa. As noted before, if x and x' are images of the same point, then x' must lie on the epipolar line associated with x. This epipolar constraint plays a fundamental role in stereo vision.

The most difficult part of stereo data analysis is establishing correspondences between the two images: deciding which points in the right image match the points in the left one. The search for these correspondences can be restricted to this epipolar line instead of the whole image by the epipolar constraint.

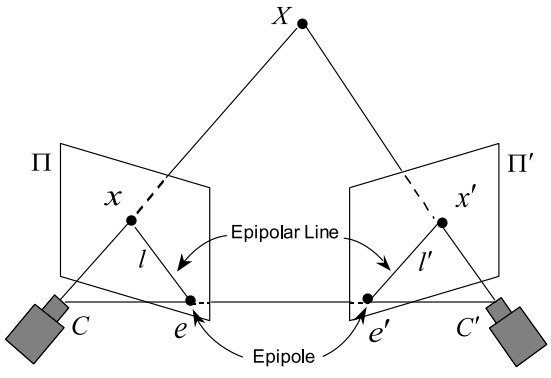


Figure A.1: Epipolar geometry.

A.2 Epipolar Constraint

Here we assume that the intrinsic parameters of each camera are known. Clearly, the epipolar constraint implies that the three vectors $\overrightarrow{Cx}$, $\overrightarrow{C'x'}$, and $\overrightarrow{CC'}$ are coplanar. Equivalently, one of them must lie in the plane spanned by the other two, or

$$\overrightarrow{Cx} \cdot [\overrightarrow{CC'} \times \overrightarrow{C'x'}] = 0 \ . \tag{A.1}$$

We can rewrite this coordinate-independent equation in the coordinate frame associated to the first camera as

$$\tilde{\boldsymbol{x}} \cdot [\boldsymbol{t} \times (\boldsymbol{R}\ \tilde{\boldsymbol{x}}')] \ , \tag{A.2}$$

where $\tilde{\boldsymbol{x}} = (u, v, 1)^\top$ and $\tilde{\boldsymbol{x}}' = (u', v', 1)^\top$ denote the homogeneous image coordinate vectors of x and x', $\boldsymbol{t}$ is the coordinate vector of the translation $\overrightarrow{CC'}$ separating the two coordinate systems, and $\boldsymbol{R}$ is the rotation matrix such that a free vector with coordinates $\boldsymbol{w}$ in the second coordinate system has coordinates $\boldsymbol{R}\ \boldsymbol{w}'$ in the first one.

Equation (A.2) can finally be rewritten as

$$\tilde{\boldsymbol{x}}^\top\ \boldsymbol{E}\ \tilde{\boldsymbol{x}}' = 0 \ , \tag{A.3}$$

where $\boldsymbol{E} = [\boldsymbol{t}]_\times \boldsymbol{R}$, and $[\boldsymbol{a}]_\times$ denotes the skew-symmetric matrix such that $[\boldsymbol{a}]_\times \boldsymbol{x} = \boldsymbol{a} \times \boldsymbol{x}$ is the cross-product of the vectors $\boldsymbol{a}$ and $\boldsymbol{x}$. The matrix $\boldsymbol{E}$ is termed the essential matrix. Its nine coefficients are only defined up to scale, and they can be parameterized by the three degrees of freedom of the rotation matrix $\boldsymbol{R}$ and the two degrees of freedom defining the direction of the translation vector $\boldsymbol{t}$.

When the intrinsic parameters are unknown (uncalibrated cameras), we can write $\tilde{\boldsymbol{x}} = \boldsymbol{A}\tilde{\hat{\boldsymbol{x}}}$ and $\tilde{\boldsymbol{x}}' = \boldsymbol{A}'\tilde{\hat{\boldsymbol{x}}}'$, where $\boldsymbol{A}$ and $\boldsymbol{A}'$ are 3×3 calibration matrices, and $\tilde{\hat{\boldsymbol{x}}}$ and $\tilde{\hat{\boldsymbol{x}}}'$ are normalized image coordinate vectors. The Longuet-Higgins relation holds for these vectors, and we obtain

$$\tilde{\boldsymbol{x}}^\top\ \boldsymbol{F}\ \tilde{\boldsymbol{x}}' = 0 \ , \tag{A.4}$$

where the matrix $\boldsymbol{F} = \boldsymbol{A}^{-\top} \boldsymbol{E}\ \boldsymbol{A}'^{-1}$, termed the fundamental matrix, is not an essential matrix generally. It has rank two, and its null space is $\tilde{\boldsymbol{e}}'$. The null space of $\boldsymbol{F}^\top$

is $\tilde{\boldsymbol{e}}$:

$$\boldsymbol{F}\tilde{\boldsymbol{e}}' = \boldsymbol{F}^\top \tilde{\boldsymbol{e}} = 0 \ , \tag{A.5}$$

where $\tilde{\boldsymbol{e}}$ and $\tilde{\boldsymbol{e}}'$ represent the positions of the epipoles. Note that $\boldsymbol{F}\tilde{\boldsymbol{x}}'$ and $\boldsymbol{F}^\top\tilde{\boldsymbol{x}}$ represent the epipolar lines corresponding to the points x' and x in the first and the second images, respectively.

A.3 Estimation of Fundamental Matrix

The fundamental matrix is defined up to scale by seven independent parameters. It can in principle be estimated from seven point correspondences. This section addresses the problem of estimating the epipolar geometry from a redundant set of point correspondences between two images taken by cameras with unknown intrinsic parameters, a process known as weak calibration.

Equation (A.4) can be rewritten using matrix elements as

$$\boldsymbol{u}_i{}^\top \boldsymbol{f} = 0 \ , \tag{A.6}$$

where

$$\boldsymbol{u}_i = [u_i u_i{}', u_i v_i{}', u_i, v_i u_i{}', v_i v_i{}', v_i, u_i{}', v_i{}', 1]^\top \ ,$$

$$\boldsymbol{f} = [f_{11}, f_{12}, f_{13}, f_{21}, f_{22}, f_{23}, f_{31}, f_{32}, f_{33}]^\top \ ,$$

$\tilde{\boldsymbol{x}}_i = (u_i, v_i, 1)^\top$ and $\tilde{\boldsymbol{x}}_i{}' = (u_i', v_i', 1)^\top$ are the i-th corresponding points, and f_{ij} is an element in the i-th row, j-th column of the fundamental matrix. Given n pairs of correspondence, we obtain

$$\boldsymbol{U}\boldsymbol{f} = \boldsymbol{0} \ , \tag{A.7}$$

where

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_1^\top \\ \vdots \\ \boldsymbol{u}_n^\top \end{bmatrix} \ .$$

Eight point correspondence is sufficient for estimating the fundamental matrix. This is termed the eight-point algorithm in case of $n = 8$. When $n > 8$ correspondences

are available, the matrix can be estimated using linear least squares by

$$\min_{\boldsymbol{F}} \sum_i (\tilde{\boldsymbol{x}}_i{}^\top \boldsymbol{F} \tilde{\boldsymbol{x}}_i{}')^2 \ . \tag{A.8}$$

Note that both the eight-point algorithm and its least-squares version ignore the rank-two property of fundamental matrices. To enforce this constraint, We construct the singular value decomposition of $\boldsymbol{F}$:

$$\boldsymbol{F} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\top \ . \tag{A.9}$$

Here, $\boldsymbol{\Sigma} = diag(\sigma_1, \sigma_2, \sigma_3)$ is a diagonal 3×3 matrix with entries $\sigma_1 \geq \sigma_2 \geq \sigma_3$, and $\boldsymbol{V}$ and $\boldsymbol{U}$ are orthogonal 3×3 matrices. The rank-two matrix $\hat{\boldsymbol{F}}$ minimizing the Frobenius norm of $\boldsymbol{F} - \hat{\boldsymbol{F}}$ is simply $\hat{\boldsymbol{F}} = \boldsymbol{V} diag(\sigma_1, \sigma_2, 0)\ \boldsymbol{U}^\top$. This is the final estimate of the fundamental matrix.

B Projective Geometry between Cameras and a World Plane

Suppose a plane Π in 3D space and points x and x' are images of a point X on the plane Π observed by two cameras which have the following projection matrices:

$$\boldsymbol{P} = [\boldsymbol{I} \mid \boldsymbol{0}],\ \boldsymbol{P}' = [\boldsymbol{R} \mid \boldsymbol{t}]\ .$$

The back-projection of the point x in the first view determines the intersection point X of the ray with the plane Π. The 3D point X is then projected into the second view.

The point x is the projection of any point on the ray $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}^\top, \rho)^\top$, where $\tilde{\boldsymbol{x}} = (u, v, 1)^\top$ denotes the coordinate of x, and ρ parameterizes the point on the ray. The 3D point X on the plane Π satisfies $\tilde{\boldsymbol{\Pi}}^\top \tilde{\boldsymbol{X}} = 0$, where the world plane Π has coordinates $\tilde{\boldsymbol{\Pi}} = (\boldsymbol{v}^\top, 1)^\top$. This determines ρ and $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}^\top, -\boldsymbol{v}^\top \tilde{\boldsymbol{x}})^\top$. The 3D point X is projected into the second view as

$$\begin{aligned} \tilde{\boldsymbol{x}}' &\cong \boldsymbol{P}' \tilde{\boldsymbol{X}} \\ &\cong [\boldsymbol{R} \mid \boldsymbol{t}] \tilde{\boldsymbol{X}} \\ &\cong \boldsymbol{R}\tilde{\boldsymbol{x}} - \boldsymbol{t}\boldsymbol{v}^\top \tilde{\boldsymbol{x}} \\ &\cong (\boldsymbol{R} - \boldsymbol{t}\boldsymbol{v}^\top)\tilde{\boldsymbol{x}}\ . \end{aligned} \tag{B.1}$$

The homography induced by the plane is

$$\tilde{\boldsymbol{x}}' \cong \boldsymbol{H}\tilde{\boldsymbol{x}} \tag{B.2}$$

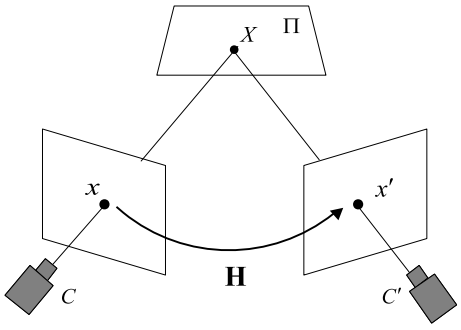


Figure B.1: The homography induced by a plane.

with

$$\boldsymbol{H} = \boldsymbol{R} - \boldsymbol{t}\boldsymbol{v}^\top \ .$$

Applying the transformation $\boldsymbol{A}$ and $\boldsymbol{A}'$ to the images, we obtain the projection matrices $\boldsymbol{P} = \boldsymbol{A}[\boldsymbol{I} \mid \boldsymbol{0}]$ and $\boldsymbol{P}' = \boldsymbol{A}'[\boldsymbol{R} \mid \boldsymbol{t}]$. Given the world plane Π that has coordinates $\boldsymbol{\Pi} = (\boldsymbol{n}^\top, d)^\top$, the resulting homography is

$$\boldsymbol{H} = \boldsymbol{A}'(\boldsymbol{R} - \boldsymbol{t}\boldsymbol{n}^\top/d)\boldsymbol{A}^{-1} \ . \tag{B.3}$$

It is defined by the plane $\boldsymbol{\Pi}$, the camera intrinsic parameters $\boldsymbol{A}$ and $\boldsymbol{A}'$, and extrinsic parameters $\boldsymbol{R}$ and $\boldsymbol{t}$.

References

[1] S. Avidan and A. Shashua, "Novel View Synthesis by Cascading Trilinear Tensors," IEEE Trans. on Visualization and Computer Graphics, Vol. 4, No. 4, pp. 293–306, 1998.

[2] R. T. Azuma, "A survey of augmented reality," Presence, Vol. 6, No. 4, pp. 355–385, 1997.

[3] R. T. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre, "Recent advances in augmented reality," IEEE Computer Graphics and Applications, Vol. 21, Issue 6, pp. 34–47, Nov./Dec. 2001.

[4] N. Babaguchi, Y. Kawai, and T. Kitahashi "Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration," IEEE Trans. on Multimedia, Vol. 4, No. 1, pp. 68–75, 2002.

[5] M. Bajura, H. Fuchs, and R. Ohbuchi, "Merging virtual objects with the real world: Seeing utlrasound," Commun of the ACM, Vol. 36, No. 7, pp. 52–62, 1993.

[6] M. Bajura and U. Neumann, "Dynamic registration correction in video-based augmented reality system," IEEE Computer Graphics and Applications, Vol. 15, No. 5, pp. 52–60, 1995.

[7] G. Baratoff, A. Neubeck, and H. Regenbrecht, "Interactive multi-marker calibration for augmented reality applications," Proc. of International Symposium on Mixed and Augmented Reality, 2002.

[8] T. Bebie and H. Bieri, "SoccerMan - Reconstructing Soccer Games from Video Sequences," Proc. of International Conference on Image Processing, pp. 898–902, 1998.

[9] T. Beier and S. Neely, "Feature-Based Image Metamorphosis," Proc. of SIGGRAPH '92, pp. 35–42, 1992.

[10] O. Bimber, B. Frohlich, D. Schmalstieg, and L. M. Encarnacao, "The Virtual Showcase," IEEE Computer Graphics and Applications, Vol. 21, No. 6, pp. 48–55, 2001.

[11] O. Bimber, "Combining Optical Holograms with Interactive Computer Graphics," IEEE Computer, January issue, pp. 85–91, 2004.

[12] G. Bleser, Y. Pastarmov, and D. Stricker, "Real-time 3d camera tracking for industrial augmented reality applications," Journal of WSCG, pp. 47–54, 2005.

[13] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, "Unstructured lumigraph rendering," Proc. of SIGGRAPH '01, pp. 425–432, Aug. 2001.

[14] J. Canny, "Computational approach to edge detection," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 8, No. 6, pp. 679–698, 1986.

[15] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel, "Free-Viewpoint Video of Human Actors," Proc. of SIGGRAPH '03, Vol. 22, No. 3, pp. 569–577, Jul. 2003.

[16] R. Cavallaro, "The FoxTrax Hockey Puck Tracking System," IEEE Computer Graphics and Applications, Vol. 17, No. 2, pp. 6–12, Mar. 1997.

[17] C. H. Chein and J. K. Aggarawal, "Identification of 3D Objects from Multiple Silhouettes Using Quadtrees/Octrees," Computer Vision, Graphics, and Image Processing, Vol. 36, pp. 100–113, 1986.

[18] S. E. Chen and L. Williams, "View Interpolation for Image Synthesis," Proc. of SIGGRAPH '93, pp. 279–288, 1993.

[19] S. E. Chen, "QuickTime VR-an Image-Based Approach to Virtual Environment Navigation," Proc. of SIGGRAPH '95, pp. 29–38, Aug. 1995.

[20] D. Comaniciu and P. Meer, "Mean shift a robust approach toward feature space analysis," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24, No. 5, pp. 603–619, 2002.

[21] D. Comaniciu, "An algorithm for data-driven bandwidth selection," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 25, No. 2 pp. 281–288, 2003.

[22] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," IEEE Trans. on Pattern Analysis and

Machine Intelligence, Vol. 25, No. 10, pp. 1337–1442, 2003.

[23] "CyberPlay," http://www.orad.co.il/.

[24] P. E. Debevec, C. J. Taylor, and J. Malik, "Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach," Proc. of SIGGRAPH '96, pp. 11–20, Aug. 1996.

[25] W. Du, H. Li, and A. Gagalowicz, "Video based 3D Soccer Scene Reconstruction," Proc. of Mirage 2003, Mar. 2003.

[26] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," IEEE Trans. on Image Processing, Vol. 12, No. 7, pp. 796–807, 2003.

[27] A. Elgammal, D. Hanvood, and L. S. Davis, "Nonparametric model for background subtraction," Proc. of European Conference on Computer Vision, pp. 751–767, Jun. 2000.

[28] A. Elgammal, R. Duraiswami, and L. S. Davis, "Efficient kernel density estimation using the fast Gauss transform with applications to color modeling and tracking," IEEE Trans. on on Pattern Analysis and Machine Intelligence, Vol. 25, No. 11, pp. 1499–1504, 2003.

[29] "EyeVision," http://www.ri.cmu.edu/events/sb35/tksuperbowl.html/.

[30] O. Faugeras and Q. Luong, "The Geometry of Multiple Images," The MIT Press, 2001.

[31] C. Fehn, P. Kauff, O. Schreer, and R. Schafer, "Interactive Virtual View Video for Immersive TV Applications," Proc. of International Broadcasting Convention, 2001.

[32] V. Ferrari, T. Tuytelaars, and L. V. Bool, "Markerless Augmented Reality with a Real-time Affine Region Tracker," Proc. of International Symposium on Augmented Reality, pp. 87–96, 2001.

[33] A. W. Fitzgibbon and A. Zisserman, "Automatic Camera Recovery for Closed or Open Image Sequences," Proc. of European Conference on Computer Vision, pp. 311–326, Jun. 1998.

[34] D. A. Forsyth and J. Ponce, "Computer Vision A Modern Approach," Prentice Hall, 2003.

[35] J. Fründ, C. Geiger, M. Grafe, and B. Kleinjohann, "The Augmented Reality Personal Digital Assistant," Proc. of International Symposium on Mixed

Reality, 2001.

[36] J. Gausemeier, J. Fruend, C. Matysczok, B. Bruederlin, and D. Beier, "Development of a real time image based object recognition method for mobile ARdevices," Proc. of International Conference on Computer Graphics, Virtual Reality, Visualization and Interaction, pp. 133–139, 2003.

[37] C. Geiger, B. Kleinjohann, C. Reimann, and D. Stichling, "Mobile Ar4All," Proc. of International Symposium on Augmented Reality, pp. 181–182, 2001.

[38] Y. Genc, S. Riedel, F. Souvannavong, C. Akinlar, and N. Navab, "Marker-less Tracking for AR: A Learning-Based Approach," Proc. of International Symposium on Mixed and Augmented Reality, pp. 295–304, Sept. 2002.

[39] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," Proc. of SIGGRAPH '96, pp. 43–54, Aug. 1996.

[40] O. Grau, T. Pullen, and G. A. Thomas, "A Combined Studio Production System for 3-D Capturing of Live Action and Immersive Actor Feedback," IEEE Trans. on Circuits and Systems for Video Technology, Vol. 14, No. 3, pp. 370–380, Mar. 2004.

[41] M. Gross, S. Wurmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. V. Gool, S. Lang, K. Strehlke, A. V. Moere, and O. Staadt, "blue-c: A Spatially Immersive Display and 3D Video Portal for Telepresence," Proc. of SIGGRAPH '03, pp. 819–827, 2003.

[42] B. Han, D. Comaniciu, and L. S. Davis, "Sequential kemel density approximation through mode propagation: applications to background modeling," Proc. of Asian Conference on Computer Vision, Jan. 2004.

[43] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2003.

[44] J.-M. Hasenfratz, M. Lapierre, and F. Sillion, "A Real-Time System for Full Body Interaction with Virtual Worlds," Proc. of Eurographics Symposium on Virtual Environments, pp. 147–156, Jun. 2004.

[45] H. Hua, C. Gao, L. Brown, N. Ahuja, and J. P. Rolland, "Using a head-mounted projective display in interactive augmented environments," Proc. of International Symposium on Augmented Reality, pp. 217–223, 2001.

[46] M. Inami, N. Kawakami, D. Sekiguchi, Y. Yanagida, T. Maeda, and S. Tachi, "Visuo-Haptic Display Using Head-Mounted Projector," Proc. of IEEE Virtual

Reality 2000, pp. 233–240, 2000.

[47] N. Inamoto and H. Saito, "Intermediate View Generation of Soccer Scene from Multiple Videos," Proc. of International Conference on Pattern Recognition, Vol. 2, pp. 713–716, Aug. 2002.

[48] N. Inamoto and H. Saito, "Fly-Through Viewpoint Video System for Multi-View Soccer Movie Using Viewpoint Interpolation," Proc. of SPIE Vol. 5150 (Visual Communications and Image Processing), pp. 1143–1151, Jul. 2003.

[49] N. Inamoto and H. Saito, "Immersive Observation of Virtualized Soccer Match at Real Stadium Model," Proc. of International Symposium on Mixed and Augmented Reality, pp. 188–197, Oct. 2003.

[50] N. Inamoto and H. Saito, "Free Viewpoint Video Synthesis and Presentation from Multiple Sporting Videos," Proc. of International Conference on Multimedia and Expo, Jul. 2005.

[51] "InMotion Technologies," http://www.inmotiontech.com/.

[52] S. Iwase and H. Saito, "Parallel Tracking of All Soccer Players by Integrating Detected Positions in Multiple View Images," Proc. of International Conference on Pattern Recognition, Vol. 4, pp. 751–754, Aug. 2004.

[53] T. Kanade, P. J. Narayanan, and P. W. Rander, "Virtualized reality: concepts and early results," Proc. of IEEE Workshop on Representation of Visual Scenes, pp. 69–76, Jun. 1995.

[54] M. Kanbara and N. Yokoya, "Geometric and photometric registration for real–time augmented reality," Proc. of International Symposium on Mixed and Augmented Reality, pp. 279–280, 2002.

[55] H. Kato, M. Billinghurst, R. Blanding, and R. May, "ARToolKit," http://www.hitl.washington.edu/artoolkit/.

[56] I. Kitahara and Y. Ohta, "Scalable 3D Representation for 3D Video in a Large-Scale Space," PRESENCE, The MIT Press, Vol. 13, Issue 2, pp. 164–177, 2004.

[57] G. Klein and T. Drummond, "Sensor Fusion and Occlusion Refinement for Tablet-based AR," Proc. of International Symposium on Mixed and Augmented Reality, pp. 38–47, 2004.

[58] J. Kollin, "A Retinal Display For Virtual-Environment Applications," Proc. of International Symposium, Digest of Technical Papers, pp. 827, 1993.

[59] T. Koyama, I. Kitahara, and Y. Ohta, "Live Mixed-Reality 3D Video in Soccer

Stadium," Proc. of International Symposium on Mixed and Augmented Reality, pp. 178–187, 2003.

[60] K. N. Kutulakos and J. Vallino, "Calibration-free augmented reality," IEEE Trans. on Visualization and Computer Graphics, Vol. 4, No. 1, 1998.

[61] M. Levoy and P. Hanrahan, "Right field rendering," Proc. of SIGGRAPH '96, pp. 31–42, Aug. 1996.

[62] B. Li and MI Sezan, "Event detection and summarization in American football broadcast video," Proc. of SPIE Conference on Storage and Retrieval for Media Databases, Vol. 4676, pp. 202–213, 2002.

[63] B. P. L. Lo and S. A. Velastin, "Automatic congestion detection system for underground platforms," Proc. of ISIMP2001, pp. 158–161, May 2001.

[64] "Lucent," http://www.lucent.com/.

[65] C. Loscos, G. Drettakis, and L. Robert, "Interactive virtual relighting of real scenes," IEEE Trans. Visualization and Computer Graphics, Vol. 6. No. 4, pp. 289–305, 2000.

[66] C. Malerczyk, K. Klein, and T. Wiebesiek, "3D Reconstruction of Sports Events for Digital TV," Journal of WSCG, Vol. 11, No. 1, 2003.

[67] R. A. Manning and C. R. Dyer, "Interpolating View and Scene Motion by Dynamic View Morphing," Proc. of International Conference on Computer Vision and Pattern Recognition, pp. 1388–1394, 1999.

[68] K. Matsui, M. Iwase, M. Agata, T. Tanaka, and N. Ohnishi, "Soccer Image Sequence Computed by a Virtual Camera," Proc. of International Conference on Computer Vision and Pattern Recognition, pp. 860–865, 1998.

[69] T. Matsuyama, X. Wu, T. Takai, and T. Wada, "Real-Time Dynamic 3D Object Shape Reconstruction and High-FidelityTexture Mapping for 3D Video," IEEE Trans. on Circuits and Systems for Video Technology, Vol. CSVT-14, No. 3, pp. 357–369, 2004.

[70] W. Matusik, C. Buehler, R. Raskar, S. Gorlter, and L. McMillan, "Image-Based Visual Hulls," Proc. of SIGGRAPH '00, pp. 369–374, Jul. 2000.

[71] W. Matusik, C. Buehler, and L. McMillan, "Polyhedral Visual Hulls for Real-Time Rendering," Proc. of Eurographics Workshop on Rendering, pp. 116–126, 2001.

[72] L. McMillan and U. Bishop, "Plenoptic Modeling: An Image-Based Rendering

System," Proc. of SIGGRAPH '95, pp. 39–46, Aug. 1995.

[73] L. McMillan, "An image-based approach to three-dimensional computer graphics," Ph.D. dissertation, Department of Computer Science, University of North Carolina, Chapel Hill, NC, 1999.

[74] S. Moezzi, L. C. Tai, and P. Gerard, "Virtual View Generation for 3D Digital Video," IEEE Multimedia, Vol. 4, Issue 1, pp. 18–26, 1997.

[75] M. Möhring, C. Lessig, and O. Bimber, "Video see-through AR on consumer cell-phones," Proc. of International Symposium on Mixed and Augmented Reality, pp. 252–253, 2004.

[76] U. Neumann, S. You, J. Hu, B. Jiang, and J. W. Lee,, "Augmented Virtual Environments (AVE): Dynamic Fusion of Imagery and 3D Models," Proc. of IEEE Virtual Reality 2003, pp. 61–67, 2003.

[77] T. H. Duy Nguyen, T. C. Thien Qui, K. Xu, A. D. Cheok, S. L. Teo, Z. Zhou, A. Mallawaarachchi, S. P. Lee, W. Liu, H. S. Teo, L. N. Thang, Y. Li, and H. Kato, "Real-Time 3D Human Capture System for Mixed-Reality Art and Entertainment," IEEE Trans. on Visualization and Computer Graphics, Vol. 11, No. 6, pp. 706–721, Nov./Dec. 2005.

[78] T. Ogi, T. Yamada, K. Yamamoto, and M. Hirose, 'Invisible Interface for Immersive Virtual World," Proc. of the Immersive Projection Technology Workshop, pp. 237–246, 2001.

[79] Y. Ohta and T. Kanade, "Stereo by Intra- and Inter-scanline Search Using Dynamic Programming, " IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 7, No. 2, pp. 139–154, 1998.

[80] M. Okutomi and T. Kanade, "A Multiple-Baseline Stereo," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 15, pp. 353–363, Apr. 1993.

[81] "Orad," http://www.orad.co.il/.

[82] W. Pasman and C. Woodward, "Implementation of an augmented reality system on a PDA," Proc. of International Symposium on Mixed and Augmented Reality, pp. 276–277, 2003.

[83] M. Piccardi, T. Jan, "Efficient mean-shift backgonmd subtraction, " Proc. of IEEE 2004 KIP, Singapore, Oct. 2004.

[84] G. Pingali, A. Opalach, Y. Jean, and I. Carlbom, "Visualization of sports using motion trajectories: Providing insights into performance, style, and strategy,"

Proc. of IEEE Visualization Conference (Vis), pp. 75–82, 2001.

[85] "The PISTE Project," http://piste.intranet.gr/.

[86] S. Pollard, M. Pilu, S. Hayes, and A. Lorusso, "View synthesis by trinocular edge matching and transfer," Image and Vision Computing, Vol. 18, pp. 749–757, 2000.

[87] M. Pollefeys, R. Koch, and L. VanGool, "Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters," Proc. of International Conference on Computer Vision, 1998.

[88] M. Potmesil, "Generating Octree Models of 3D Objects from their Silhouettes in a Sequence of Images," Computer Vision, Graphics, and Image Processing, Vol. 40, Issue 1, pp. 277–283, 1987.

[89] S. Prince, A. D. Cheok, F. Farbiz, T. Williamson, N. Johnson, M. Billinghurst, and H. kato, "3D live: Real Time Captured Content for Mixed Reality," Proc. of International Symposium on Mixed and Augmented Reality, pp. 7–13, Sept. 2002.

[90] "Princeton Video," http://www.pvi-inc.com/.

[91] H. L. Pryor, T. A. Furness III, and E. Viirre, "The Virtual Retinal Display: A New Display Technology Using Scanned Laser Light," Proc. of Human Factors and Ergonomics Society, 42nd Annual Meeting, pp. 1570-1574, 1998.

[92] "QuesTec," http://www.questec.com/.

[93] R. Raskar, G. Welch, K. L. Low, and D. Bandyopadhyay, "Shader Lamps: Animating real objects with image-based illumination," Proc. of Eurographics Rendering Workshop, pp. 89–102, 2001.

[94] J. P. Rolland, L. D. Davis, and Y. Baillot, "A Survey of Tracking Technologies for Virtual Environments," Fundamentals of Wearable Computers and Augmented Reality, W. Barfield and T. Caudell, eds. , Lawrence Erlbaum, Mahwah, N.J., pp. 67-112, 2001.

[95] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV Baseball programs," Proc. of ACM Multimedia '00, pp. 105–115, 2000.

[96] H. Saito, S. Baba, and T. Kanade,, "Appearance-Based Virtual View Generation From Multicamera Videos Captured in the 3-D Room," IEEE Trans. on Multimedia, Vol. 5, No. 3, pp. 303–316, 2003.

[97] I. Sato, Y. Sato, and K. Ikeuchi, "Acquiring a radiance distribution to super-

impose virtual objects onto a real scene," IEEE Trans. on Visualization and Computer Graphics, Vol. 5, No. 1, pp. 1–12, 1999.

[98] I. Sato, Y. Sato, and K. Ikeuchi, "Illumination distribution from shadows," Proc. of International Conference on Computer Vision and Pattern Recognition, pp. 306–312, Jun. 1999.

[99] K. Satoh, S. Uchiyama, H. Yamamoto, and H. Tamura, "Robust Vision-Based Registration Utilizing Bird's-Eye View with User's View," Proc. of International Symposium on Mixed and Augmented Reality, pp. 46–55, 2003.

[100] K. Satoh, S. Uchiyama, and H. Yamamoto, "A Head Tracking Method Using Bird's-Eye View Camera and Gyroscope," Proc. of the International Symposium on Mixed and Augmented Reality, pp. 202–211, 2004.

[101] "Scidel," http://www.scidel.com/.

[102] S. M. Seitz and C. R. Dyer, "View Morphing," Proc. of SIGGRAPH '96, pp. 21–30, Aug. 1996.

[103] Y. Seo and K. Hong, "Calibration-free augmented reality in perspective," IEEE Trans. on Visualization and Computer Graphics, Vol. 6, No. 4, pp. 346–359, 2000.

[104] J. Shade, S. Gortler, L.-W. He, and R. Szeliski, "Layered depth images," Proc. of SIGGRAPH '98, pp. 231–242, Jul. 1998.

[105] H.-Y. Shum and L.-W. He, "Rendering with Concentric Mosaics," Proc. of SIGGRAPH '99, pp. 299–306, Aug. 1999.

[106] H.-Y. Shum, S. B. Kang, and S.-C. Chan, "Survey of Image-Based Representations and Compression Techniques," IEEE Trans. on Circuits and Systems for Video Technology, Vol. 13, No. 11, pp. 1020–1037, Nov. 2003.

[107] G. Simon, A. W. Fitzgibbob, and A. Zisserman, "Markerless Tracking using Planar Structures in the Scene," Proc. of International Symposium on Augmented Reality, pp. 120–128, Oct. 2000.

[108] A. State, G. Hirota, D. Chen, W. Garrett, and M. Livingston, "Superior augmented reality registration by integrating landmark tracking and magnetic tracking," Proc. of SIGGRAPH '96, pp. 429–438, Aug. 1996.

[109] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," Proc. of International Conference on Computer Vision and Pattern Recognition, pp. 246–252, Jun. 1999.

[110] R. Subbaraoyz, P. Meery, and Y. Gencz, "A Balanced Approach to 3D Tracking from Image Streams," Proc. of International Symposium on Mixed and Augmented Reality, pp. 70–78 , Oct. 2005.

[111] N. Sugano, H. Kato, and K. Tachibana, "The Effects of Shadow Representation of Virtual Objects in Augmented Reality," Proc. of International Symposium on Mixed and Augmented Reality, pp. 76–83, Oct. 2003

[112] "Symah Vision," http://www.symah-vision.fr/.

[113] R. Szeliski and H.-Y. Shum, "Creating Full View Panoramic Image Mosaics and Texture-Mapped Models," Proc. of SIGGRAPH '97, pp. 251–258, Aug. 1997.

[114] "The Matrix," http://whatisthematrix.warnerbros.com/.

[115] R. Y. Tsai, "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses," IEEE Journal of Robotics and Automation, Vol. RA-3, No. 4, pp. 323–344, Aug. 1987.

[116] "2D3," http://www.2d3.com/.

[117] J. Underkoffler, B. Ullmer, and H. Ishii, "Emancipated pixels: real-world graphics in the luminous room," Proc. of SIGGRAPH '99, pp. 385–392, Aug. 1999.

[118] L. Vacchetti, V. Lepetit, and P. Fua, "Stable real-time 3D tracking using online and offline information, " IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 26, No. 10, pp. 1385–1391, 2004.

[119] S. Vedula, P. W. Rander, H. Saito, and T. Kanade, "Modeling, Combining, and Rendering Dynamic Real-World Events From Image Sequences," Proc. of International Conference on Virtual Systems and Multimedia, Vol. 1, pp. 326–322, 1998.

[120] S. Vedula, S. Baker, and T. Kanade, "Spatio-temporal view interpolation," Proc. of Eurographics Workshop on Rendering, pp. 65–76, 2002.

[121] D. Wagner and D. Schmalstieg, "First steps towards handheld augmented reality," Proc. of International Conference on Wearable Computers, pp. 127–136, 2003.

[122] K. W. Wan, X. Yan, X. Yu, and C. Xu., "Robust Goalmouth Detection for Virtual Content Insertion," Proc. of ACM Multimedia, pp. 468–469, 2003.

[123] J. R. Wang and N. Parameswaran, "Survey of Sports Video Analysis: Research Issues and Applications," Proc. of Pan-Sydney Area Workshop on Visual Information Processing, Vol. 36, pp. 87–90, 2003.

[124] Y. Wexler and A. Shashua, "On the Synthesis of Dynamic Scenes from Reference Views," Proc. of International Conference on Computer Vision and Pattern Recognition, pp. 1576–1581, 2000.

[125] C. Wren, A. Azarhayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, pp. 780–785, 1997.

[126] H. Wuest, F. Vial, and D. Stricker, "Adaptive Line Tracking with Multiple Hypotheses for Augmented Reality," Proc. of International Symposium on Mixed and Augmented Reality, pp. 62–69 , Oct. 2005.

[127] J. Xiao, C. Rao, and M. Sha, "View Interpolation for Dynamic Scenes," Proc. of Eurographics 2002.

[128] S. Yaguchi and H. Saito, "Arbitrary Viewpoint Video Synthesis from Multiple Uncalibrated Cameras," IEEE Trans. on Systems, Man and Cybernetics, PartB, Vol. 34, No. 1, pp. 430–439, 2004.

[129] X. Yan, X. Yu, and T. S. Hay, "A 3D Reconstruction and Enrichment System for Broadcast Soccer Video," Proc. of International Conference on Multimedia and Expo, pp. 746–747, 2004.

[130] X. Yu and D. Farin, "Current and Emerging Topics in Sports Video Processing," Proc. of International Conference on Multimedia and Expo, 2005.

[131] J. Zauner and M. Haller, "Authoring of Mixed Reality Applications including Multi-Marker Calibration for Mobile Devices", Proc. of Eurographics Symposium on Virtual Environments, 2004.

[132] Z. Zhang, "A flexible new technique for camera calibration," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, No. 11, pp. 1330–1334, 2000.