Temporal and Spatial Vision Augmentation for Perceiving Key Moments in Virtual Reality Sports

Submitted to
Graduate School of Information Science and Technology
Osaka University

January 2023

Tao TAO

Thesis Committee:

Prof. Haruo Takemura (Osaka University)
Prof. Tatsuhiro Tsuchiya (Osaka University)
Prof. Noriyuki Miura (Osaka University)
Assoc. Prof. Yuki Uranishi (Osaka University)

List of Publications

Journals

1. T. Tao, P. Ratsamee, J. Orlosky, Y. Uranishi, and H. Takemura. MomentViz: An Interactive 3D Vision Augmentation Framework for Rapid Motion. *Transactions of the Virtual Reality Society of Japan*, 25(2):158-168, 2020.

International Conferences

Under Review

1. T. Tao, P. Ratsamee, C. Liu, Y. Uranishi, and H. Takemura. SmartVP: Viewpoint Optimization Based on Individual Preference for Watching 3D Boxing Punch Videos. 2023.

Peer-reviewed

1. T. Tao, P. Ratsamee, Y. Uranishi, J. Orlosky, M. Higashida, and H. Takemura. An interactive 4D vision augmentation of rapid motion, *Proceedings of the 9th Augmented Human International Conference*, 1–4, Feb. 2018.

Domestic Conferences

Non-peer-reviewed

1. T. Tao, P. Ratsamee, Y. Uranishi, T. Mashita, K. Kiyokawa, H. Takemura, and T. Fukuda. SmartVP: An autonomous Viewpoint Selection for Watching a Boxing Game in Virtual Reality, *SCI'21 research presentation*, TS12-02-4, 2021.

Thesis

- T. Tao. H4D Visualization: An Interactive Visualization with High Frame Rate and Time-Control in 3D for Rapid Motion, *Master's Thesis, Graduate School of Information Science and Technology, Osaka University*, Feb, 2018.

Abstract

Along with the high-speed development in the technology and commercialization of virtual reality (VR) devices, spectating VR sports has become popular. Compared with the traditional spectating style: spectating sports matches in a stadium live, spectating VR sports has several advantages. For instance, spectating VR sports only requires one VR device, allowing spectators to enjoy the sports matches anywhere, anytime, and from any free viewpoint with a more immersive experience. On the other hand, spectating VR sports still has several limitations. For example, high bandwidth requirements, general VR sickness, misperceiving key moments, etc. This work focuses on the problem of misperceiving key moments. The key moment in sports is a short and critical moment in a sports action, including the information that can influence or determine the action result. However, there is easy to misperceive the key moment when spectating sports matches in VR because of the temporal and spatial limitations caused by the rapid and occluded actions. This work proposes temporal and spatial vision augmentation frameworks to improve the spectator's ability to perceive the key moment in spectating VR sports. This work includes two parts: 1) Time-control interactive 3D visualization framework for rapid actions; 2) viewpoint optimization based on individual preference for occluded actions.

Many sports, like boxing, basketball, and so on, involve rapid actions. However, human eyes commonly do not have enough time to perceive and process such rapid actions. This temporal limitation leads spectators to misperceive the key moment in spectating VR sports. While much research got exemplary achievements in visualizing a rapid scene, less work involved VR rapid sports scenes. This work presents an interactive 3D vision augmentation framework called MomentViz, which allows for both high frame rate recording and interactive Time-control in 3D space. The system is designed to allow users to freely spectate rapid 3D actions from different viewpoints and control time from any free viewpoint. This work starts from 3D data collection and 3D reconstruction. Then the slow time frame and original time frame are set in the same place, and the original time frame overlaid the slow time one. Through a VR HMD, the users can select any area to control the frame time by raycasting with a controller. As a result, the overlaid slow time frame can emerge through the stencil technique in that user-selected sub-region. This method allows the user to control the time in their desired area. A simple pilot experiment is conducted to verify the necessity and performance of MomentViz. Eight participants are recruited to join the experiment. They are asked to observe the rapid motion in VR with four conditions, including **2D** visualization, **2D Time-control** visualization, **3D** visualization, and **3D Time-control** Visualization (MomentViz). By the analysis of results, MomentViz outperforms other groups. From the participants' feedback, it is found that the low FPS recording limits the video quality and influences the performance of MomentViz. In order to overcome this limitation, a high-speed RGB camera is implemented in improved MomentViz for high FPS RGB data recording. A user study experiment is conducted to evaluate the empirical performance of this improved MomentViz. Twelve participants join the experiment and are asked to judge the rapid basketball videos. Three groups are compared and evaluated, including **3D** visualization (original time videos), **3D Time-control** MomentViz (not using the high-speed camera), and **H3D Time-control** MomentViz (using the high-speed camera). Results showed H3D Time-control MomentViz group outperforms the others in terms of accuracy, required views, and subjective user experience.

In addition, in these sports, the players always keep moving, and their poses are usually dynamic. It leads to the key moment of these actions that will be occluded from some viewpoints. Moreover, finding an excellent viewpoint to perceive the key moment from an occluded action in VR is challenging to the spectators. Though some research has tried to solve this viewpoint optimization issue, all of them ignore the individual preference of spectators and provide all the spectators with a monotonous viewpoint. This work introduces a novel method to select an optimal viewpoint for watching punching moments in VR. This method can handle different spectator preferences. In this method, a visibility model is customized, which utilizes eight bounding boxes to account for the visibility of upper body parts. Then a neural network classification model is utilized to reproduce the optimal viewpoint selection based on the features of body parts visibility, punch side, and punch offset. An experiment collects the spectators' preferred viewpoints from 24 participants under three controlled preference conditions: seeing the punch arm form clearly (**AF**), seeing the facial expression clearly (**FE**), and no additional restriction (**None**). With a new analysis method for this viewpoint optimization issue, the prediction accuracy of the trained model is evaluated as the scientific performance. The accuracy results are 64.93% for None, 68.61% for AF, and 76.48% for FE, respectively. A user study is conducted to evaluate the empirical performance of this method. Twenty-one participants are recruited to participate in the user study, and their task is to find their preferred viewpoint depending on different preference conditions (**AF** and **FE**). Two groups are divided to compare, including the **SmartVP** group and the **Without SmartVP** group. The subjective and objective evaluation results show that the SmartVP group outperforms the Without SmartVP group.

The results of this work show the promise and potential of VR sports broadcasting applications. The author hopes this work can encourage the development of portable MR devices for use in watching sporting events both by broadcasting and live in the stadium.

Acknowledgments

As the final outcome of my academic journey, this dissertation highly summarizes the results of my research over several years. It could not be completed without the support and help of my advisors, other laboratory staffs, family, and friends.

I would like to thank my supervisors Prof. Haruo Takemura and Assist. Prof. Photchara Ratsamee for their support, comments and guidance during my whole Ph.D. program. Their kind helps also are an important factor in the success of this work.

I also would like to express my appreciation to Prof. Kiyoshi Kiyokawa for inviting me to join Takemura laboratory and giving me a chance to start this work. My gratitude also goes out to Assoc. Prof. Yuki Uranishi for the support of experiment preparation. Another thank from me goes out to Assoc. Prof. Jason Orlosky for revising my paperwork and advising on my writing skill. I appreciate the help of my friend, the OB of Takemura laboratory, Dr. Chang Liu, for the inspirational comments on my study.

I also would like to thank the secretary of Takemura laboratory, Ms. Hiroko Takahashi. As an international student, I got a lot of help from her, including university documents and government procedures. I want to thank all the members of the Takemura Laboratory during my laboratory life for correcting my Japanese and sharing the various interesting cultures. Their kind friendship supports me in studying in a foreign country alone. In particular, I really would like to express my gratitude to my friend Harn Sison, who always helped me test the pilot experiment or be a model in data collection trials.

I want to express a special appreciation to the Ishibashi boxing gym for providing the boxing ring and boxers for actual boxing data recording. I also would like to thank everyone who participated in the experiments presented in this dissertation.

Finally, I wish to express my most gratitude to my family, who supported me the whole time, always encouraged me to keep forward to the goal and never stopped believing in me.

Tao Tao
Osaka University
January 2023

Contents

List of Tables

List of Figures

Chapter 1
Introduction

1.1 Background

Sports is an activity involving physical exertion and skill in which an individual or a team competes against another or others for entertainment [Jarvie (2013)]. It has a long history and attracts millions of people around the world. To the report, the global TV audience was 3.5 billion for the 2018 world cup in Russia, which is predicted to increase to 5 billion for the 2022 Qatar world cup [Burns (2022)]. Sports fans can spectate the sports match mainly through two types: spectating the match in the stadium live or watching the broadcasting by TV and Internet.

Compared with broadcasting, spectating live at a stadium is more attractive for most of the spectators. The reasons can be attributed to these as follows: 1). Watching the game in an actual gym, stadium, or theater can provide spectators with excitement and give personal experience at the venue. 2). Interaction such as cheering or rooting for the player in the stadium is hard to replace since the spectators' cheers directly reach the players. The spectators and players can share the excitement directly in real-time through body touching or eye contact. 3). Socializing with other supports in the stadium is also attractive and an effective way to make new friends with the same interest. On the contrary, spectating sports matches live in a stadium has several limitations. For instance, spectating sports matches in a stadium obviously requires the spectators to go to a certain place, a sports stadium. Sometimes this stadium is so far that the spectators should take lots cost to get there. In addition, spectating sports in a stadium also limits the spectators' time. The spectators are required to ensure a certain time available for that match. Moreover, a seat is usually fixed when spectating sports in a stadium. It leads to the spectators' viewpoints being unfree, sometimes resulting in the occluded view.

In recent years, the development of Virtual Reality (VR) devices and services has been outstanding. This achievement provides other possible solutions to enjoy and spectate sports matches without the abovementioned limitations.

1.1.1 Development of Virtual Reality

Virtual Reality (VR) is an interactive and immersive (with the feeling of presence) experience in a simulated (autonomous) world [Zeltzer (1992)]. This

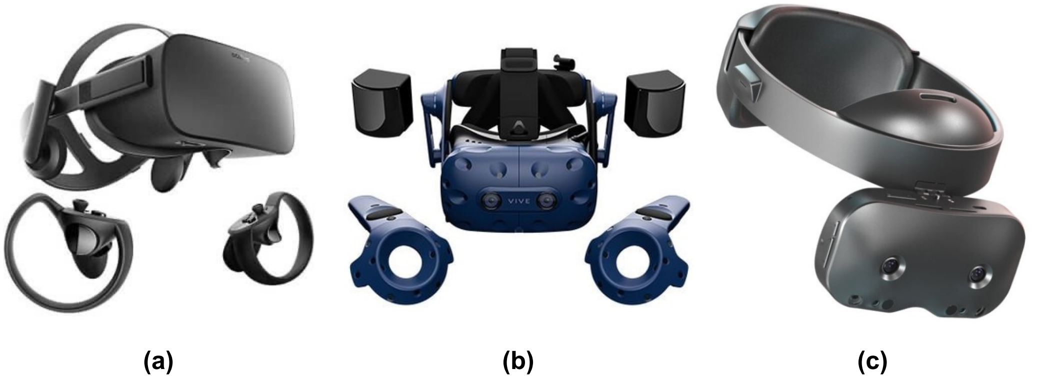


Figure 1.1: Some VR devices: (a) Oculus Rift (Released in 2016 by Oculus). (b) HTV Vive (Released in 2016 by HTV). (c) Lynx R1 (Released in 2021 by Lynx).

idea can be traced back to the 1800s, almost the beginning of practical photography. In the 1960s, a head-mounted display (HMD) was designed by Heilig, which can be regarded as the origin of VR devices. After a hard time in the 1970s and early 1980s, the first commercial HMD called EyePhones was introduced in the late 1980s. The term "virtual reality" was first used in the mid-1980s by Jaron Lanier, the founder of VPL Research. VPL Research concentrates on developing VR devices, including goggles and gloves [Burdea and Coiffet (2003)]. Since then, VR has become a popular topic both in the academic and industrial worlds. Many researchers and companies have pioneered low-cost, high-quality, more portable devices. In the 21st century, VR technology has kept high-speed development and has become commercially available. By 2016, more than 230 companies were developing VR-related products. This year, Oculus released the VR HMD Oculus Rift, and HTC shipped the first major commercial VR device with sensor-based tracking. With the fruitful release of major commercial VR devices, VR technology enters a new epoch. In recent years, as the new term: mixed reality (MR) has come to the front, the commercial HMD tends to MR, which mixes Augmented Reality and Virtual Reality. Lynx releases their first MR production, Lynx-R1, which can easily switch the VR and AR environment with one HMD device (Figure 1.1).

As a result of Virtual Reality development, VR has been applied to various fields. It covers data and architectural visualization, modeling, and designing, training and education, remote operating, cooperating working, and entertainment [Mazuryk and Gervautz (1999)]. The new concept of Metaverse frequently dominates the hot news and has become a quietly popular topic worldwide. Metaverse is a universal, immersive virtual world that unites all VR users in science fiction. This world is a platform that provides almost the whole services mentioned above. Several major IT companies, such as

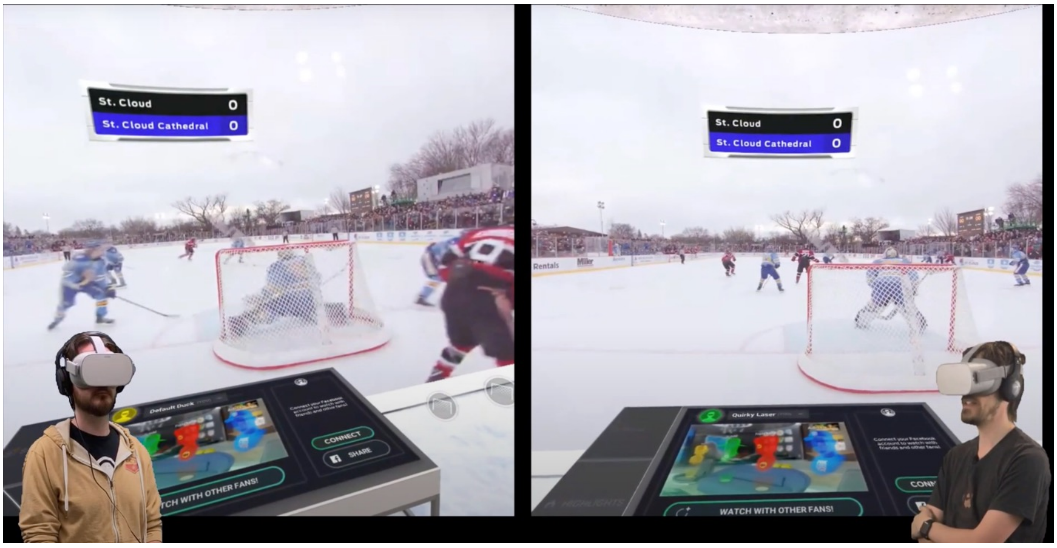


Figure 1.2: Spectating VR sports case: Spectating an ice hockey matches in VR through a VR device (screen shot from the video of FOX Sports[1]).

Facebook, Microsoft, and Google, show interest in this topic.

1.1.2 Spectating Virtual Reality Sports

As an important part of the entertainment service, spectating sports matches in VR has flourished and provides a novel and fashional enjoyment style of sports spectating. Spectating Virtual Reality sports uses a VR device to spectate a real sports match broadcasting in a VR environment. Compared with the traditional style of spectating sports matches in a stadium, this new style just requires a VR device. The VR device allows spectators to enjoy sports matches anywhere, even at home. With the VR sports data rebroadcast or replay, the spectators enjoy the sports anytime and never worry about the time missing again. Moreover, VR sports provide an immersive 3D VR environment, and the spectators can spectate the sports matches from any free viewpoint.

Some pioneer sites, such as FOX Sports [1] and Subvrsive [2], provide the VR sports stream service. Since 2020, the NBA has had contracts with the VR device company Oculus to provide VR spectating services. The Japanese professional baseball team SoftBank Hawks also provides the VR live stream of all the home matches. Enjoying a sports match via VR is not an out-of-reach dream (Figure 1.2).

However, spectating VR sports still has several issues, such as high bandwidth requirements, general VR sickness, key moment misperceiving, and so

[1] https://www.youtube.com/watch?v=T3Ip9s2WvCc

[2] https://subvrsive.com/blog/showtime-championship-boxing-preview-thurman-vs-garcia

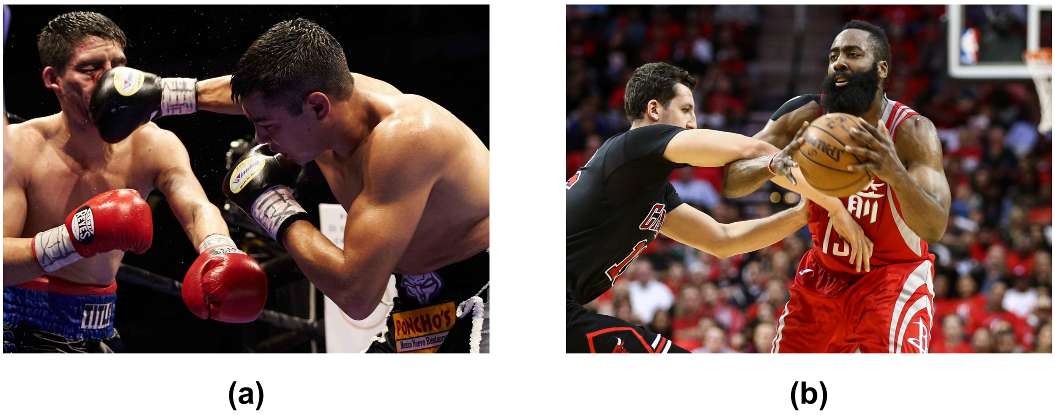


Figure 1.3: Two cases of key moment in sports actions: (a) The key moment in boxing punch actions can determine whether this punch hits the target or is dodged; (b) The key moment in basketball ball stealing actions can judge whether this stealing is foul.

on.

In order to keep the portability, the current major VR devices use the wireless network. When the coding mode of 4K video, the bandwidth demand is at least 840Mbps, and if the spectators want to enjoy the 3D experience of the VR video with 120 FPS, it is at least 4.2Gbp [Yang et al. (2019)]. This high bandwidth requirement is challenging the current wireless network technology.

Another issue in spectating VR sports is VR sickness. Various factors affect the user's VR sickness, such as excessive motion mismatch [Kim et al. (2018)], a wide FOV [Arthur (2000); Lin et al. (2002)], time lag [Geršak et al. (2020)], etc. The excessive motion mismatch between the user's view and physical feeling leads to a high degree of sensory conflicts. In particular, there are a lot of exceeded acceleration or rapid turning actions in sports. These actions enlarge the motion mismatch, which will exacerbate VR sickness.

This work focuses on the problem of key moment misperceiving, which will be described in the following subsection.

1.1.3 Key Moment Misperceiving in VR Sports

The **Key Moment** in sports is a short and critical moment in a sports action, including the information that can influence or determine the action result. For instance, in a boxing punch action, the key moment is the moment that can determine whether this punch hits the target or is dodged. In a basketball ball stealing action, the key moment is the moment that can judge whether this stealing action is foul because of the hand hitting .

Though the key moment is very significant for sports spectating, the conventional visualization technology in VR sports spectating will easily miss the key moment since the temporal and spatial limitations.

Temporal Limitation Many sports, like boxing, basketball, and so on, involve rapid actions. Human vision is one of the most fundamental and significant biological systems for processing information. However, this highly evolved version system still has limited capabilities when perceiving and processing rapid actions in such sports. For instance, the average punch speed of a male amateur boxing player could be reached at 8 meters per second [Kimm and Thiel (2015)]. The naked eye does not have enough time to follow such rapid punch actions and perceive the key moment. In addition, because of the persistence of vision [Hardy (1920)], human visual speed is limited to the rough equivalent of a 25Hz sampling rate. It means there is approximately a 0.3-meter moving offset between two visual samplings. This limited visual sampling usually results in missing the key moment of these actions.

Spatial Limitation In many sports, such as boxing, the players always keep moving and rotating, and their positions and poses are usually dynamic. Hence, the key moment occlusion frequently occurs in such sports. Â However, VR sports spectating allows the spectators to spectate the actions from any free viewpoint. Since there is quantitive information in a free VR environment, the spectators usually feel difficult and confused to find a good viewpoint to perceive the key moment.

These limitations (Figure 1.4) prevent the spectators from perceiving the key moment in VR sports actions and reduce the user experience of VR sports spectating.

1.2 Philosophy

To address these limitations, this work designs and evaluates a temporal and spatial vision augmentation framework for spectating sports matches in VR. There are several challenges in this work, as follows:

- How to help the human vision to process the key moment in rapid actions?
- How to help the human vision to sample the key moment in rapid actions?
- How to guide human vision to find an optimal viewpoint to perceive the key moment from an occluded view?
- How to handle the individual preference in viewpoint selection?

The philosophy of this work is that the VR sports spectating system can improve the spectator's ability to perceive the key moment in VR. This system

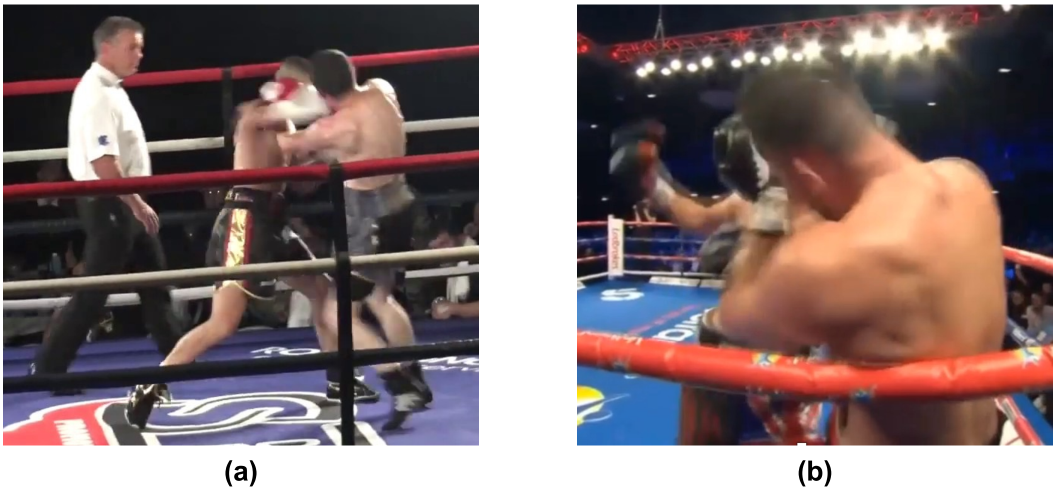


Figure 1.4: Issues when spectating the sports matches in a stadium by the naked eye: (a) The limited naked eye does not have enough time to process the rapid actions resulting in misperceiving. (b) Finding a good viewpoint is required to perceive the key moment in occluded actions.

should be easy to process the key moment in rapid actions. It can allow the spectators interactively Time-control the VR sports videos. In addition, it also can provide a high frame rate recording to sample the key moment. On the other hand, this system should also easily find the key moment in occluded actions. This system can optimize individual preference-based viewpoints to guide the spectators with a good viewpoint.

Like nowadays, when audiences go to watch 3D movies, the movie cinema or theme park will provide 3D glasses. One possible future application case of this work could be VR sports spectating glasses. When sports fans go to the stadium to spectate a sports match live, they can borrow the MR device from the organizers or use the device they own. The spectator can commonly enjoy the live match's atmosphere with the AR mode. In contrast, when crucial actions occur, the spectators can switch the glasses to the VR mode to observe the 3D sports actions replay immersively. This VR sports replay can allow the spectators to control the time in any area and provide an optimal viewpoint depending on the spectators' preference. With this application, the spectators can simultaneously enjoy the excitement of live sports matches and breakthrough temporal and spatial limitations.

1.3 Contribution

This work focuses on resolving the key moment misperceiving that results from the rapid or occluded sports actions and presents interaction methods that allow spectators to spectate the sports match by breaking through temporal and spatial limitations. It can play a role in sports that evolves multiple players, rapid actions, and dynamic moving, for instance, boxing, soccer, basketball, judo, etc.

Two interaction methods (MomentViz and SmartVP) for these two issues are introduced in this work: 1) MomentViz provides a visualization framework that allows the spectators to observe rapid 3D actions from free viewpoints in VR and control the time in any given area. 2) SmartVP can select an optimal viewpoint automatically for watching punching moments in VR that can handle different spectator preferences.

MomentViz: Time-control in 3D for Spectating Rapid Action This work introduces MomentViz, a novel method to control time in any area from the free viewpoint in VR. This method provides a systematic framework including 3D actions data collection, 3D fusion, and time-control interaction. RGB camera and depth camera are utilized for the 3D data collection. After the camera calibration, two types of information can be fused to reconstruct the 3D model. With the stencil technology, time-control interaction is provided for the users. The contributions of this work can be summarized as follows:

1. Design a systematical framework for interactively visualizing the rapid actions in VR;
2. Propose a novel method with the stencil technology to control the time in any area of 3D information;
3. Conduct a pilot to verify the performance of this method, and the results show MomentViz outperforms the other conventional visualization groups.
4. Improve the FPS of data recording by utilizing a high-speed camera and fusing the RGB and depth data.
5. Conduct a user study experiment to evaluate the performance of improved version, and the results verify the new version gets better performance.

SmartVP: Viewpoint Optimization for Individual Preference This work chooses the boxing match as a case, and proposes a novel interactive method that can provide the spectators with an optimal viewpoint to observe the boxing punch actions by their preferences. In this method, a visibility model is designed to account for the visibility of boxers' upper body parts. Then Utilizing the visibility, punch side, and punch offset as features, user-selected preferred candidate viewpoint class as label, a neural network classification model is trained. In summary, the contributions of this work include the following:

1. Design and conduct a experiment to collect the user preferred viewpoints, which considering two different spectator preferences in watching boxing punch video

2. Propose to extract the visibility of body parts, punch side and punch offset as the features;

3. Propose the method to label the continuous user optimal viewpoint selections to discrete candidate viewpoint classes;

4. Design the training model for optimal viewpoint selection training and propose an evaluation method to evaluate the training results for the optimal viewpoint selection problem.

5. Conduct a user study experiment to evaluate the system's performance, and the results show SmartVP can provide optimal viewpoints corresponding with the individual preference.

1.4 Dissertation Outline

This dissertation is structured into six chapters as follows:

Chapter 1 introduces the background of this work, discusses the issue that the rapid or occluded actions will reduce the user experience in sports match spectating, and summarizes the contributions of this work.

Chapter 2 provides a large-scale survey, which includes the visualization of 3D human actions, perceiving the rapid actions, and viewpoint selection for the occluded view. The limitations of existing work are discussed, and the position of this work is defined among the conventional studies.

Chapter 3 introduces MomentViz, a novel visualization method to freely control the time for a 3D rapid action in VR. The framework, technology, and implementation are described in this chapter. A pilot experiment is conducted to verify this visualization method's performance, and the statistical analysis results are discussed.

Chapter 4 introduces an improved version of MomentViz. From the feedback of the pilot experiment described in Chapter 3, the fact is discovered that the low FPS video recording influenced the user experience of spectating the rapid actions. To push this limitation, a high-speed RGB camera replaces the common RGB camera in the improved version. The details of new technology and implementation for this version is presented here. A user study experiment is designed and conducted to evaluate the new version. A discussion of statistical analysis results is presented in the final.

Chapter 5 introduces the SmartVP, a novel viewpoint optimization method that can provide the optimal viewpoint depending on the individual preference when spectating boxing punch actions in VR. This method trains a neural network classification model using the boxer body parts visibility, punch side, and punch offset as the features and the user-selected preferred viewpoint from an experiment as the label. A user study is conducted to evaluate the empirical performance of this method. Both the scientific and empirical evaluation results are discussed here.

Chapter 6 is the conclusion of this dissertation. In this chapter, the contributions and findings of this work are summarized. This chapter also points out the existing limitations and challenges in future work.

Figure (1.5) shows the overview of this dissertation.

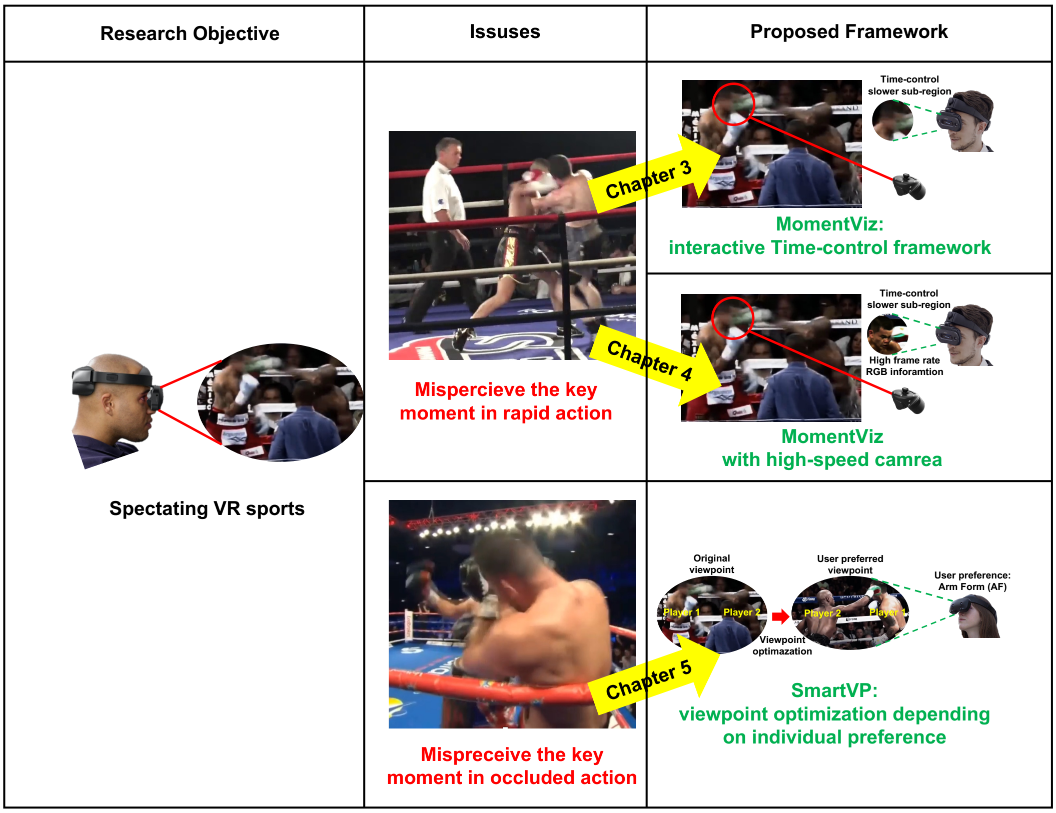

Research Objective
Issuses
Proposed Framework
Spectating VR sports
Mispercieve the key moment in rapid action
Chapter 3
Time-control slower sub-region
MomentViz: interactive Time-control framework
Chapter 4
Time-control slower sub-region
High frame rate RGB inforamtion
MomentViz with high-speed camrea
Mispreceive the key moment in occluded action
Chapter 5
Original viewpoint
Player 1
Player 2
User preferred viewpoint
Player 2
Player 1
Viewpoint optimazation
User preference: Arm Form (AF)
SmartVP: viewpoint optimization depending on individual preference

Figure 1.5: The overview of the dissertation: Chapter 3 and Chapter 4 introduce the MomentViz, which resolves the temporal limitations, and Chapter 5 introduces SmartVP, which deals with the spatial limitation.

Chapter 2

Related Work

Recently, there has been much work on visualizations that assist with understanding human motion [Watanabe et al. (2009)]. In this chapter, we will mainly discuss literature on methods for Time-control 2D visualization and current 3D visualization, viewpoint selection for human actions and 360 degree camera viewpoint adaption.

2.1 Time-control in 2D Visualization

In order to help understand rapid motion, some visualization methods have tried to control the time dimension. Vollmer et al. [Vollmer and Möllmann (2012a)] utilized a slow motion method for visualizing non-compressible liquids and the oscillating droplets. In the same year, slow motion was also used for visualizing gas thermodynamics by Vollmer et al. [Vollmer and Möllmann (2012b)] (Figure 2.1). Another novel attempt of time-control in the real world was proposed by Potthast et al. [Potthast (2013)], who developed a unique helmet that allows the user to perceive the real world in slow motion. Slow motion is also being applied to education nowadays, such as Vollmer et al.'s [Vollmer and Möllmann (2018)] work in physics education.

Despite this recent work on controlling time for improving scene understanding, this kind of motion has yet to be successfully reconstructed into 3D models. As a result, users are not able to observe motion data from free viewpoints.

2.2 Dynamic 3D Visualization

Compared with conventional 2D visualization, 3D visualization provides the user with additional degrees of freedom for viewpoint selection so that the user can gain more information from various angles. Since it can help a user trace complex motions clearly and in a better spatial context, dynamic 3D visualization is especially useful in many fields, like for the Hawk-Eye system used by sports referees. Other significant work on real-time dynamic 3D visualization has been presented in the last several years, such as that of Newcombe et al. [Newcombe et al. (2015)] who succeeded in reconstructing and tracking non-rigid scenes in real-time, using RGB-D scans captured from commodity sensors. Their approach reconstructs the geometry of the scene while simultaneously estimating a volumetric 6D motion field that warps the estimated

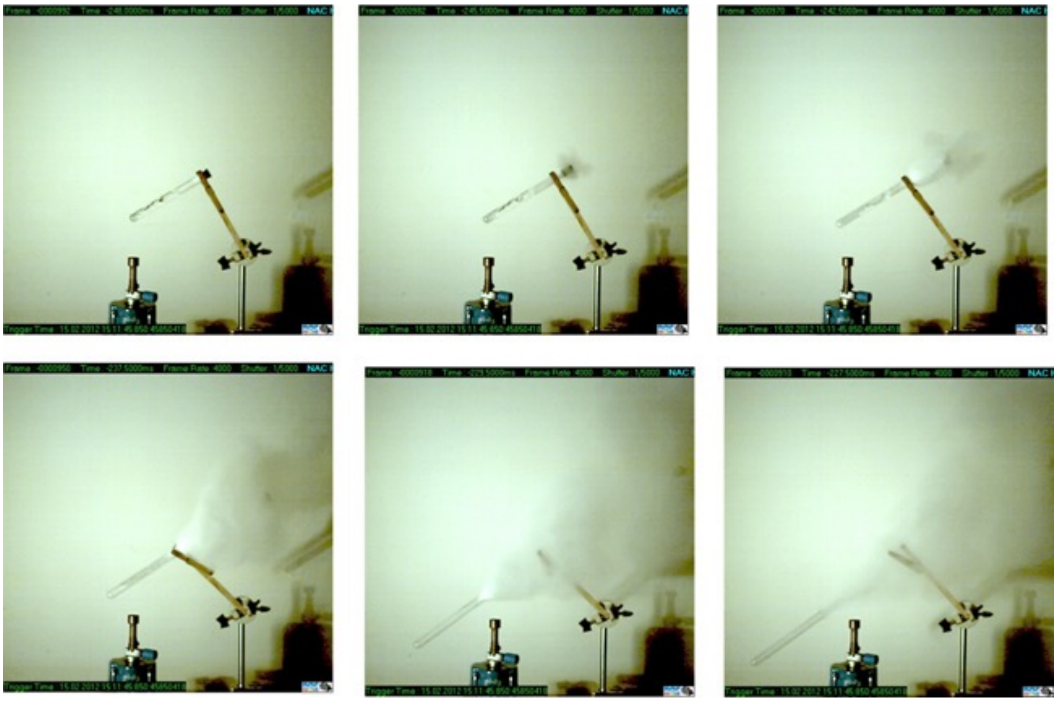
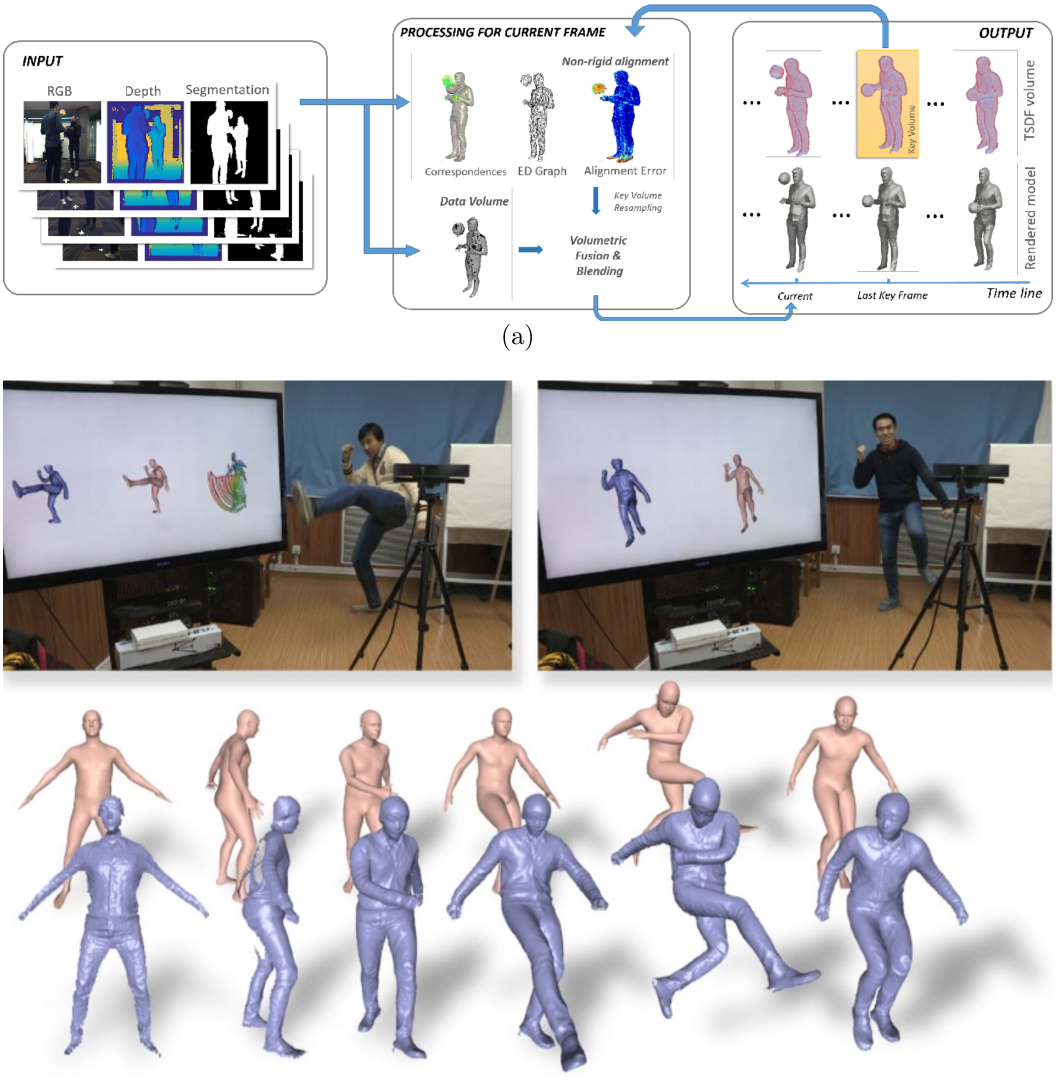


Figure 2.1: Ejection of a rubber stopper from a test tube due to vapour pressure: Sequence of snapshots recorded with 4000 Hz/shutter 1/5000 s at the following times: t = 0, + 2.5 ms, + 5.5 ms, + 10.5 ms, + 18.5 ms, + 20.5 ms [Vollmer and Möllmann (2012b)].

geometry into a live frame. The approach is applicable to a wide range of moving objects and scenes. Another similar study called Fusion4D (Figure 2.2 (a)) was presented by Dou et al. [Dou et al. (2016)]. They proposed a new pipeline for live and multi-view performance capture, generating temporally coherent high-quality reconstructions in real-time. Yu et al. [Yu et al. (2018)] achieved similar results that could reconstruct detailed geometry, non-rigid motion and the inner human body shape in real time from one single depth camera using a double layer representation (Figure 2.2 (b)). Other some similar work [Guo et al. (2017); Innmann et al. (2016); Slavcheva et al. (2017); Yu et al. (2017)] exists for dynamic 3D visualization.

Though these implementations help visualize motion in 3D by providing the user with improved viewpoint selection, none of them explicitly involve time-control to help visualize or improve the understanding of rapid motion. These previous methods motivated us to address the gaps in time-control research by incorporating the 3D information combined with an interactive component.



(a)

(b)

Figure 2.2: Two research about 3D human reconstruction: (a) Real-time Performance Capture of Challenging Scenes [Dou et al. (2016)]; (b) Real-time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor [Yu et al. (2018)].

2.3 Viewpoint Selection for Human Actions

Since human action is complex and unpredictable [Assa et al. (2008); Goldman (2015)], tracing human action from a good viewpoint is quite difficult for a spectator. To help humans to observe humans actions more clearly and easily, several attempts tried to solve this problem [Doubek (2005)]. Compare with the traditional viewpoint selection research which targets general objects, the one that targets human action started late. Nonetheless, some work flourished in the current 10 years.

Dmitry Rudoy and Lihi Zelnik-Manor [Rudoy and Zelnik-Manor (2010)] utilized a method that evaluated the three-dimensional shapes of human action induced by their silhouettes in the space-time volume. The experiment results presented their method provides intuitive results which match common conventions. Chong Huang et al. [Huang et al. (2018)] also achieved viewpoint selection research about an autonomous drone cinematography system for human action. In their research, their used 3D skeleton data to design a real-time dynamical camera planning strategy (Figure (2.3)). Another similar work using the drone was published by Sena Kiciroglu et al. [Kiciroglu et al. (2020)]. They proposed an algorithm that the key idea is estimating the uncertainty of the 3D body pose. One more new work was presented by Mojtaba Ahangar Arzati and Siamak Arzanpour [Arzati and Arzanpour (2021)]. In their research, they came up with an RL-based method. The camera viewpoint was identified as the key parameter in the accuracy of monocular 3D human pose estimation. The results improved the performance of 3D human pose estimation. In Wei Cheng's work on human body 3D reconstruction [Cheng et al. (2018)], they also determined the real-time next-best-view selection by information gain calculation in Truncated Signed Distance Function(TSDF) volume.

Though the previous work can help to adapt a viewpoint for human action autonomously, their work is still limited to simple and slow human actions. Moreover, all of them have not considered the spectator desire and preference about viewpoint.

2.4 360 Degree Camera Viewpoint Adaption

The 360 degree camera has a field of view that covers approximately the entire sphere or at least a full circle in the horizontal plane. Recently, it is very popular among all generations, and in a lot of fields, especially in sports. A lot of online video sharing sites like Youtube are already supporting the 360 degree camera videos.

The 360 degree camera video can provide more information including whole the 360 degree field of view of a scene at one moment. But for so large information of a scene, it's difficult to find a good viewpoint to observe for the

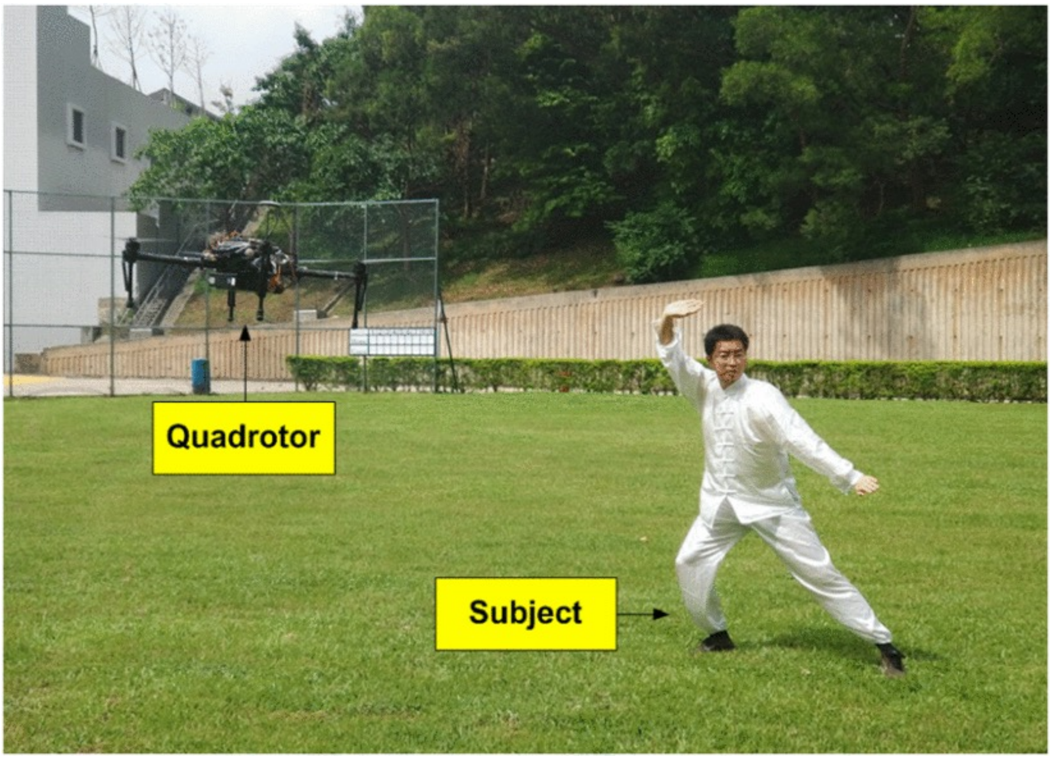


Figure 2.3: Viewpoint selection for human actions by using a drone [Huang et al. (2018)].

user. Hou-Ning Hu et al. [Hu et al. (2017)] got a successful attempt on the 360 degree sports video piloting by utilizing a deep learning-based agent. The experiment results reveal that their method is better than the conventional method (Figure 2.4). But as their work just explored the viewpoint selection in 360 degree camera video, no 3D reconstruction means some information will be lost for a sports scene, which can not be caught by a single camera.

Another popular topic in 360 degree camera viewpoint adaption is viewport prediction. The goal of this topic is to resolve the contradiction between the limitation of Internet bandwidth and computational resources and the entire high resolution 360 degree camera videos. Tile-based and user behavior learning-based are the two most popular types of methods. Lan Xie et al. [Xie et al. (2018)] proposed a Cross-user Learning based system to improve the precision of viewport prediction. According to their research, the users have similar region-of-interest when watching the same video, so they came up with the idea of exploiting cross-users' ROI behavior to predict viewport. Their results showed the Cross-user Learning based system outperformed the other tile-based methods. Another similar work was published by Yucheng Zhu et al. [Zhu et al. (2021)]. They proposed a traditional method to adapt 2D saliency models and design a CNN-based model to better predict visual saliency. Their experiment results also showed effective performance.

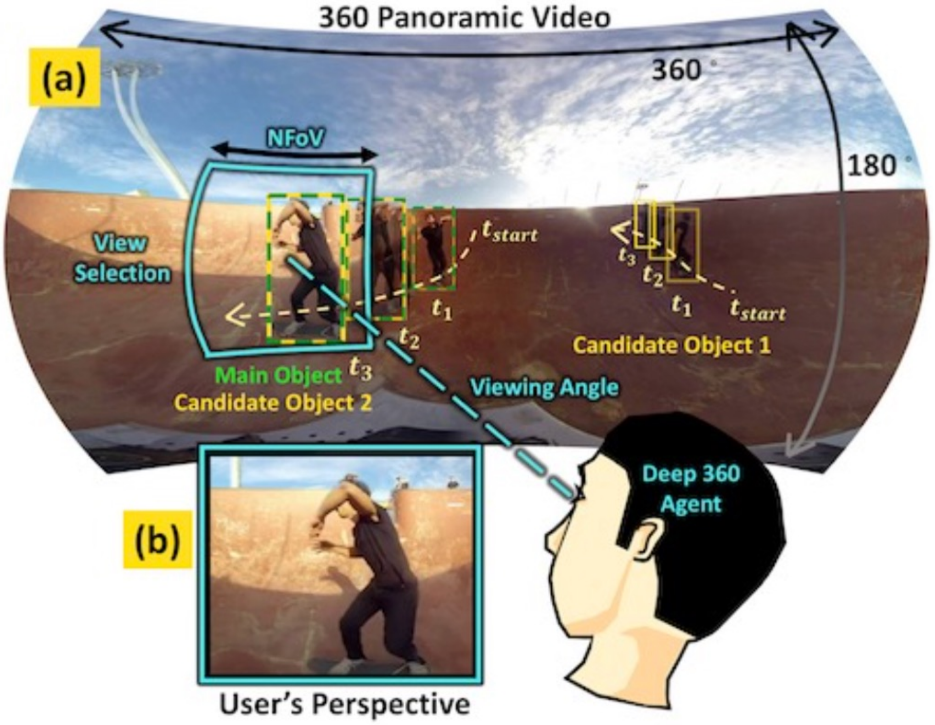


Figure 2.4: An attempt of viewpoint selection for 360 degree sports videos [Hu et al. (2017)].

One more work attributed by Yuanxing Zhang et al. [Zhang et al. (2019)] presented a deep reinforcement learning based framework, which included any possible feature, given any quality of experience (QoE) objective. They also got a good result about providing high QoE and adapting many real-world playback conditions. Yanan Bao et al. [Bao et al. (2016)] devoted to another interesting work which is based on the prediction of the user head motion. The simulation result that reduced the bandwidth consumption by 45% proved good performance. In addition, a lot of other work [Corbillon et al. (2018); Fu et al. (2020); Zhao et al. (2019)] contribute to this topic.

However, all the work introduced above showed satisfactory results. Because of the different goals, their work just explored the viewpoint adaption in 360 degree camera video, no 3D reconstruction means some information located at the back of the viewpoint will be lost for a boxing punch scene.

2.5 Chapter Summary

Several attempts have evolved the Time-control in 2D, and high-speed cameras are utilized in work. However, their attempts ignore the 3D information, which is very important in a sports match. In addition, some other pioneers tried several new methods to visualize the 3D information. Nevertheless, these methods mainly focus on 3D visualization. No Time-control means these methods are difficult to apply to rapid sports videos. Some previous work tries to select the optimal viewpoint for human actions. Nevertheless, this work focus on simple scenes and videos with a single person. It is impossible to be utilized in a multiple sports match with complex human actions. It also ignores to handle individual preferences of spectators. Moreover, viewpoint adapting in 360 degree video also already gets some promising achievements. However, 360 degree video only covers the information from one 360 degree camera. No 3D reconstruction means this information can not cover the whole sports match area.

This work covers the entire 3D information of sports matches by 3D reconstruction, provides the Time-control to help perceive the rapid actions, and suggests the optimal viewpoint depending on individual preference to prevent the view occlusion.

Chapter 3
Interactive Visualization for Rapid Actions

3.1 Introduction

Spectating sports in VR is fun and has a high potential to break through the temporal and spatial limitations for the spectators. However, even with this cutting-edge and fashionable technology, perceiving rapid sports action is a challenging task with conventional visualization methods.

To address this issue, this work proposed a novel visualization framework called MomentViz for spectating rapid sports actions in VR. This work utilizes one RGBD camera to record the 3D human action data and uses commercial software to visualize the 3D data in VR with point cloud style. A novel method is proposed in this work to allow the users to control the sports action video time in a user-selected sub-region area. The implementation details are also introduced in is chapter. In order to evaluate the performance of this method, a pilot study that observes rapid actions in VR is conducted. Eight participants are recruited to take part in the pilot. Four groups, including 2D original time videos, 2D time-control videos, 3D original time videos, and 3D time-control videos (MomentViz), are tested by the participants. The pilot study evaluates the system performance from accuracy, the number of views, task completion time, and subjective questionnaires. The pilot study results verify that MomentViz outperforms the conventional visualization methods in spectating rapid actions in VR. The main contributions of this work can be summarized as follows:

- Propose a novel framework for interactively visualizing rapid actions in VR;
- Implement the system with the proposed framework, and conduct a pilot study to test the performance of the framework,
- From the results, it is found that the proposed framework outperforms other conventional visualizations.

The remainder of this chapter presents the framework details, the pilot design, the evaluations, and the findings from the results.

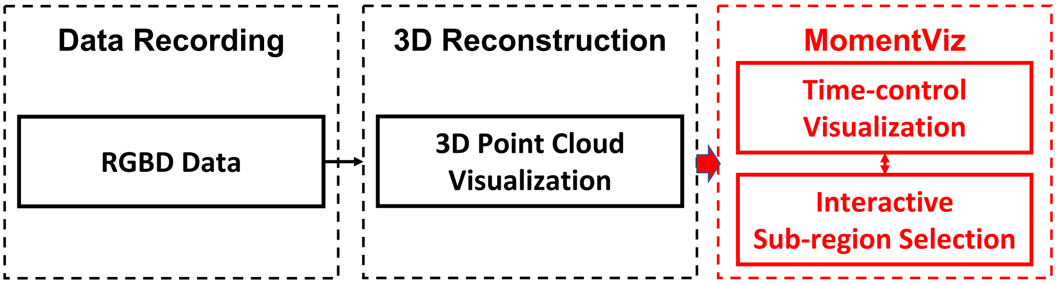

Data Recording
RGBD Data
3D Reconstruction
3D Point Cloud Visualization
MomentViz
Time-control Visualization
Interactive Sub-region Selection

Figure 3.1: The framework of MomentViz.

3.2 Methodology

This work proposed a novel visualization method called MomentViz, which allows the user to control time speed in any area from the free viewpoint in VR. Figure 3.1 presents the system framework of MomentViz. This work mainly focuses on the red dashed box, which includes the Time-control method and interaction function.

3.2.1 Data Preparation

Since 3D reconstruction is not the topic of this work, some know-how technology is utilized directly in this work.

The first step of 3D visualization is 3D motion data recording. Various types of methods can capture human motion credibly and effectively [Van der Kruk and Reijne (2018)], which includes an optical system with markers [Kirk et al. (2004); Lee and Yoo (2017); Meyer et al. (2014); Miyata et al. (2004)], traditional RGB camera [Guerra-Filho (2005); Nakano et al. (2020); Hasler et al. (2009a); Rosenhahn et al. (2006)], RGBD camera [Han et al. (2013); Napoli et al. (2017)], mechanical motion capture [Brigante et al. (2011); Liu et al. (2020); Tognetti et al. (2014)], magnetic systems [O'Brien et al. (1999); Fourati et al. (2012)], and so on. In order to prevent the markers from influencing the players' actions and implement the system easily in the laboratory, the method using a commercial RGB-D camera is chosen in this work. The next step is 3D fusing the RGB-D data from the RGB-D camera. Though it is a challenge to reconstruct the 3D human model, there are several studies attempted in current years, some of them are already introduced in Chapter 2 [Dou et al. (2016); Newcombe et al. (2015); Bhatnagar et al. (2020)]. However, considering the time cost and main topic, commercial software for 3D visualizing with point cloud is utilized in this work.

3.2.2 Time-control Visualization

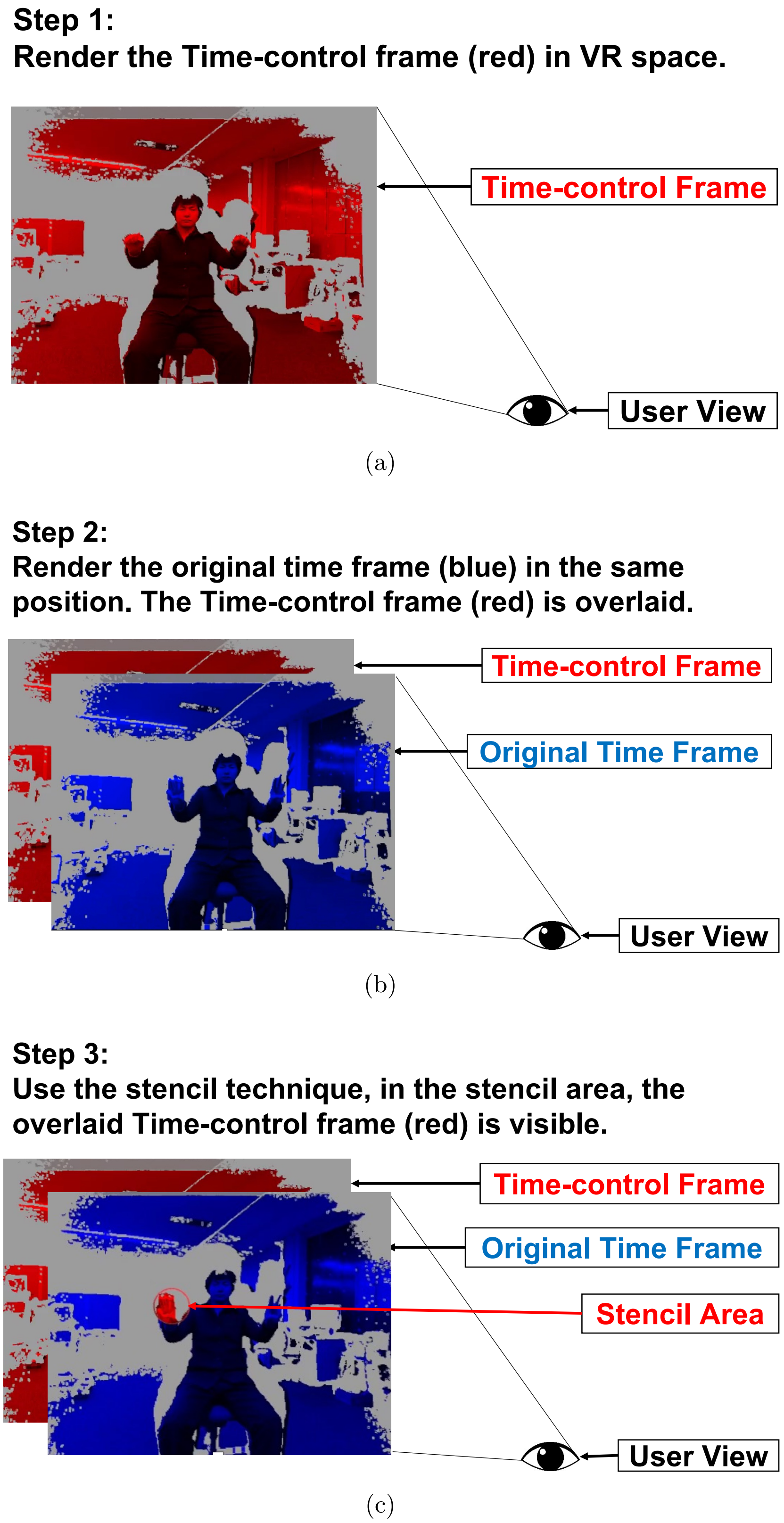
Inder order to address the issue that the naked human eye is difficult to process rapid actions, a novel method is proposed in this work with an interaction paradigm that allows the user to control a specific area of the 3D action videos. This method of Time-control utilizes two parallel 3D action frames, which include both the original time frames and the slow time frames. When a user selects the area to be time-controlled, a circular stencil provides a window into the time-controlled slow time frames, as outlined in Figure 3.2. The detailed process of this method is presented as follows:

1. The first step: render the time-control slow time frames in the VR. The rate of slow time can be set in advance by the users. In order to distinguish the different time frames, the slow time frames are rendered in red in the Figure (Figure 3.2(a));

2. The second step: render the original time frames at the same position as the slow time frames. Since the rendering order, the slow time frames will be overlaid by the original time frames and become to be invisible. In the Figure, the original time is rendered in blue (Figure 3.2(b));

3. The third step: the users can choose the area by the input devices where they want to focus. In this sub-region area the user-selected, the stencil technology is utilized. With this technology, the rendering order is reversed in this particular area, and the slow time frames appear from the back. In the Figure, the person shows the same finger gesture from 0 to 5 in two hands simultaneously. In the original time frame, the left hand shows three fingers up (the blue frame), but in the sub-region of the slow time frame, the right hand still holds a fist (Figure 3.2(c)).

As the result of showing both the original speed stream and partial slow speed stream simultaneously, this method allows a user to view the desired event in slow motion and be aware of the surrounding environment in real time. This solution avoids potential peripheral occlusion of ongoing events, which will be especially important for use at outdoor sporting events in the future. Moreover, a user can change the scale and rate of the Time-control to adjust to any dynamic scene.



Step 1:
Render the Time-control frame (red) in VR space.

(a)

Step 2:
Render the original time frame (blue) in the same position. The Time-control frame (red) is overlaid.

(b)

Step 3:
Use the stencil technique, in the stencil area, the overlaid Time-control frame (red) is visible.

(c)

Figure 3.2: The process of the Time-control interaction method: (a) First step, set the Time-control frame; (b) Second, set the original time frame on the same position, then the Time-control frame will be overlaid by the original time frame; (c) Final step, use the stencil technique to control the time in sub-region interactively.

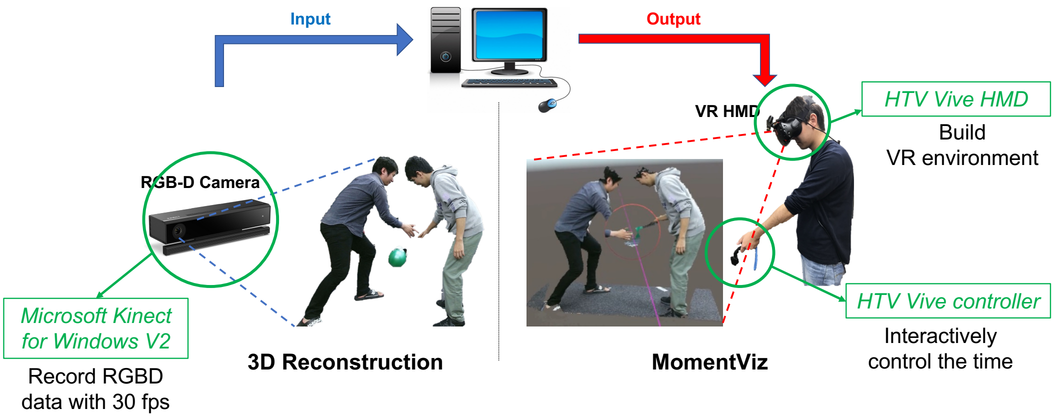


Figure 3.3: The implementation of MomentViz: left is the data preparation process, using a standard RGBD camera for 3D data recording (with 30 fps); right is the MomentViz visualization process, using VR HMD for visualizing the 3D data in VR environment, and using a controller for interactively selecting Time-control sub-region area.

3.3 System Implementation

This section describes the implementation of the MomentViz visualization system, including the system overall, hardware, and software utilized in the implementation.

3.3.1 System Overview

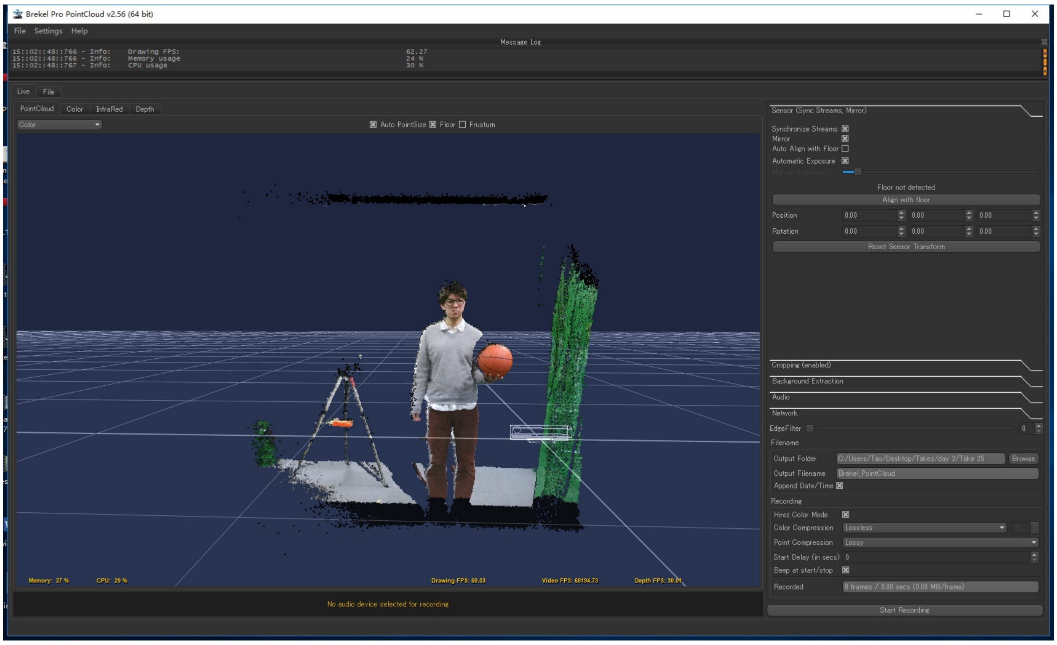
According to the framework of MomentViz, described in the previous section, this system can be divided into two parts, data preparation and MomentViz visualization, which corresponds to the input and output (Figure 3.3).

In the data preparation part, a standard RGBD camera is utilized for recording the 3D data. By using commercial software, the obtained 3D data can be visualized in 3D point cloud style in VR via a VR device. In the MomentViz visualization part, by using the proposed visualization method, the users can choose the sub-region area for controlling the frame time by one controller of the VR device.

3.3.2 Hardware

At this stage, as this work focuses on indoor sports like basketball, boxing, and so on, an RGB-D is chosen for getting the 3D data because of the less cost and lower computer specs requirement.

One Microsoft Kinect for Windows V2 is utilized for recording 3D data in this work. This RGBD camera can run at 30 FPS, 1920 × 1080 resolution color, and 30 FPS, 512 × 424 resolution depth. As the depth range of Kinect



```
15::02::48::766 - Info:    Drawing FPS:          62.27
15::02::48::766 - Info:    Memory usage          24 %
15::02::48::767 - Info:    CPU usage             30 %
```

☒ Auto PointSize ☒ Floor ☐ Frustum

Synchronize Streams ☒
Mirror ☒
Auto Align with Floor ☐
Automatic Exposure ☒
Manual Exposure

Floor not detected

Position 0.00 0.00 0.00
Rotation 0.00 0.00 0.00

EdgeFilter 0

Output Folder C:/Users/Tao/Desktop/Takes/day 2/Take 25
Output Filename Brekel_PointCloud
Append Date/Time ☒

Hirez Color Mode ☒
Color Compression Lossless
Point Compression Lossy
Start Delay (in secs) 0
Beep at start/stop ☒
Recorded 0 frames / 0.00 secs (0.00 MB/frame)

Memory: 27 % CPU: 29 % Drawing FPS: 60.03 Video FPS: 60194.73 Depth FPS: 30.01

No audio device selected for recording

Figure 3.4: Using software Brekel Pro PointClound v2 for recording and visualizing 3D data: one Kinect V2 is utilized for recording the 3D data, and Brekel can save the 3D data as several file formats in real time.

V2 is from 0.5m to 8.0m, multiple cameras placed around an indoor sports arena can cover the whole area (Table 3.1). For instance, four cameras can almost cover an entire boxing ring (about 6×6m). Aligning multiple cameras is not the focus of this work, and much research exists on camera alignment [Córdova-Esparza et al. (2016); Kim et al. (2017); Yang et al. (2013)]. This work utilizes one set consisting of an RGBD camera to catch the 3D data and test the performance of this visualization framework. The system is developed with Unity and the HTC Vive Plugin, and the computer specs are an Intel Core i7-6700K 4.00GHz CPU, NVIDIA GeForce GTX 1080 GPU, 32G RAM, and Windows 10 Enterprise Operation System (Table 3.2)

3.3.3 Software

This work utilizes Brekel Pro PointClound v2 (BPC) [1] to record and visualize the depth information. Brekel provides an integrated motion capture system that can record videos and save the 3D data in real time by using SDK from Kinect SDK and OpenNI and NITE. This software can record the 3D dynamic scenes by using Kinect V2 and output the recorded information with a variety of formats, including Alembic, OBJ sequence, PDB and PDC for Maya, Brekel Real-time File (BRF) for Unity, etc. (Figure 3.4). In this system, the recorded 3D data is output in the BRF format, which consists of

[1] https://brekel.com/pointcloud_v2/

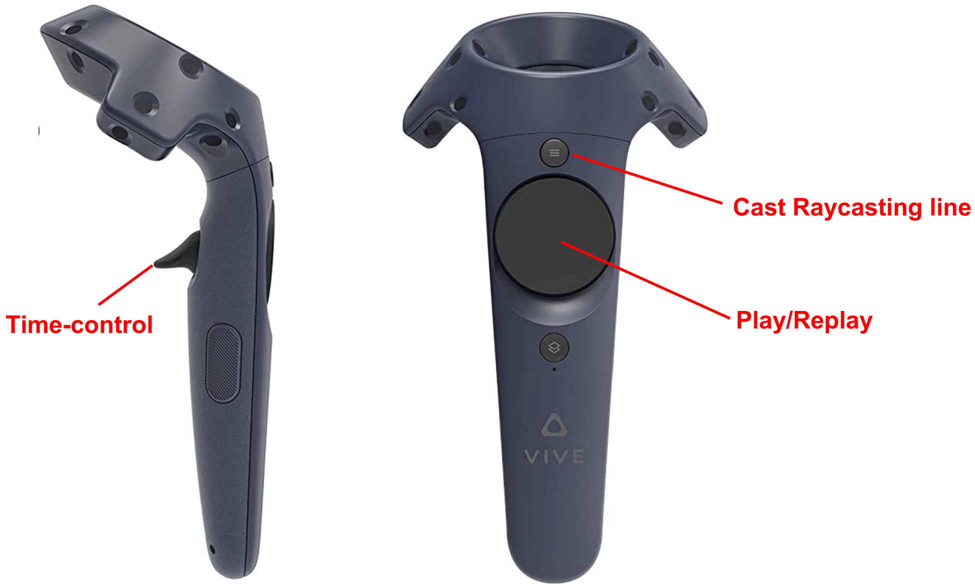


Figure 3.5: The HTC Vive controller setting for interaction: the Touchpad is set for pltheay or replay the videos; Menu button is set for casting a Raycasting line; the Trigger is set for the time-control.

the point cloud information, start time, etc., of every frame. Brekel provides the plugins for Unity. With these plugins, BPC can easily input the BRF file, which includes the depth information, into Unity, and then render the 3D dynamic scenes with Shader. Because of this reason, this work uses Unity as the primary tool for VR development.

3.3.4 VR Setting

HTC VIVE [2] is chosen as the VR device in this work since it can be simply developed in Unity. A space with 320mm×250mm is set for the VR environment area. The users can immerse themselves in the VR environment in this area and utilize one HTC Vive controller to interact with the MomentViz visualization system. The users can press the Touchpad button to start or restart videos. When the videos start to play, the controller generates a raycasting line immediately. The users can move the raycasting line to choose the area they want to control the time. When they determine the expected area, they can keep pressing the trigger button to start the time-control, while releasing the trigger will stop it (Figure 3.5).

[2] https://www.vive.com/

Table 3.1: Specifications of Kinect for Windows V2

Item	Detail
Name	Kinect for Windows V2
Color Resolution	1920×1080
Color FPS	30 fps
Depth Resolution	1920×1080
Depth FPS	512×424
Range of Depth	0.5 8.0m
Range of Detection	0.5 4.5m
Depth Angle (Horizontal)	70°
Depth Angle (Vertical)	60°

Table 3.2: Specifications of computer in implementation

Item	Detail
CPU	Intel(R) Core(TM) i7-6700K 4.00GHz
GPU	NVIDIA GeForce GTX 1080
RAM	32GB
OS	Windows 10 Enterprise

3.4 Pilot Study

In order to determine the necessity of Time-control and verify the performance of Time-control in a 3D VR environment, a pilot study is conducted.

3.4.1 Experiment Setup

A total of 8 participants are recruited with an age range from 23 to 27. The task is to watch a quick typing scene and answer the sequence of the letters as quickly and correctly as possible (Fig.3.6). One Kinect V2 records all the scenes, and the distance between the camera and the typist is about 0.5m. All the character keys are stuck with the key name on white background for easier recognition.

Every participant performs four visualization methods, including 2D, 3D, Time-control 2D, and Time-control 3D. To reduce the influence of the experiment order, the participants are divided into two groups, and every group includes 4 participants. One group starts the experiment with 2D visualizations, while the other group tests the 2D visualizations following the 3D visualizations.

The participants are asked to observe six videos for every method. In every video, there is a letter sequence typing of 6 letters in random order. Every video lasts for approximately 5 seconds. The participants are allowed to replay the videos if they need. To control the experiment time, the number of total views is limited to 5 per video. After watching each video, participants are then asked to answer the typing order orally. The answered letter sequences, task completion time, and the number of total views are recorded. The participants are asked to fill out a user experience questionnaire after finishing all the videos.

In the 2D visualizations, participants are asked to finish the task on a monitor. The keyboard and mouse are utilized for operating the experiment. For the details, the key "R" is set to start or replay the scene. In the 2D Time-control visualization, the mouse is set to move the time-control area. If the mouse keeps still for more than 1 second, slower time-control will start automatically, and when the mouse starts to move, the time-control will end. The 3D visualizations are implemented in VR by HTC Vive. In this VR environment, they can move and observe the 3D videos from any preferred viewpoint. An HTC Vive controller is utilized for experimenting. The touchpad is set for starting or replaying the videos. In the Time-control 3D visualization, the menu button is set to cast a raycasting line, and the trigger is set for the time-control. The rate of Time-control is five times slower in this experiment.

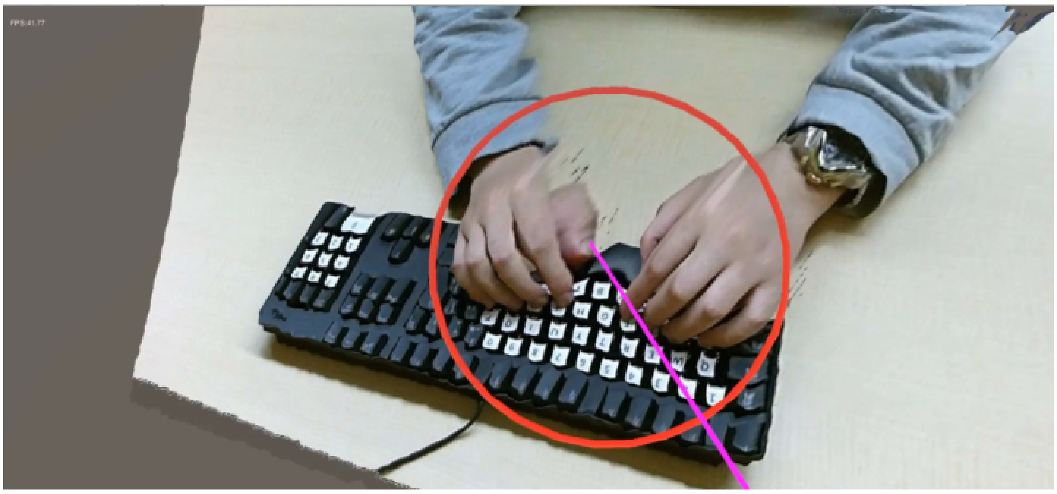


Figure 3.6: A pilot study experiment of observing rapid typing using Time-control 3D.

3.4.2 Experiment Results

The accuracy, the number of total views, and the completion time of all the methods are analyzed for objective evaluation. After tasks, a subjective evaluation of the user experience questionnaire is also conducted. The experiment results are shown as follows:

Accuracy As shown in Fig.3.7(a), Time-control 3D had the highest average accuracy, with 4.854 correctly recognized letters (the maximum number is 6). The analysis of variance (ANOVA) is utilized to compare the four methods [Tabachnick and Fidell (2007)]. The Holm method [Holm (1979)] is chosen for p-value adjustment to evaluate the results.

This analysis finds significant differences between the Time-control 3D and 3D with $p < 0.001$. It also finds significant differences between Time-control 3D and Time-control 2D with $p < 0.01$. Cohen's d [Rice and Harris (2005)] is calculated as effect size by using the means and standard deviations of the two groups. Cohen's d is 0.51 between the Time-control 2D and Time-control 3D. Moreover, between the 2D and Time-control 3D, Cohen's d is 2.24, and between the 3D and Time-control 3D, Cohen's d is 3.08.

Number of Views In order to control the total experiment time, the maximum views are limited to a total of 5. Then the average number of total views is calculated for every method. Similar to the accuracy analysis, ANOVA is also used here to compare the four methods, and the Holm method is used for p-value adjustment to evaluate the results. The results are shown in Figure 3.7(b).

From the results, it can be known that in 2D and 3D visualizations without time-control, every participant uses the maximum available views to observe the quick typing in all videos. In the Time-control visualizations, the 3D has a mean of 3.583 views, which is lower than the 2D view average of 4.146. There is a significant difference between Time-control 3D and each of the other methods. All the p-values are less than 0.001.

Completion Time The task completion time of every video is recorded. This completion time is defined as the start of the first view of the video to the time a participant answers the letter sequence. After calculating the average value for each method, ANOVA is utilized to analyze the values. The results are shown in Figure 3.7(c). As can be seen in this graph, there is a significant difference between 2D and Time-control 3D with $p < 0.001$. At the same time, a significant difference is also found between 3D and Time-control H3D with $p < 0.001$. Another finding is that the average completion time values of both Time-control methods are approximately twice as much as the other methods. Since the Time-control rate is five times slower, the double completion time is considered an expected result.

Questionnaire All participants are asked to complete a questionnaire after the experiment. This questionnaire consists of 16 questions, including each participant's basic information, HMD experience, and subjective experience of this framework. Most participants have experience using HMD. The result of the question "Would you like to use the 3D Time-control visualization for this task" shows in Figure 3.8 (a). From the results, 75% of the participants want to use the 3D Time-control visualization, including 37.5% strongly agree and 37.5% agree. For the task "Rank the four visualizations by overall preference", the results are shown in Figure 3.8 (b). More than half of the participants (5 participants) consider the 3D Time-control visualization the best of the four visualizations. Furthermore, 2 participants rank the 3D Time-control visualization in second place, and no participant rank it in last place.

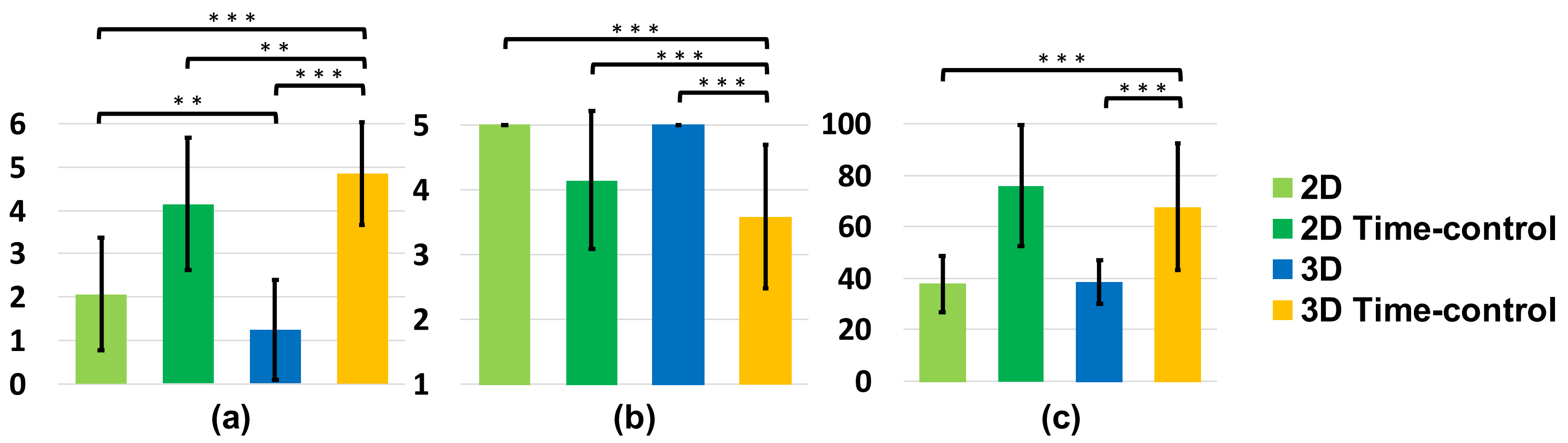
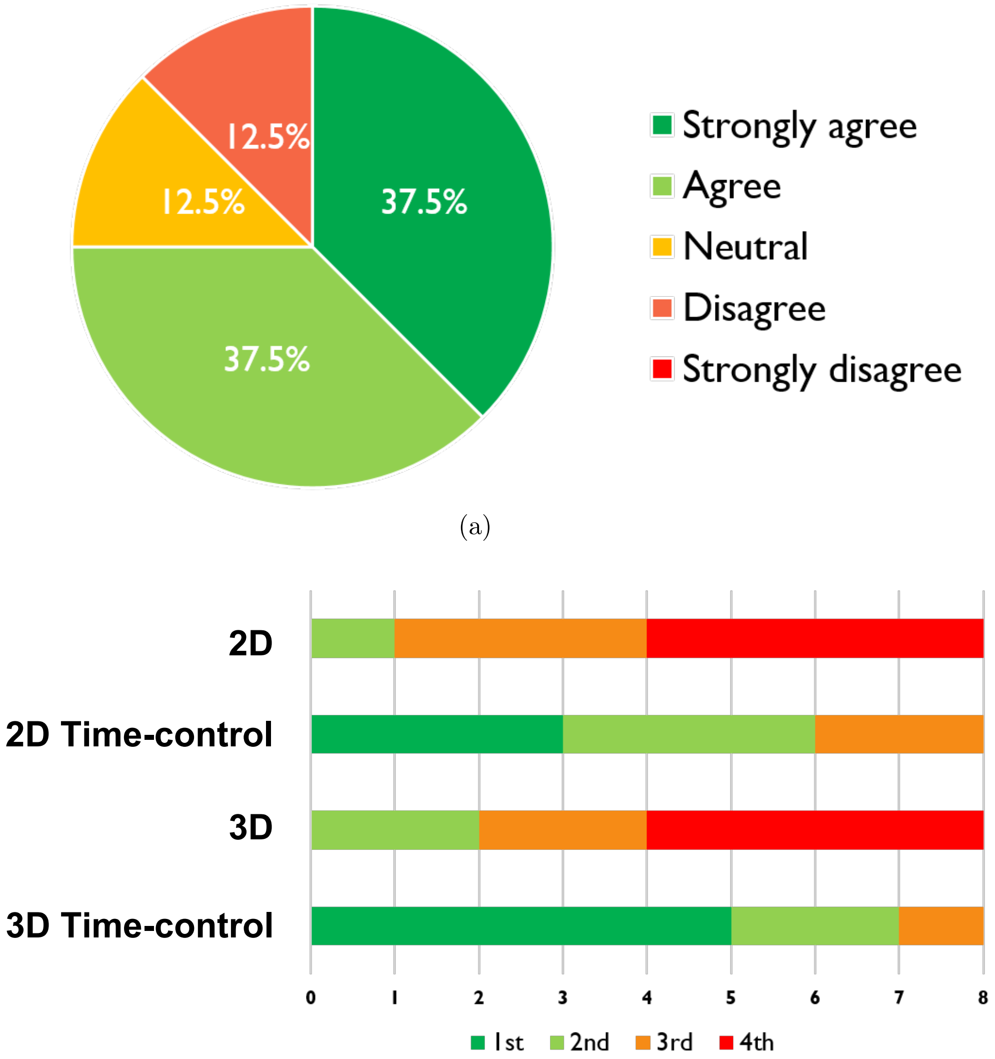


Figure 3.7: The results of pilot experiment: (a) The accuracy of correctly recognized letters. (F = 82.05, df = 3) (b) Number of views required to determine the letter sequence for each method, limited to 5 total views.(F = 38.99, df = 3) (c) Completion time required to determine the letter sequence for each method.(F = 55.38, df = 3) (***: $p < 0.001$, **: $p < 0.01$)



(a)

(b)

Figure 3.8: The results of questionnaire in pilot experiment: (a) The results of the question "Would you like to use the 3D Time-control visualization for this task?"; (b) The results of task "Rank the four visualizations by overall performance".

3.4.3 Discussion

The experiment results found that in non-time-control visualizations, the accuracy of 3D is lower than that of 2D. One potential reason for this is that for a participant observing a rapid scene, it is difficult to find the best viewpoint to observe a target from such a massive amount of information in the free VR environment, which requires more time in the 3D environment.

Furthermore, it is found that the accuracy of Time-control 3D is higher than 2D Time-control. It could be interpreted that in 3D Time-control, the participant has enough time to find a better viewpoint to observe the target. For this experiment task, the participant can view from the same point as the camera viewpoint to observe the position of the pressed key clearly and use the reverse direction viewpoint to understand the key label easily.

Finally, the feedback from the participants noted the difficulty in finding a viewpoint in the 3D VR environment and unclear detail of motion due to the low frame rate of the RGB-D sensor.

From the pilot study, it is realized that it is necessary to utilize Time-control in the 3D environment of the difficulty of the best viewpoint finding. However, it is also doubtful that having a Time-control of low FPS information in the 3D environment is still insufficient.

3.5 Chapter Conclusion

In this work, propose a novel 3D Time-control visualization framework called MomentViz is proposed. This framework can augment the vision for rapid actions by an interactive time-control visualization method. This method renders two different time frames at the same position, and the original time frames overlay the slower time frames. With the technology of stencil, the slower time frame can be revealed in the user-selected sub-region area. To evaluate the performance of this framework, a pilot experiment of spectating rapid typing actions is conducted. Eight participants are recruited to join the pilot experiment. This framework is evaluated objectively and subjectively by analyzing the accuracy, the number of views, the completion time, and the feedback of the user questionnaire. The results show that this framework outperforms the other conventional visualizations. The findings from the feedback motivate the author to improve the 3D Time-control visualization by using high frame rates data.

Chapter 4

Interactive Visualization with High FPS for Rapid Motion

4.1 Introduction

Through the pilot experiment, the performance of MomentViz is verified. Using Time-control visualization, the users can get a better experience in spectating rapid actions. It handles the issue that naked human eyes can not process rapid actions. However, the limitations of the naked human eyes for spectating rapid actions have not been resolved entirely. From the feedback of the pilot questionnaire, a fact is revealed that the low frame rate standard RGBD cameras also have a higher possibility to miss sampling the crucial action moment. The previous framework of MomentViz can not solve the problem that the naked eye will misperceive the crucial moment of rapid action.

In order to address this remainder issue, an improved framework of MomentViz is proposed here. In the improved version, a high-speed camera is utilized instead of the standard RGBD camera to record the high frame rate RGB frame data. After camera calibration, coordinates unifying, and 3D fusion, the 3D rapid action frames with high frame rate RGB and standard depth data are reconstructed. The visualization, which allows the users to control the frame time in a sub-region area, is also applied in this improved framework of MomentViz. The implementation and a user study evaluation experiment are conducted to evaluate the performance of this improved MomentViz. Twelve participants are recruited to join the experiment, and all of them tested the three groups: the conventional 3D visualization (**3D**), the previous framework of MomentViz (**Time-control 3D**), and this improved MomentViz using a high-speed camera (**Time-control H3D**). The task of this user study is to spectate a rapid action of stealing the basketball in defense and judge whether this stealing is foul. The aspects of the D prime value of judgments, the number of views, task completion time, and subjective questionnaire are analyzed to evaluate the performance. The results show that the improved MomentViz performs best among the three groups. This new framework of MomentViz using a high-speed camera can increase the user experience from the previous framework. The main contributions of this part can be summarized as follows:

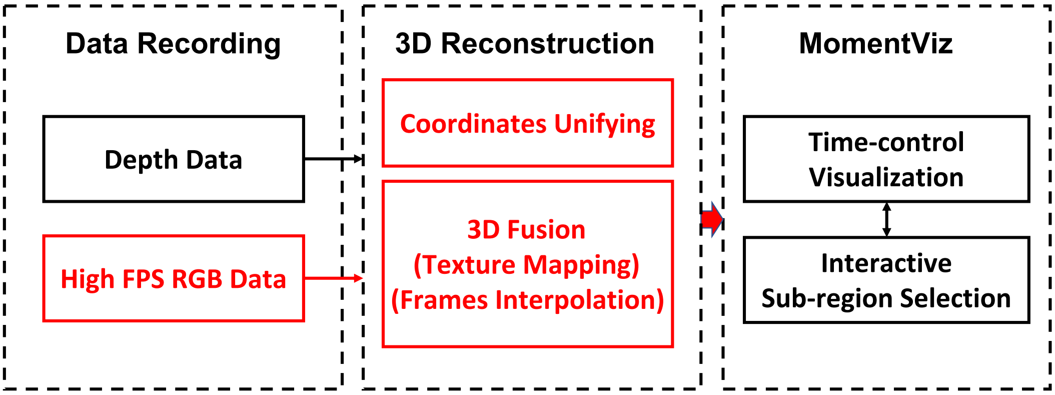


Figure 4.1: The framework of improved MomentViz: red box presents the new changes from the original version.

- propose an improved framework of MomentViz by using a high-speed camera;
- conduct a user study evaluation experiment to test the performance of improved MomentViz;
- The results verify that the improved MomentViz could increase the user experience from the previous framework of MomentViz.

The details of this improved framework, the user study design, the result analysis, and the findings are presented in the remainder of this chapter.

4.2 Methodology

From the results of a pilot study, it is found that the low sampling rate of standard RGBD cameras has a high probability of missing recording the crucial moment since the short moment. According to these findings, an improved version of MomentViz is proposed in this work. The high-level overview of the improved MomentViz system is shown in Figure 4.1. The red box shows the improved parts from the previous version. In order to resolve the issue that the crucial moment action frame may miss sampling in a low FPS recording, instead of the standard FPS RGBD camera, this improved version of MomentViz utilized a high-speed RGB camera to catch the crucial frame. With the calibration of two kinds of cameras, the RGB data from the high-speed camera and the depth data from the standard RGBD camera can be aligned, and the transform matrix can be calculated. After that, the different coordinates data can be unified with the transform matrix. Finally, map the two aligned data for 3D fusion.

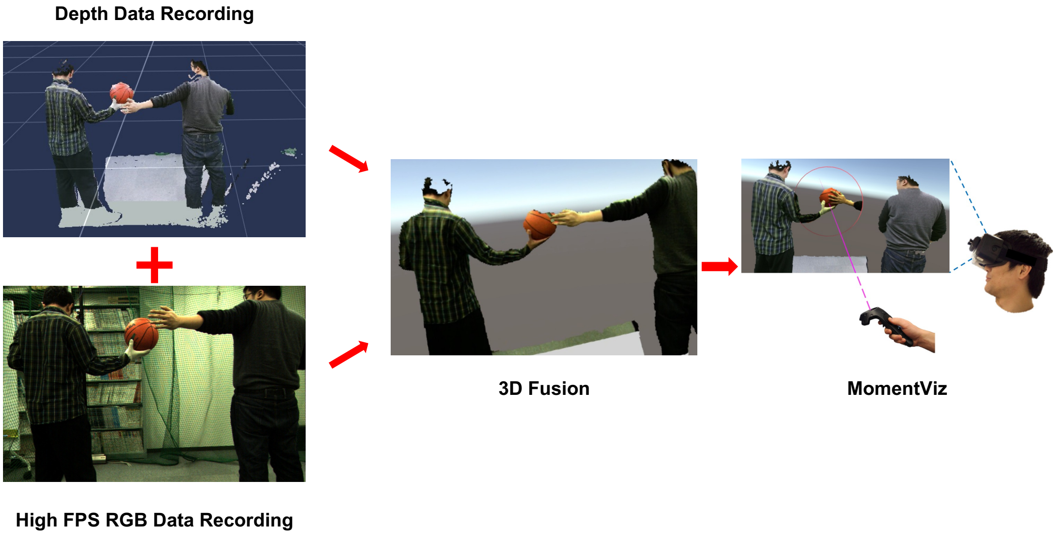


Figure 4.2: Images outlining the process of the improved MomentViz, including the recording and import of depth (RGB-D) data, recording and calibration of high FPS camera frames, mapping and fusion of high FPS frames to 3D data, and interactive control of the user's sub-region of interest.

4.2.1 Coordinates Unifying

Since two kinds of cameras (RGB-D camera and High-Speed camera) are utilized in this work, every camera and the VR environment have different coordinates. Calibrating the two cameras and unifying the coordinates is required. Figure 4.3 shows the different coordinates of this framework. The details of the coordinates unifying step can be presented as follows:

1. Using the standard RGB-D camera to record the rapid action depth data. With commercial software, the recorded depth data can be saved and exported to the point cloud format, including the x, y, and z coordinates. Then map the point cloud coordinates to the virtual world in the VR environment. So the transform matrix from depth camera coordinates to the virtual world coordinates can be calculated and denoted as $M_{D \to V}$.

2. Many studies on the topic of camera calibrations have existed [Zhang (2000); Placht et al. (2014); Strauß et al. (2014); Wang et al. (2007)]. The standard RGBD camera and the high-speed camera can be calibrated using the checkerboard method. The transform matrix from the depth coordinates to the high-speed camera coordinates also can be calculated, which is donated as $M_{D \to H}$.

3. According to the two known transform matrices, the late unknown transform matrix from the high-speed camera coordinated to the VR virtual

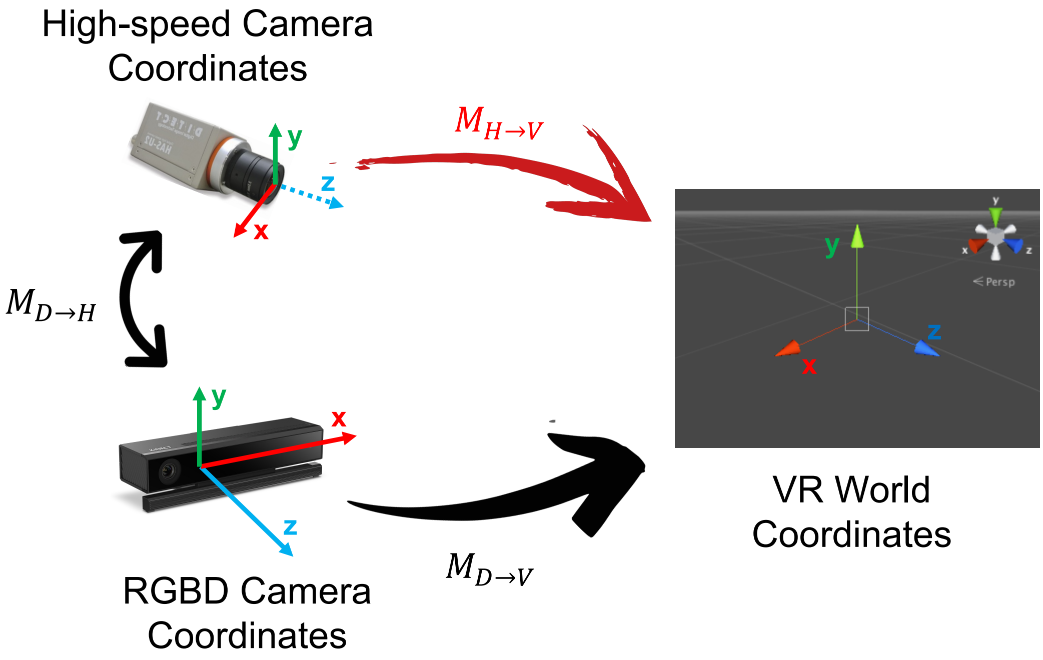


Figure 4.3: The unifying method for different coordinates: (1). get transform matrix $M_{D\to V}$ by depth data; (2). get transform matrix $M_{D\to V}$ by cameras calibration; (3). calculate transform matrix $M_{H\to V}$ using the known two transform matrix.

world coordinates is also can be calculated easily by the equation:

$$M_{H\to V} = M_{D\to H}^{-1} M_{D\to V} \tag{4.1}$$

4.2.2 3D Fusion

After unifying the coordinates of different cameras, it is possible to fuse these two kinds of data to reconstruct the 3D videos. Compared with the previous framework, this framework has a higher ability to prevent the misperceiving of a crucial action frame. For instance, when the users control the time as five times slower, in the previous framework, as the low recording rate, the system only can extend one frame's lasting time without providing more new information. While in this improved framework, as the high-speed camera is utilized, the new frame data can be played slower, which has a higher possibility of containing a crucial action moment frame.

However, a troublesome problem arises in this improved framework. Since the high-speed and standard cameras have different recording frame rates,

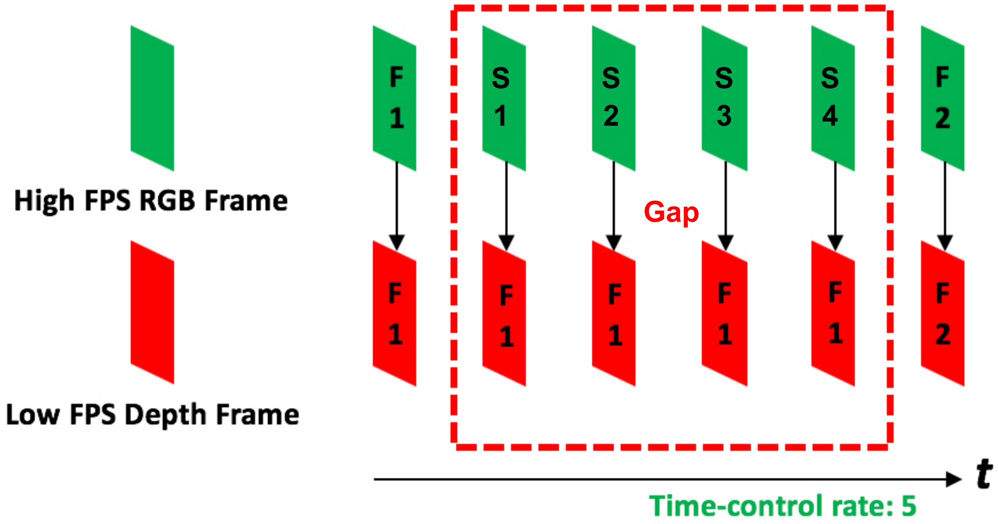


Figure 4.4: The interpolation method for controlling the time slower: In the case of controlling time to 5 times slower, use surplus RGB frames from the high-speed camera for interpolating the slower time; duplicate the previous depth frame for interpolating the slower time. There are slight gaps between the added surplus RGB frames and the duplicated depth frames.

there is a gap between the fewer depth frames and the more high-speed RGB frames. Several studies focus on interpolating this gap [Hasler et al. (2009b); Xu et al. (2005); Lewis et al. (2000)].

As the high-speed camera has a very high frame rate, and the interval between the data gap is very short, this work duplicates the previous frame data for all the frames in the gap. Figure 4.4 explains this method. This figure uses the same instance with five times slower Time-control. Here, the red blocks represent depth data frames from the standard low recording frame rate RGBD camera, and the green blocks are the high frame rate RGB frames from the high-speed camera. When the frame time is controlled as five times slower, as RGB data from the high-speed camera still has a surplus between the original Frame 1 and Frame 2, the new frames with the same time interval can be interpolated into this extended time, which is drawn as I1 to I4. On the other hand, the low-depth data can not provide more frames between Frame 1 and Frame 2 for interpolation. This work makes a compromise by duplicating the depth data of Frame 1 for the four new RGB data. This compromise can not prevent the slight gap between the high FPS RGB data and the low FPS depth. Considering the interval between these two frames is very short, the slight gap has little influence on the user experience.

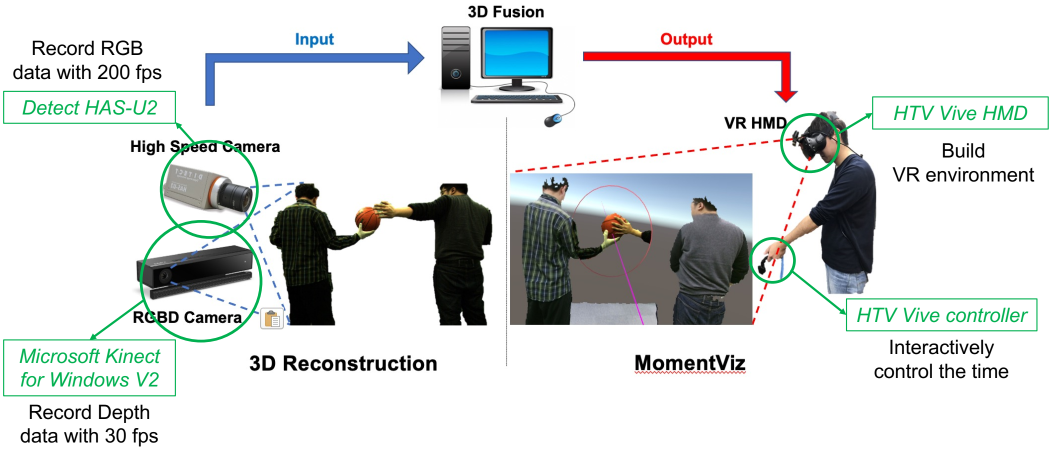


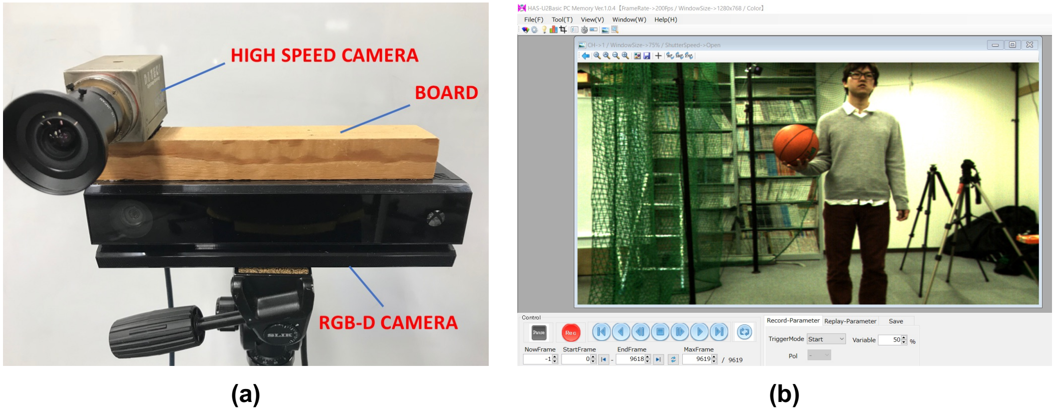
Figure 4.5: The implementation of improved MomentViz: left is the data preparation process, using a high speed RGB camera for high frame rate RGB data recording (with 200 fps), and a standard RGBD camera for depth data recording (with 30 fps); right is the improved MomentViz visualization process, which is as same as the original MomentViz visualization.

4.3 System Implementation

The implemented system of this improved framework is shown in Figure (4.5) In this improved framework, one high-speed camera called Detect HAS-U2 [1] is utilized to record the RGB data. This camera can provide a maximum of 7500 FPS and a maximum of 2592×2048 resolution color information recording. The detailed specification of Detect HAS-U2 is introduced in Table 4.1. This implementation uses 200 FPS and 1280×768 resolution for recording the dynamic scene. This high-speed camera is mounted on the standard RGBD camera (Kinect) for easy and exact calibration. A board is mounted between two cameras to prevent the occlusion caused by the two cameras' lenses. The setup is shown in Figure 4.6 (a).

At the same time, the software called HAS-U2Basic PC Memory attached to the HAS-U2 high-speed camera is utilized to record the high frame rate RGB data (Figure 4.6 (b)). This software provides a video in AVI format. In the implementation of this work, the open source library Open-CV 2.4 is utilized for extracting the image sequences from the recorded high frame rates video. The next step is transforming the RGB data coordinates to the Unity virtual world coordinates using the transform matrix calculated in the previous steps. Finally, import the high frame rates RGB and standard depth data into Unity, and map them with the unified Unity virtual coordinates.

[1] https://www.ditect.co.jp/products/camera/has$_u2.html$



HAS-U2Basic PC Memory Ver.1.0.4 【FrameRate->200Fps / WindowSize->1280x768 / Color】

CH->1 / WindowSize->75% / ShutterSpeed->Open

NowFrame -1 StartFrame 0 - EndFrame 9618 MaxFrame 9619 / 9619

TriggerMode Start Variable 50 %

Pol -

Figure 4.6: The hardware and software of improved MomentViz: (a) A high-speed camera (HAS-U2) mounted on a standard RGBD camera (Kinect V2), a board is between the camera for preventing the lens overlaid. (b) The software (HAS-U2Basic PC Memory) is utilized for high frame rates RGB data recording.

Table 4.1: Specifications of Detect HAS-U2

Item	Detail
Name	Detect HAS-U2
Sensor	1 Inch CMOS
Resolution	320×20 2592×2048
FPS	100 7500 fps
Lens Mount	C Mount
Shutter Speed	Max:10Î¼s
Sensor Sensitivity (550nm)	7.7V/lux.s
Camera Power Input	5V
Power Consumption	Below 4.5W
Weight	About 210g
Size	44×44×81.5mm

4.4 Evaluation Experiment

To evaluate the performance of MomentViz, a user study experiment is conducted. In this experiment, the participants are asked to judge a simulated basketball video to distinguish whether the player is fouling when stealing the ball. All the scenes are recorded from one Kinect coupled with a high-speed camera. The distance between the player and the cameras is about 5m. The main goal is to test the effectiveness of this improved framework of MomentViz.

4.4.1 Experiment Setup

Twelve participants are recruited to participate in the user study experiment, including nine males and three females aged 23 to 29. Participants are asked to watch the enactments of a basketball player attempting to steal the ball from the opponent (Figure 4.7). The task is to judge whether or not the steal would result in a foul (whether the defender hit any part of the attacker's hand while stealing the ball). The crucial moment of stealing the ball continued for approximately 0.03 seconds (6 frames), which was not long enough for the eye to grasp fully.

Three groups with different visualizations are tested, including the 3D visualization without time-control, Time-control 3D visualization, and Time-control H3D visualization (using the high-speed camera). The experiment is a repeated measures design, and eight scenes are tested for every visualization, randomized to alleviate ordering effects. All trials are conducted within a controlled VR environment. Participants wear an HTC Vive to watch the enacted scenes, and they use one of the HTC Vive controllers to finish the task.

All the scenes last approximately 3 seconds. The Time-control rate is set as five times slower in this experiment. Similarly, the maximum number of views is limited to 5 for every scene to control the total experiment time. Participants can make a judgment with less than five views if they are already confident enough to make a correct judgment.

4.4.2 Experiment Results

The results of the experiment are analyzed on both quantitative and subjective measures. Quantitative measures included the d prime value of judgments, the number of total views, and the task completion time, whereas subjective results are from a questionnaire.

D Prime Value of Judgments As the judgments are binary values with "yes" and "no", the d prime value (d') [Stanislaw and Todorov (1999)] is chosen in this analysis instead of accuracy, which is more suitable for analyzing

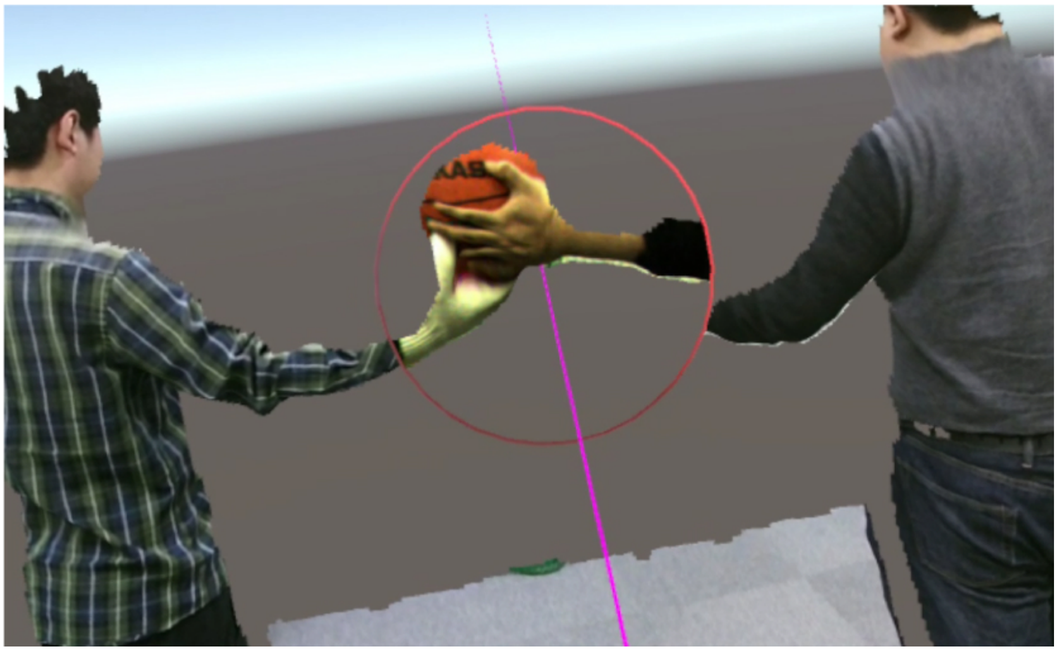


Figure 4.7: mage taken from the experiment videos of a player enacting a basketball foul with different visualization methods using Time-control H3D.

in this scenario. The d prime values of the three visualizations are calculated for each participant. In order to avoid the hit rate being 1 or 0, the upper limit and lower limit are adjusted slightly to 0.99 and 0.01. Then the mean and the standard deviation are calculated. The ANOVA is utilized to analyze the multiple-group results.

As shown in Figure 4.8(a), Time-control H3D has the highest d prime value of judgments with an average of 3.302 in all eight scenes. Similar to the pilot study, an analysis of variance (ANOVA) revealed a significant effect between 3 visualizations, which is confirmed with the Holm method for adjustment to evaluate the average d prime value.

From this analysis, it is found that there is a significant difference between 3D and Time-control H3D with $p < 0.05$. Cohen's d values are also calculated as the effect size. The Cohen's d is 1.27 between the 3D and Time-control H3D. On the other hand, there is no significantly different between 3D and Time-control 3D.

Number of Views In order to control the total experiment time, the maximum views are limited to a total of 5. The average number of views for every visualization is calculated. After the ANOVA analysis for comparison of the three visualizations and the Holm method for p-value adjustment, the results for the number of views are shown in Figure 4.8(b).

It is found that Time-control H3D required the lowest number of views, coming in at 2.552 times. In addition, significant differences between 3D and

Time-control 3D are found in this case, with $p < 0.001$. Moreover, there are significant differences between the 3D and Time-control H3D with $p < 0.001$. Cohen's d values are also calculated as effect size. It is 0.62 between 3D and Time-control 3D. Furthermore, between 3D and Time-control H3D, Cohen's d is 0.83. However, no significant difference is found between Time-control 3D and Time-control H3D.

Completion Time The task completion time for every scene is also recorded. It is defined as the start of the first view of the scene to the time a participant makes a judgment. Each visualization's average completion time is calculated, and ANOVA is used to analyze the results. The detail of the results is shown in Figure 4.8(c). There are significant differences between 3D and both Time-control visualizations with $p < 0.001$. For the average completion time values, 3D is 15.15s, Time-control 3D is 32.79, and Time-control H3D was 32.21s. These results can be regarded as favorable since both Time-control 3D and Time-control H3D are only twice as long as 3D despite the videos being five times slower.

Questionnaire Finally, the subjective questionnaire is checked, which the participants filled out after all the trials. The questionnaire consisted of 15 questions, including demographics, experiment experience, subjective evaluation, and comments. From the results, a quarter of all the participants have no experience with an HMD, and 58.3 percent of participants do not feel tired during the experiment. For the subjective questions, the answers are set in 5 levels as follows:

- Label 1: Strongly disagree;
- Label 2: Disagree;
- Label 3: Neutral;
- Label 4: Agree;
- Label 5: Strongly agree.

The result of the question "Can you see the foul moment definitely" is shown in Figure 4.9(a). It can be found that the median values of both Time-control 3D and Time-control H3D are 4. These results exceed the median value of 3D visualization, which is 2. From the results of the Friedman test and post-hoc Nemenyi test [Pohlert (2014)], it is found that there is a significant difference ($p = 0.012 < 0.05$) between 3D and Time-control 3D. There is also a significant difference ($p = 0.001 < 0.01$) between 3D visualization and Time-control H3D.

The result of the question "Would you like to use this mode for this task?" is depicted in Figure 4.9(b). Same as in the previous question, the median values of both Time-control 3D and Time-control H3D are 4. They are higher than the 3D visualization, which is 2. By similarly utilizing the Friedman and post-hoc Nemenyi test, the differences in multiple groups are tested. In the results, there is a significant difference ($p = 0.0217 < 0.05$) between 3D and Time-control 3D, and also a significant difference ($p = 0.0044 < 0.01$) between 3D and Time-control H3D.

The result of the task "Rank the three visualizations by overall preference" is shown in Figure 4.9(c). Labels 1, 2, and 3 mean the first, the second, and the third place, respectively. In this figure, it can be known that Time-control H3D achieved the best results, for which the median value is 1. On the contrary, 3D has the worst results, for which the median value is 3. Time-control 3D is in the middle.

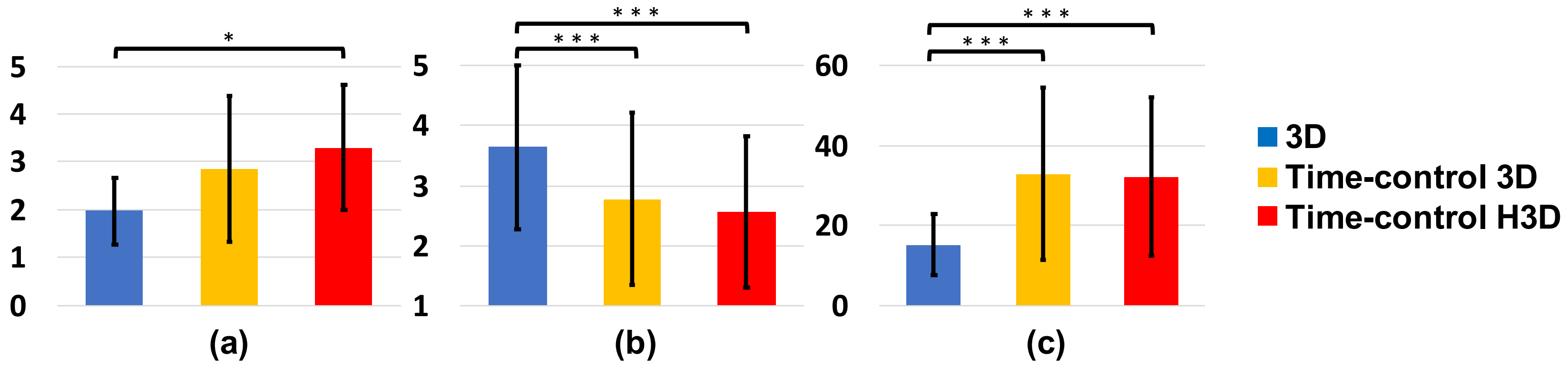


Figure 4.8: The results of evaluation experiment: (a) D prime values in judgments for each method. ($F = 3.66$, $df = 2$) (b) Number of views required to make a judgment for each method.($F = 17.30$, $df = 2$) (c) Completion time required to make a judgment for each method.($F = 31.58$, $df = 2$) (***: $p < 0.001$, *: $p < 0.05$)

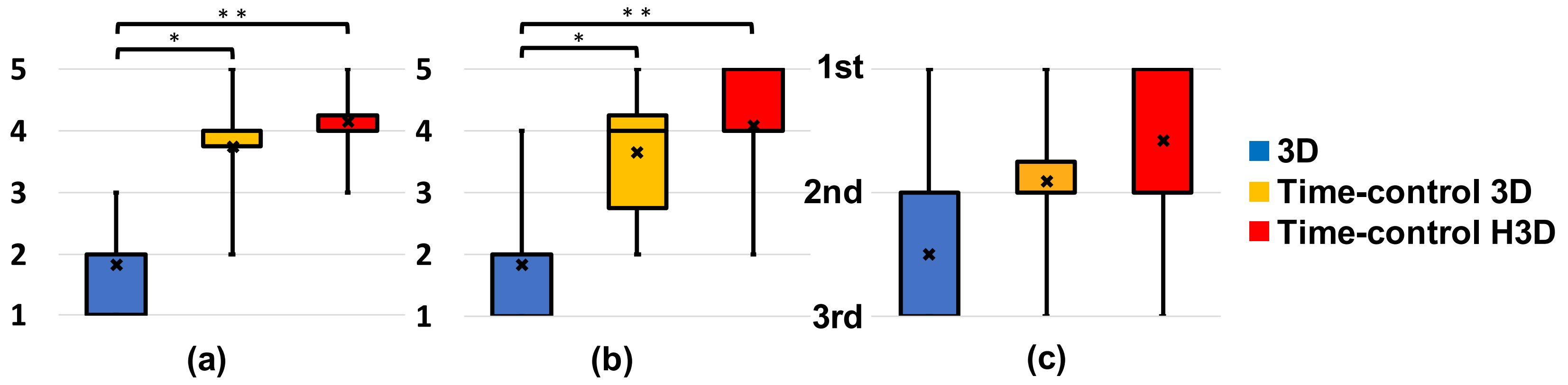


Figure 4.9: The results of questionnaire : (a) The question: Can you see the foul moment definitely? ($F = 39.35$, $df = 2$) (b) the question: Would you like to use this mode for this task? ($F = 17.07$, $df = 2$) (c) Participant rankings of the 3 methods by overall performance. (**: $p < 0.01$, *: $p < 0.05$)

4.5 Discussion and Future Work

According to the experiment results, it can be found that Time-control H3D outperformed conventional 3D visualization, which is good evidence that Time-control visualizations can perform well with rapid motion. The reason is likely due to the rapid speed of action. The users needed more time to find a better viewpoint to observe the dynamic scenes in the 3D visualization. However, in this Time-control experiment, the time dimension could be controlled more efficiently to resolve this issue. Therefore, participants have enough time to find a better viewpoint. As in the example presented in Figure **??**, (a) and (b) shows that in the viewpoint from below, it is easier to see that the defender's little finger touched the attacker's hand, and these Time-control visualizations can afford the users enough time to find this viewpoint.

Moreover, for the typing action in the pilot experiment, the action of a finger moving or pressing already can reveal typing the key. Participants need not judge whether the typist really pressed the key and just need to observe the pressing position. However, for the basketball foul judgment scene, participants should judge whether the player touched another player's hand, which is more complex than the typing scene. Hence, it can be inferred that the low FPS Time-control 3D would not be enough for a complex sports action. On the other hand, time-control H3D, which fuses high FPS frames with 3D data, can provide participants with more detailed information and still keep good performance. One typical example of the differences between these two methods can also be observed in Figure **??**. In this case, the camera with the low sampling rate prevented the crucial frames from being recorded during the foul. (c) Shows the blurry frames that likely caused some judgment confusion. On the other hand, (d) shows that Time-control H3D provides a higher probability that the crucial moment will be recorded. The d prime value results also can verify the inference. In the evaluation experiment, there is no significant difference between 3D and Time-control 3D, which means the Time-control 3D is not enough for this complex action. However, Time-control H3D has a significant difference with 3D, and it verifies well that Time-control H3D has better performance in rapid motion, especially in some complex sports scenarios.

Several remaining challenges still exist, which are expected to be improved in future versions. First, although Time-control H3D provides sufficient time to find an appropriate viewpoint to observe movements, it is still difficult for users to quickly find the best viewpoint in the VR environment. An automatic viewpoint optimization is desired in such a great amount of information in a free VR environment. Secondly, participants mentioned that the deviation between high FPS frame color information and low FPS frame depth information caused some blur, which likely influenced the system's performance. In addition, the hue difference between the low FPS color frames and the high FPS color frames also influenced the user experience. More appropriate frame interpolation and color enhancement methods are expected to be explored in the following work. Finally, this work chooses slow motion as the case of Time-control. In the actual situation, rapid motion is also expected in some cases. For example, much time is void in sports like boxing or baseball. Keeping rapid motion in Time-control can allow the spectators to save time to skim this low valuable information.

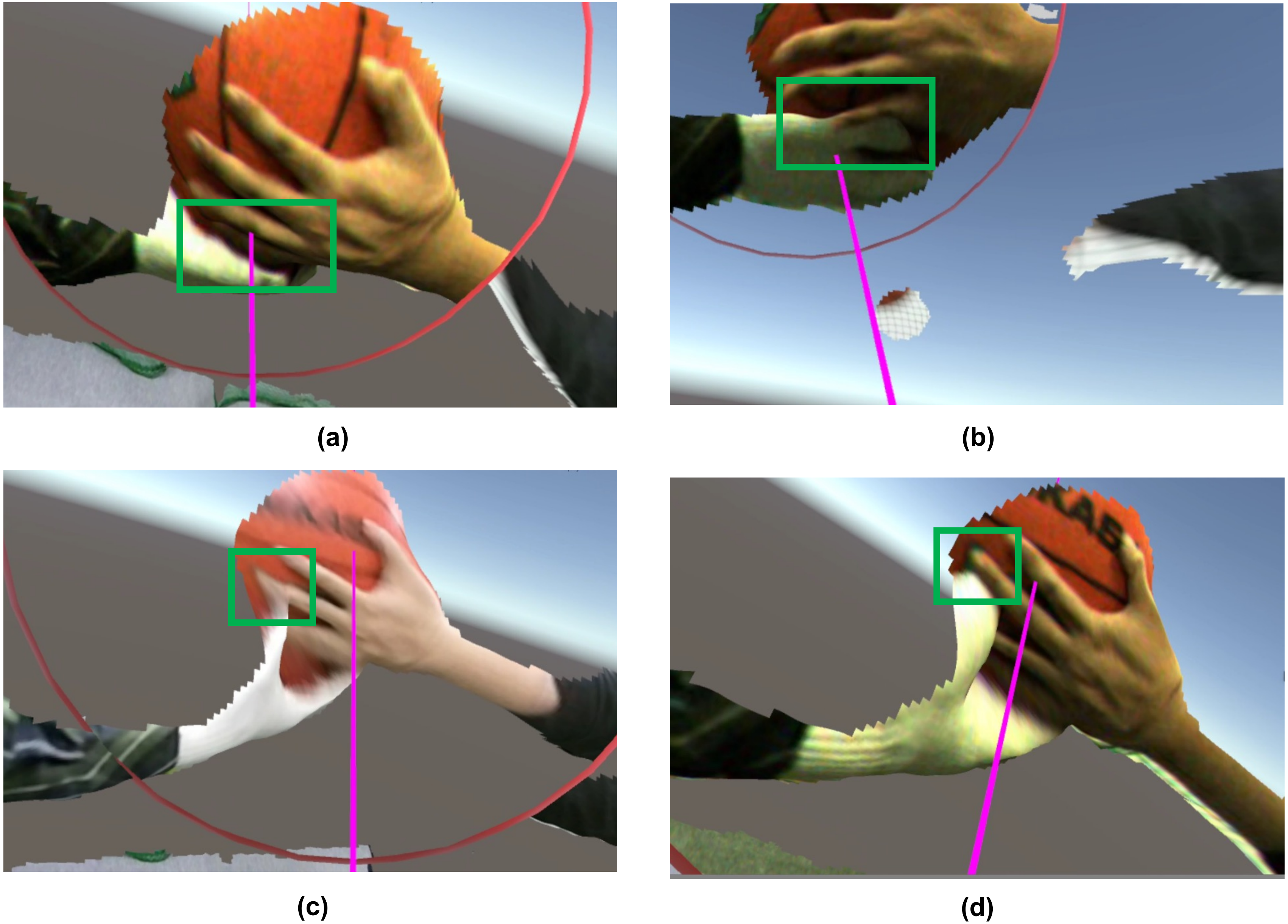


Figure 4.10: Two cases of the findings from the user study experiment: (a) the hitting moment can not be seen clearly from the front; (b) the hitting can be seen clearly from below; (c) the hitting moment can not be seen clearly in Time-control 3D visualization; (d) the hitting can be seen clearly in Time-control H3D (Green rectangles indicate the difference of crucial part for judgment in different visualizations).

4.6 Chapter Conclusion

In this work, the previous framework of MomentViz is improved by using a high-speed camera to get the high frame rate data. This improved framework not only provides time-control visualization but also provides higher frame rate RGB information which has a higher possibility to record the crucial action moment action frames. The camera calibration coordinates unify, and 3D fusion is processed sequentially to reconstruct the 3D frames. A user study experiment with 12 participants is conducted to evaluate the performance of the improved framework. The task is to spectate the simulated basketball videos and judge the rapid actions of stealing the ball. The d prime of judgment, the number of views, and the completion time are analyzed as the quantitative evaluation, while the user experience questionnaire is checked for the subjective evaluation. The user study experiment results suggest that the improved Time-control framework outperforms the previous framework and can increase the user experience when spectating rapid sports videos.

Chapter 5

Viewpoint Optimization Based on Individual Preference

5.1 Introduction

Along with the development of VR technology, watching sports matches in VR already comes true. However, the issue that the crucial actions may be occluded from some bad viewpoint angle when spectating the sports also exists in VR. Moreover, from the finding of Chapter 3, it is more difficult for the spectators to find a good viewpoint in the 3D space of the VR environment. Optimizing the viewpoint for spectating sports videos in VR is an urgent need.

As noted in Chapter 2, conventional viewpoint selection methods ignore the individual preference of spectators, select the viewpoints by the directors or research rules, and provide the same view to all the spectators who may have different preferences. In order to break through these limitations, this work proposes a novel optimal viewpoint selection method for watching sports actions that depend on individual preferences. This work chooses boxing as a target case since, in boxing matches, the frequency of boxers moving and rotating leads to more situations where the crucial action is occluded. This work uses a neural network classification model for optimal viewpoint training. In the training model, the visibility of the boxer's upper body parts, the punch side, and the punch offset are extracted as the features; the user-selected preferred viewpoints are collected through an experiment as the training label. To get the visibility of boxer upper body parts, a method that raycasting the boundary box to check the visibility is proposed. With the rule of least visibility difference (LVD), the user-selected viewpoints can be classified into finite candidate viewpoint classes, which are set in advance.

The accuracy of training results is evaluated as the scientific evaluation. In order to verify the empirical performance of this work, a user study is conducted. Twenty-one participants are recruited to join the experiment. All the participants test the system optimal viewpoint suggestion group (SmartVP) and random common viewpoint group (No SmartVP) in this experiment. The hypothesis is: 1). the SmartVP can reduce the moving steps for the user to find a preferred viewpoint. 2). the viewpoints SmartVP suggested can get higher user-satisfied scores. The user study results verify that the hypothesis is accepted and gives strong evidence that this framework can optimize

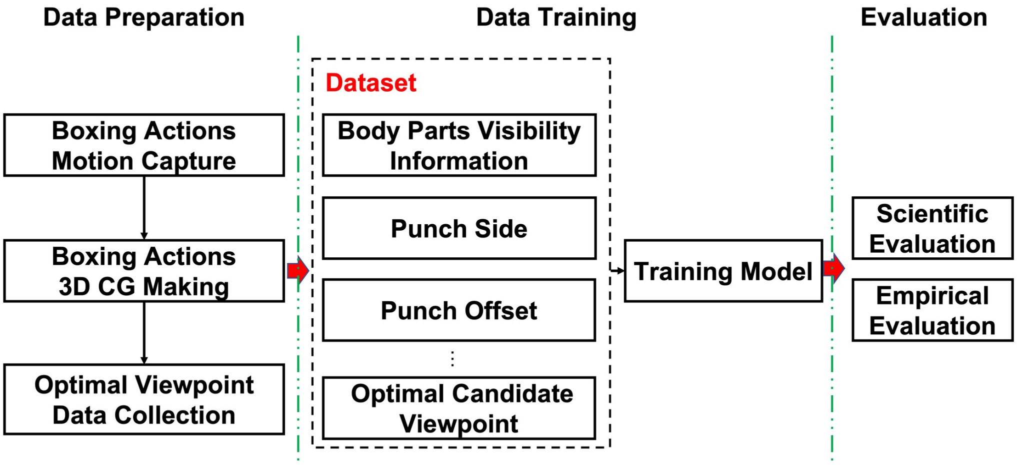

Data Preparation
Data Training
Evaluation
Dataset
Boxing Actions Motion Capture
Boxing Actions 3D CG Making
Optimal Viewpoint Data Collection
Body Parts Visibility Information
Punch Side
Punch Offset
Optimal Candidate Viewpoint
Training Model
Scientific Evaluation
Empirical Evaluation

Figure 5.1: The high-level overview of framework, includes three steps with data preparation, data training and evaluation.

the viewpoint for spectating sports occluded actions depending on individual preference.

The main contributions of this work can be summarized as follows:

- Design and conduct an experiment to collect the user-preferred viewpoints;
- Extract the features and propose a method to get visibility;
- Propose a method to label the continuous user-selected viewpoints;
- Design the training model for optimal viewpoint selection training;
- Conduct a user study experiment to evaluate the system's performance.

The remainder of this chapter presents the framework details, the user study design, the evaluations, and the findings from the results.

5.2 Methodology

This method utilizes complete 3D information and considers the personal preference of the spectator for watching a boxing punch action. It has an integral and systemic neural network framework, which includes three steps: data preparation, data training, and result evaluation. Figure 5.1 shows a high-level overview of the framework.

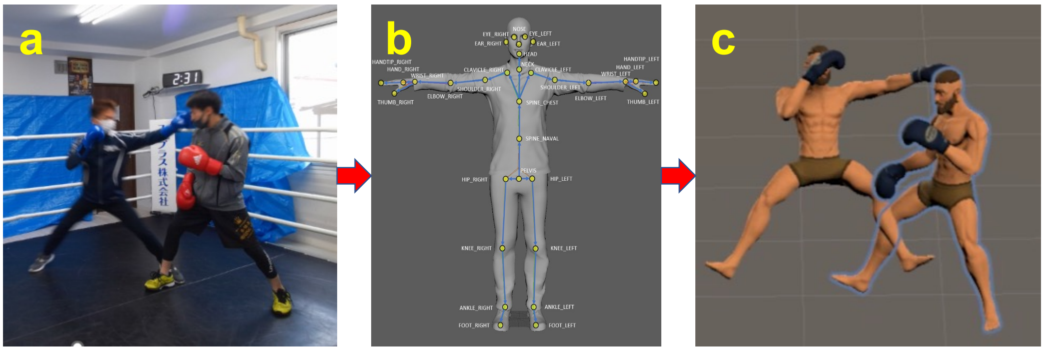


Figure 5.2: Make the 3D video from the real boxing punch scene (a) record the real boxing scene with a depth camera. (b) extract the skeleton data. (c) use skeleton data to make the 3D boxing video.

5.2.1 Data Preparation

Preparing the training data is the first step in this work, including boxing punch actions 3D information as the training features values and the optimal viewpoints spectators selected as the training label values.

In order to get honest and reliable boxing punch actions 3D information, two boxers are recruited to simulate the actual boxing punch scenes. The marker-less method depth camera is utilized to avoid the markers affecting the boxers' performance. However, there are some research and technology about 3D reconstruction for humans with depth data [Dou et al. (2016); Newcombe et al. (2015)]. As 3D reconstruction is not the main topic of this work, moreover, for the goal of this work, the exact pose is more required than the reality of the 3D model. Considering the reason above, instead of natural human body 3D reconstruction, 3D avatar boxing punch scenes using exacted skeleton data is a better choice (Figure 5.2).

To collect the optimal viewpoint data which the spectators selected when they watched the boxing punch actions, we designed and conducted an experiment in VR. We set some experiment limitations to simplify the problem and control the experiment time. The distance and height of the user viewpoint are constrained. The viewpoint height is fixed at 1.5 meters, which is approximately the height of the boxers' hands. The scene center is forced to be set to the center of two boxers, taking it as the center of a circle and 1.5 meters as the radius. On this circle, which we call the View Ring, experiment participants can freely select any angle as their optimal viewpoint. The values are continuous in the scope from $0° \sim 360°$ (Figure 5.3).

5.2.2 Features

Depending on the possibility of influencing the viewpoint selection for spectating boxing punch action videos, these features are extracted for data training

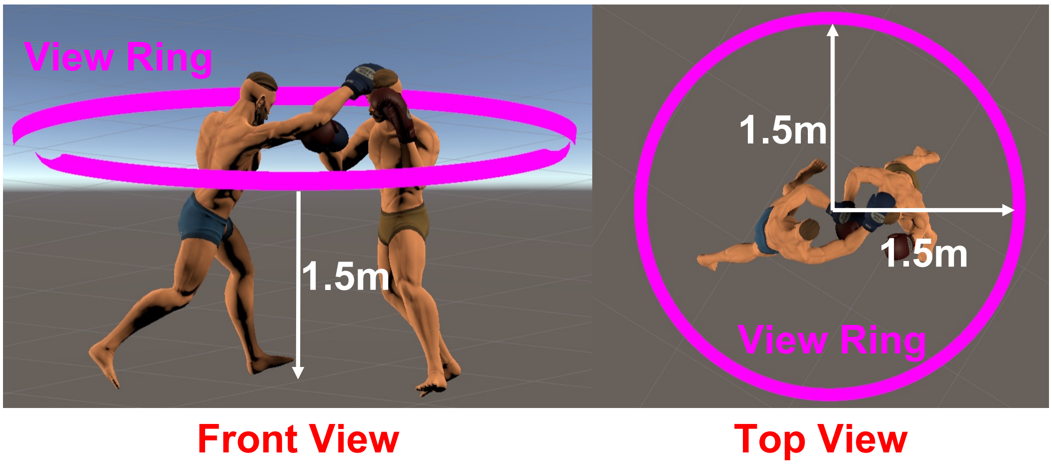


Figure 5.3: The size of the view ring (the height is 1.5m, the radius is 1.5m, the ring center is the center of two boxers, viewpoint focuses on the ring center). Experiment participants can turn the viewpoint around the center of this ring to choose their preferred viewpoints.

from the 3D punch action scenes:

Visibility (V) The visibility of boxers' body parts should be the most important factor for viewpoint selection. Unlike the viewpoint selection method for general human actions, different body parts should have different important weights for watching boxing punch actions. The boxer's upper body is separated into eight parts: head, body, left hand, left arm, left forearm, right hand, right arm, and right forearm. The main idea is that calculate the visibility of every part from the viewpoint and then get a series of visibility data composed of 16 body parts from 2 boxers.

The raycasting approach is utilized to reduce computation consumption and increase precision instead of the traditional computer graphic image recognition method. The first step is binding a box collider for every part, tagging the collider with its part name, and then setting 12 checkpoints on the box vertexes of each part (Figure 5.4). The next step is casting the rays from the viewpoint to all checkpoints and getting the collision information. Then check the collider tags of all the checkpoints which the raycasting returned. If the returned collider tag is the same as the box tag, it means this checkpoint is not occluded (Figure 5.5). The final step is summing the visible checkpoint counts for every part. These visible checkpoint counts represent the visibility of parts.

Punch Side (PS) The punch side indicates which hand casts this punch. As the punch side has a possibility to influence the viewpoint from which side, it is also essential for viewpoint selection. As a result, it is extracted for the

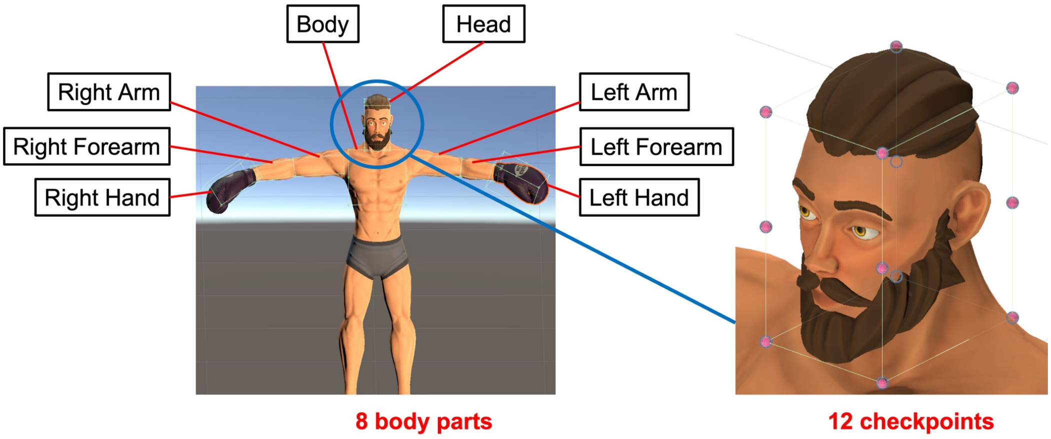


Figure 5.4: A model for visibility calculation by using boundary box: divide the boxerâs upper body into 8 parts, bind a box collider to each part, and set 12 checkpoints for every box vertex (pink points in right figure). Check the number of visible checkpoints to present the visibility of each part.

model training. In order to combine the categorical data and numerical data, one-hot is utilized for encoding. The left punch is encoded as 1, and the right punch is encoded as -1. This feature is marked manually in advance (Figure 5.6).

Punch Offset (PO) A large offset reveals that the punch is dodged in this scene, and the dodged punch usually brings less occlusion, so multiple viewpoints have a chance to be selected by the spectators. For this reason, the punch offset also is considered to influence the viewpoint selection results. The method to extract this feature is simply getting the center position of the hand tip checkpoints and the center of the head checkpoints, then automatically calculating the distance of these two center positions. (Figure 5.6).

5.2.3 Label Design

From the user experiment mentioned in the previous section, it can get the optimal viewpoint data of every check punch action frame. Since the participants can select any viewpoint freely on the view ring, the collected data is continuous. However, as one particular occluded bad viewpoint angle may be in the middle of two good viewpoint angles, the viewpoint quality is not linear and continuous on the view ring by its spatial position. Moreover, the continuous label values need enormous data sets for training, which is challenging to realize in this work.

In order to solve these difficulties, the continuous viewpoint data is expected to transform into the discrete viewpoint data in the scope of a finite set. The idea for this issue is to separate the view ring into equal angles, set

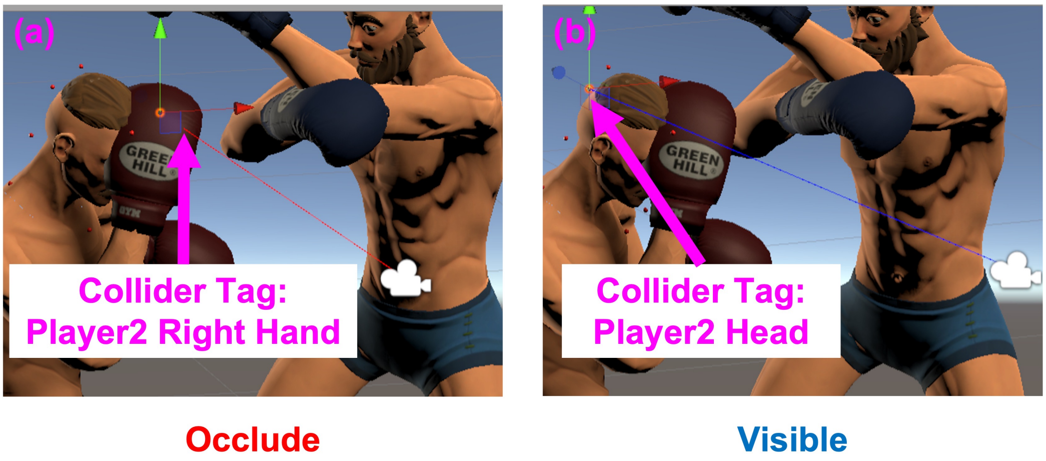


Figure 5.5: Raycasting approach for checking the visibility of two checkpoints on the head. (a) The collider tag is "right hand", which is different from the checkpoint position "head" and means occluded. (b) The collider tag is "head", same as the tag of checkpoint position, and means visible.

the virtual candidate viewpoints on the divided angles, and finally classify all the participants selected optimal viewpoint data into these candidate viewpoints by the rule. More candidate viewpoints can get more precise viewpoint classification, meanwhile needing more data sets to train the models. In practice, finite candidate viewpoints should be set by the data set size. After the classification, all the participant-selected optimal viewpoints belong to several classes, and these are used as label values for the training.

For the viewpoint classification rule, choosing the closest Euclidean Distance to label the optimal candidate viewpoints to the candidate viewpoint classes is the most straightforward idea. Since the viewpoint quality is not linear in the spatial position on the view ring, even the closest Euclidean distance can not ensure having a similar view if the closest candidate viewpoint gets the occluded view by coincidence. So instead of the spatial position, the rule that finds the least difference in visible checkpoint counts is proposed in this work. This rule's essence is focusing on the view contents themselves. It is assumed that the least difference of all boxer body parts visibility can represent the most similar view of the boxing punch action scene. As the previous section explained, the raycasting approach gets the visible checkpoint count series of boxers' body parts from the optimal viewpoint participants selected. Using the same way, it also can get different visible checkpoint count series from each candidate viewpoint. Then compare the optimal viewpoint visibility series and all candidate viewpoints visibility series, and calculate the sum difference of every body part by absolute values for each candidate viewpoints. Finally, pick up the one with the least difference, and set it as the label value. The equation can be represented as:

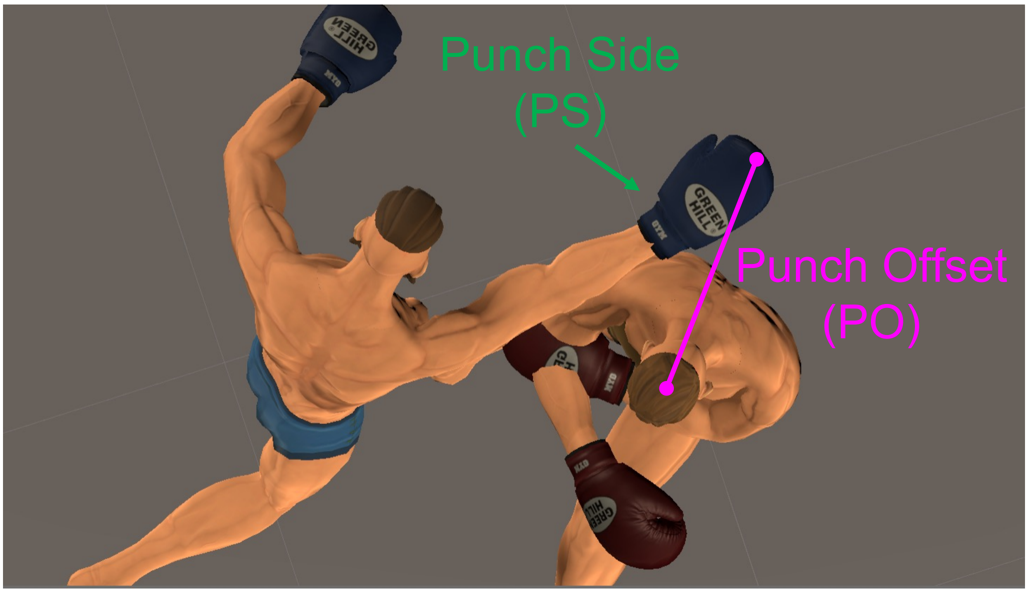


Figure 5.6: The features of punch side and punch offset: punch side indicates which side of hand casts the punch, punch offset presents the distance from the punch to the head. This two features are marked manually in advance.

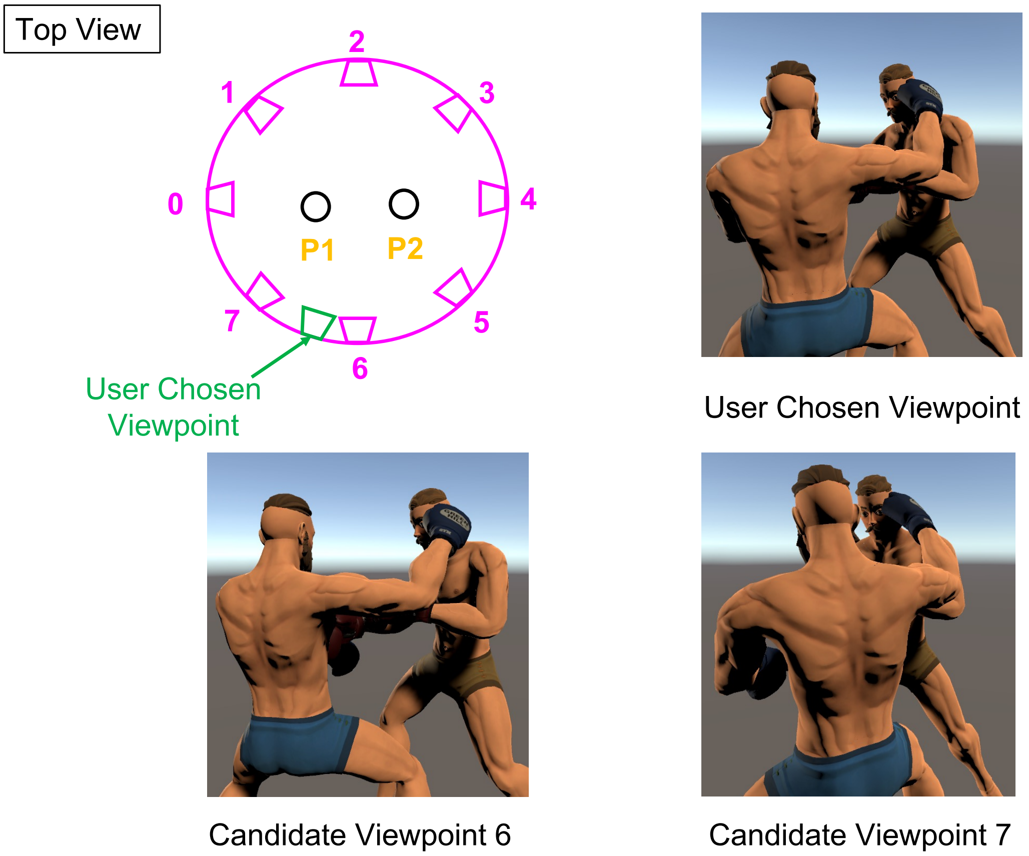
$$i^* = \underset{i=(0,\ldots,C_{vp})}{\arg\min}\ Dif(i) \tag{5.1}$$

where

$$Dif(i) = \sum_{j=1}^{16} |\bar{V_j} - V_{i\_j}| \tag{5.2}$$

is the visibility difference function. i is the number of the candidate viewpoint, j is the number of the body part (2 boxers, a total of 16 body parts), C_{vp} is the count of all candidate viewpoints, $\bar{V_j}$ is the visibility of j body part in the optimal viewpoint participant selected, and V_{i_j} is the visibility of j body part in the candidate viewpoint i.

For example, In Figure 5.7 (a case of 8 candidate viewpoints), the red viewpoint is the optimal viewpoint participant selected from Euclidean distance, and candidate viewpoint 6 is the closest one, but if compared with the visibility difference, candidate viewpoint 7 has the least difference. So, in this case, viewpoint 7 is regarded as the label value. This rule also has its limitation, as the body parts have different important weights for viewpoint selection in our supposition, so classifying the viewpoint to one with less difference in total parts but more difference in some important parts will result in the error. However, the limitation just exists in the few cases, which will not seriously affect the training results.



User Chosen Viewpoint

Candidate Viewpoint 6

Candidate Viewpoint 7

Figure 5.7: A case labeling the user-selected viewpoint to the candidate viewpoint by visibility difference (candidate viewpoint 6 is closest to the spectator selected viewpoint, but candidate viewpoint 7 has the least visibility difference)

5.3 Implementation

The approach is realized by using a neural network model. This section describes the implementation of data preparation and training.

5.3.1 Punch Data Collection

One Azure Kinect was utilized in the implementation to capture the actual boxing punch motion, which has an RGB camera with a maximum of 3840×2160 resolution in 30FPS and a depth camera with a maximum of 1024×1024 resolution in 30FPS. With the help of a boxing gym, two boxers were hired to simulate the real boxing punch videos. The helpers, a former professional boxer and another an amateur boxer, are experienced and familiar with any boxing punch. The recording was conducted in the ring of a boxing gym. The camera was set at the corner of the boxing ring, and the height was

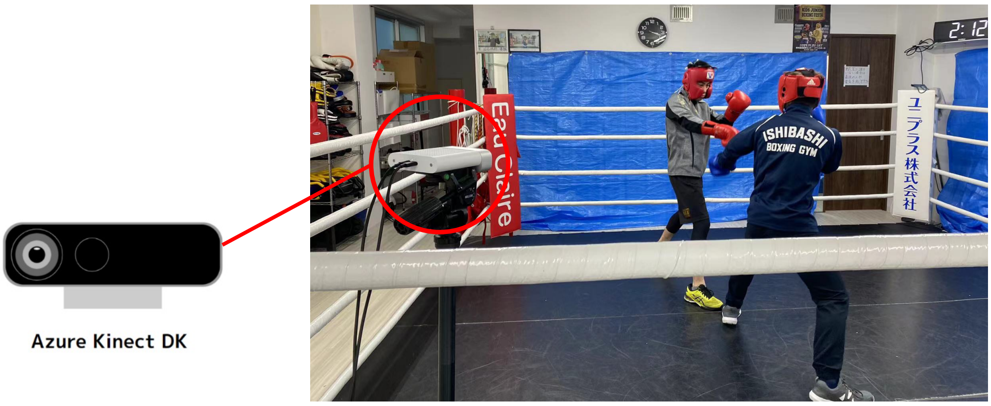


Figure 5.8: The Azure Kinect depth camera setting for boxing punch scene recording: one Azure Kinect depth camera is set in the corner of the boxing ring.

approximately the same as the boxer's hands (Figure 5.8). Sixteen boxing videos were recorded successfully, including job, cross, hook, and uppercut, four different boxing base punch types. The punch target was limited to the head; some punches were designed to hit the target, and others were dodged. Azure Kinect can trace 32 human body joints' depth data (both position and quaternion) directly [1]. The pose temporal data was obtained using the simple middle interpolation method of occluded joint data with neighbor frames.

5.3.2 Labelling Experiment

The labeling experiment is implemented in the VR environment by HTC Vive. Twenty-four participants (16 males and 8 females, ages 19 to 25) are recruited to participate in our experiment. Considering the experiment time, we chose 16 boxing punch videos covering all four punch types and randomly divided them into two groups. Every group has eight videos, and 24 participants would be divided into these two groups equally. Every punch video extracts three punch-hitting moment frames for the user to choose the preferred viewpoint. Finally, three different preferences are set in the experiment as follows:

- No Preference (**None**): This task is to see the punch clearly, and distinguish whether this punch was hit or dodged.
- Preference Arm Form (**AF**): The participants are asked to see the punch-hitting moment clearly, and the punch arm form as clearly as possible.
- Preference Facial Expression (**FE**): The participants are asked to see the punch clearly, meanwhile could see the punched boxer's facial expression as clearly as possible.

[1] https://docs.microsoft.com/en-us/azure/kinect-dk/body-joints

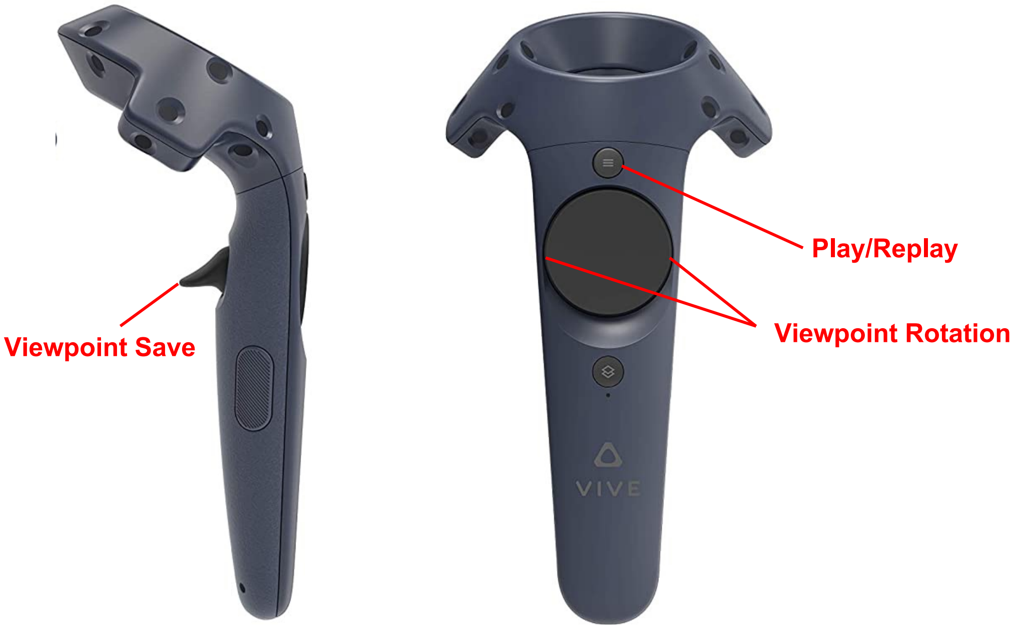
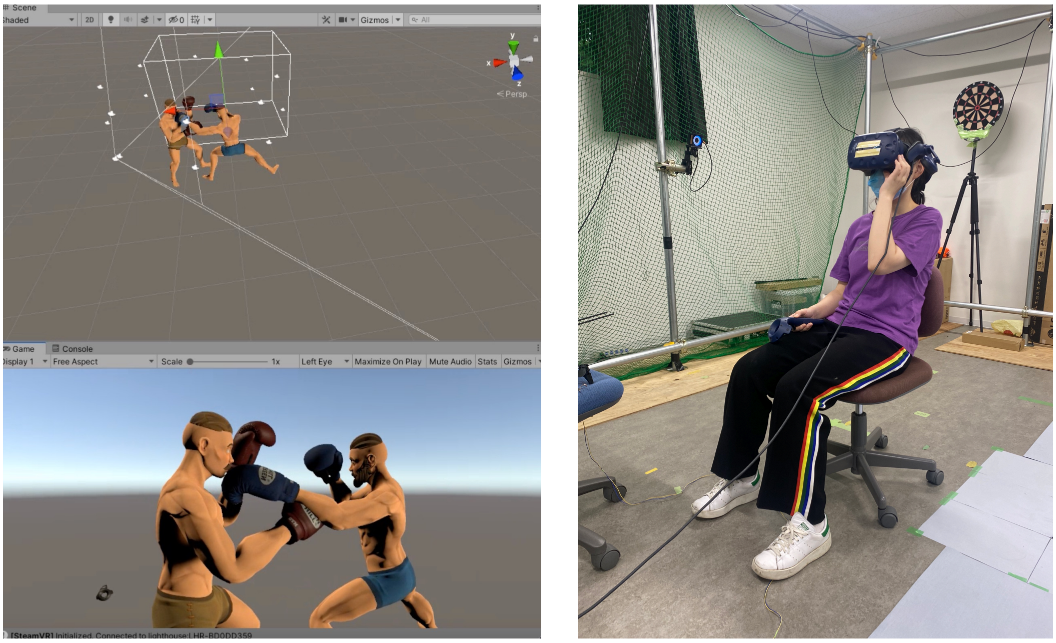


Figure 5.9: Using one HTC Vive controller for viewpoint selection in experiment: Menu button is set for playing and replaying the videos; Touchpad is set for rotating the viewpoint on the view ring; Trigger is set for saving the selected preferred viewpoint.

Group None is the control group. AF and FE are two cases of spectator preferences. The participants are asked to watch the punch action videos and check several given frames to select the optimal viewpoint according to the above tasks. The participants sit in the Vive space during the experiment, and the default viewpoint is set to face the center of the view ring. The participants use a Vive controller to control viewpoint rotation and select the viewpoint (Figure 5.9). Every participant should finish all three preference groups. Moreover, to reduce randomness in participant selections, every video was asked to finish three times in random order. The total experiment time of every participant is 120 minutes, including several breaks (Figure 5.10).

5.3.3 Neural Network Model Design

Since the viewpoint selection problem is complex and nonlinear, and the data set has a scale of thousands, the neural network classification model is chosen for data training. The goal is that by using the proposed features, the trained model can predict the optimal viewpoint of any boxing punch action scene for the spectators based on their preferences. In this model, the main idea is that input the scene punch features (punch side and punch offset) and every body part visibility from all the candidate viewpoints; the output value is the optimal viewpoint class. For the label of training data, considering the



[SteamVR] Initialized. Connected to lighthouse:LHR-BD0DD359

Figure 5.10: A labeling experiment for collecting the user preferred viewpoint in different preference condition: left top is scene view of the experiment program, left bottom is the view in VR device which the user can see, right shows that a participant takes part in the experiment who is sitting on a chair and using a controller to select preferred viewpoint.

experiment time and cost, eight candidate viewpoints are set in advance as the case in this work, which 45° between two candidate viewpoints. The candidate viewpoints are fixed at the position on the view ring and named from viewpoint $0 \sim 7$ in order (Figure 5.11).

Figure 5.12 shows the training data format details. OVP refers to the optimal viewpoint class, which transforms from the participant-selected optimal viewpoint, and the value is in scope from candidate viewpoint 0 to candidate viewpoint 7. V represents the visibility of every body part from a particular candidate viewpoint, as we set 8 candidate viewpoints. This type of data has eight groups, and every group has 16 numbers, which include two boxers with eight body parts (head, body, left hand, left arm, left forearm, right hand, right arm, right forearm) for each boxer. This information involves the view features of every candidate's viewpoint. The last two numbers PS and PO denote the scene's punch side and offset, which covers the frame features. With the scene punch features, this model is expected to learn the important weight of every body part visibility, understand the interrelation between the body parts visibility of every candidate viewpoint, and the result of the optimal viewpoint class.

The advantage of this model is that it is straightforward to use and does not need any other additional post-processing in the output values. On the

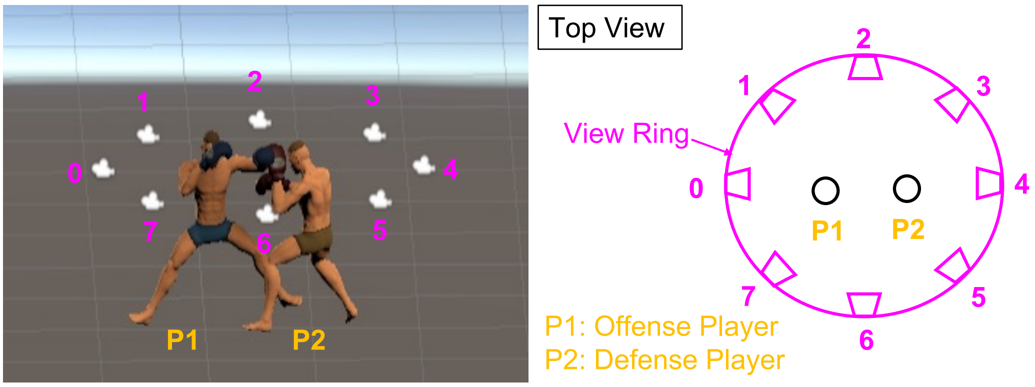


Figure 5.11: The candidate viewpoints setting in this implementation: 8 candidate viewpoints are set on the view ring evenly, and 45 degrees between the two candidate viewpoints.

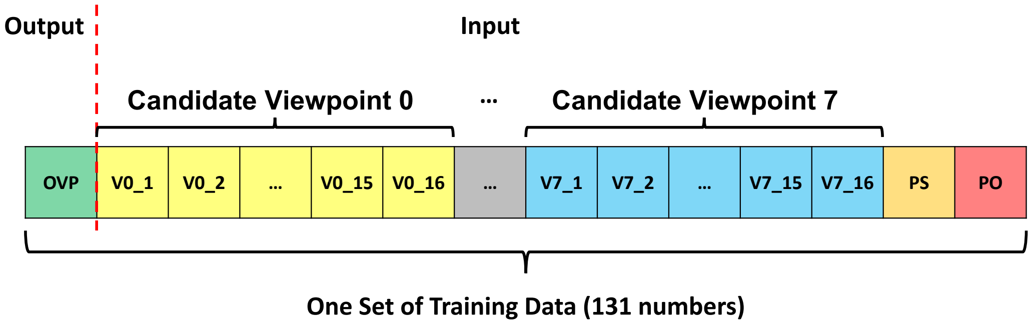


Figure 5.12: The format of one set of training data in the classification model: OVP is the optimal viewpoint class, V is body part visibility including 2 boxers total 16 parts, PS is punch side, PO is punch offset.

other hand, the training result is strongly influenced by the training data set. If some particular candidate viewpoint classes have little or a loss of training data, these classes will have a high possibility to result of insufficient training.

5.3.4 Data Training

The python scikit-learning library is utilized for implementation and training. In this model, every selection by the participants is regarded as one set of training data. 24 participants are separated into two groups. Every group checks 16 videos, and every scene has three frames for checking. Every frame is checked three times by each participant, so there are a total of 1,728 sets of training data for every experiment task. After multiple trials, the following parameter setting yields the best training results (Table 5.1). The solver is set as "adam" because of the large amount of data set, over 1000. Since the internal relations between input and output are multilevel and indirect and the input dimension is large, four levels of hidden layers are set, and (200, 100, 50, 50) neural nodes for every hidden layer, respectively. For the activation function, "logistic sigmoid" is chosen because of its good performance in classification problems. Max iteration times are set to be 10,000. K-fold cross-validation is utilized to separate the training and test data, and the k is set as 8 in this implementation. The average training time is less than one minute.

5.4 Evaluation and Discussion

In order to evaluate the performance of SmartVP, the scientific evaluation of the machine learning training result and the empirical evaluation of a user study result are analyzed.

5.4.1 Scientific Evaluation

According to the different additional preferences in the data collection experiment, three groups with None, AF, and FE are trained. After the training, the model can predict the optimal viewpoint from candidate viewpoint classes for any boxing punch scene. Since the viewpoint selection problem does not have any ground truth, all the label values depend on the objective selection of participants, and no scene can get a uniform selection from all the participants. Simply comparing the predicted optimal viewpoint and the most selected viewpoint can not evaluate this problem precisely.

A method called Ratio Highest Accuracy (RHA) is proposed to evaluate the results more precisely. This method first picks up the highest selection count in each frame, then calculates the RHA for every candidate viewpoint by calculating the ratio to the highest score. For instance, the case 1 in Table 5.2, the RHA of viewpoints 0, 2, 3, and 7 are 0%, and the RHA of viewpoint 1 is 66.7%, which is calculated by 8/12. Similarly, the RHA of viewpoint 6 is 33.3%. Viewpoints 4 and 5 can get 100% accuracy. In case 2, viewpoint 2 gets the RHA as 91.7%, which can keep its real value well. The final results of this model with different preferences are obtained through this method (Table 5.3).

Table 5.1: Parameter setting of data training

Solver	adam
Activation function	logistic
Hidden layer levels	4
Neutral nodes	(200,100,50,50)
Max iteration	10000
K-fold	8

Table 5.2: Two cases of the experiment participants' optimal viewpoint selections.

	VP0	VP1	VP2	VP3	VP4	VP5	VP6	VP7
case 1	0	8	0	0	**12**	**12**	4	0
case 2	0	5	11	2	0	4	**12**	2

Table 5.3: The results of data training

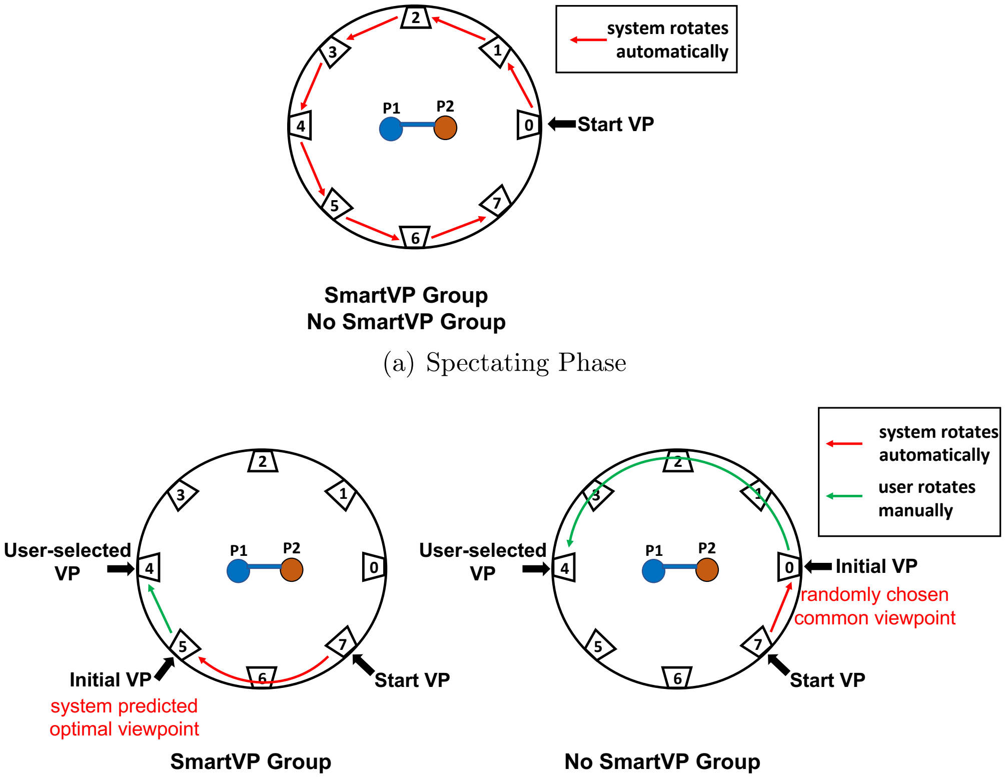
None	Arm Form (AF)	Facial Expression (FE)
64.93%	68.61%	**75.31%**

5.4.2 User Study

In order to evaluate the practical performance of SmartVP, a user study by implementing a recommendation viewpoint system for boxing punch video was designed and conducted. This user study utilized eight punch videos that had not been used for training. Four videos were used for Preference AF, and the other four were for Preference FE. The punch moment frame was extracted for prediction, and the predicted value was regarded as the optimal viewpoint for this punch video. Twenty-one university students aged 18 to 25 were recruited to participate in the user study, including 10 males and 11 females. All the participants were asked to finish the tasks of two groups, one was applying SmartVP, and the other was the control group without SmartVP. Each video was tested in two groups. The experiment process has two phases.

- **Spectating Phase:** Phase 1 is the Spectating Phase. In this phase, the participants were asked to observe the punch video from each candidate's viewpoint and find the best viewpoint depending on different preferences (Figure 5.13 (a)).

- **Selecting Phase:** Phase 2 is the Selecting Phase. In this phase, firstly, the system will choose an initial viewpoint. The SmartVP group chose the predicted optimal viewpoint as the initial, while the control group chose a random one from the other seven common candidate viewpoints (not the optimal viewpoint). Then the participants were asked to move the viewpoint from the initial viewpoint to the best viewpoint they selected in phase 1 (Figure 5.13 (b)).



(a) Spectating Phase

(b) Selecting Phase

Figure 5.13: Two phases of user study: (a) Spectating Phase: in this phase, two groups will be the same cooperating, and the system will rotate the viewpoint from candidate viewpoint 0 to 7. From each candidate viewpoint, the punch video will replay once. The users are asked to spectate the videos from all the candidate viewpoints and try to compare all of them.; (b) Selecting Phase: in this phase, the system will automatically rotate the viewpoint to the initial viewpoint first. In the SmartVP group, this initial viewpoint is the optimal viewpoint, while in the No SmartVP group, this initial viewpoint is a random common viewpoint. Then the users are asked to rotate the viewpoint from the initial viewpoint to the viewpoint they prefer.

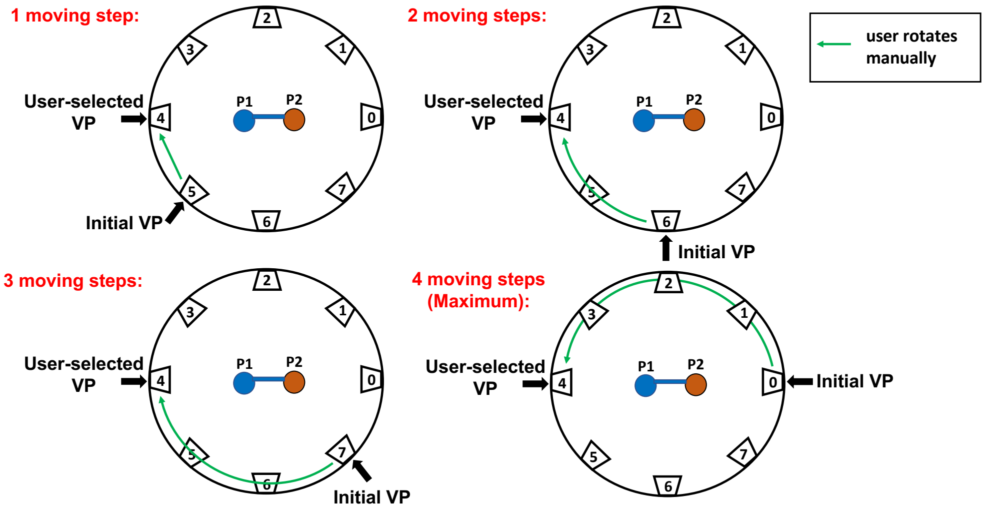


Figure 5.14: The rule of moving step count: Moving to the neighbour candidate viewpoint counts one moving step, and the maximum moving step is 4.

Moving the viewpoint to the nearby candidate viewpoint counts as one moving step, so the maximum moving step for every video is 4 (in this case, two viewpoints have an angle of 180 degrees) (Figure 5.14). The participants could replay the video from any candidate's viewpoint in this phase. Twenty-one participants were divided into three groups, and in every group, 7 participants got a different one from the 7 common candidate viewpoints as their control group's initial viewpoint. When one group's tasks finished, the participants were asked to score the performance of initial viewpoints from 1 to 7. A higher score means better performance of viewpoint. An ethical review was reviewed by the university IRB (Institutional Review Board) in advance, and all the participants were asked to fill out a questionnaire after the experiment.

Hypothesis From the scientific results of the training model evaluated in section 5.4.1, this method is excepted to have the potential to optimize the viewpoints in the practical application. So the hypothesis of the user study can be set as follows:

H1 Both Preference AF and FE, the participants in SmartVP group needs fewer moving steps to reach the best viewpoint.

H2 Both Preference AF and FE, the initial viewpoints of the SmartVP group are more satisfied by the participants.

5.4.3 Empirical Evaluation

The user study results are shown in Figure 5.16. For the moving step (Figure 5.16 (a)), in Preference AF, the average moving step of SmartVP and No

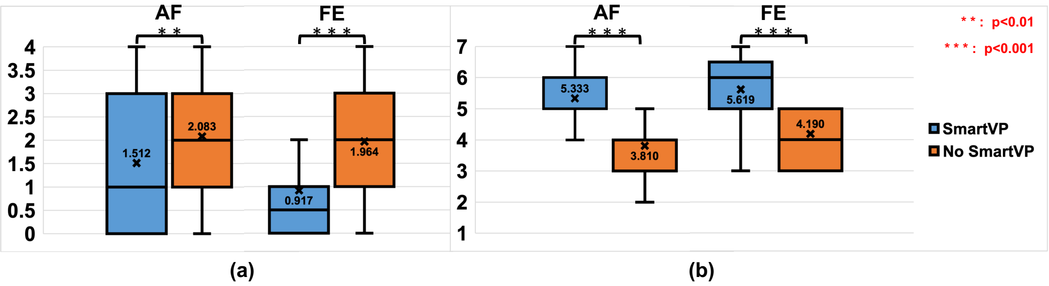


Figure 5.15: The user study results: (a) the moving steps from the initial viewpoint to the user-selected best viewpoint. (b) the user-graded score of the initial viewpoints. (The initial viewpoints in the SmartVP group are system-recommended optimal viewpoints, and in the No SmartVP group, they are the random common viewpoints.

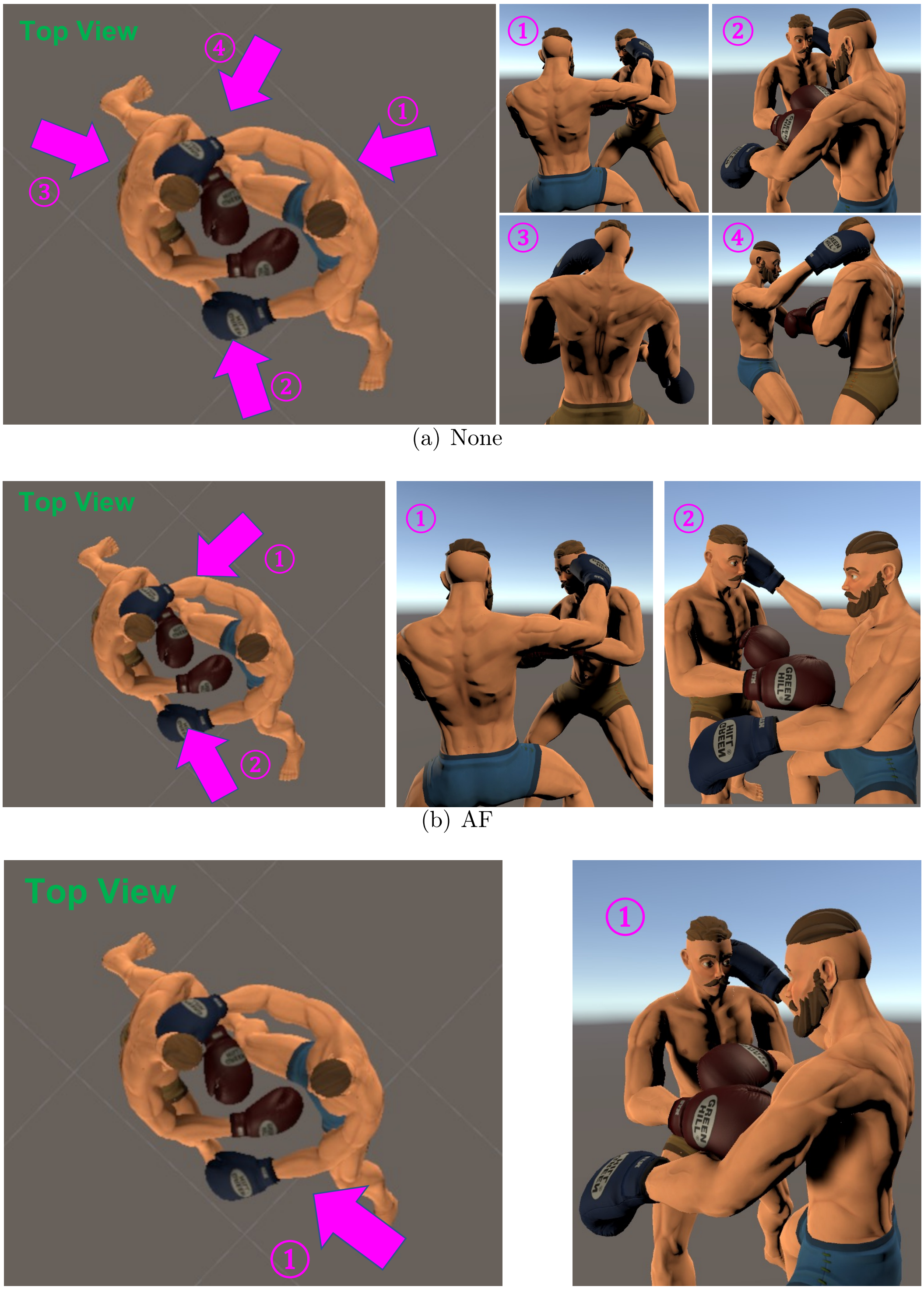
SmartVP are 1.512 and 2.083, respectively, and the p-value is 0.0033 by the Wilcox test. In Preference FE, the average moving step of SmartVP and No SmartVP are 0.917 and 1.964, and the p-value is smaller than 0.001. For the initial viewpoint participants graded (Figure 5.16 (b)), in Preference AF, the average scores of SmartVP and No SmartVP are 5.333 and 3.81, while in Preference FE, the scores of these two groups become 5.619 and 4.19. The p-values are lower than 0.001 in both preferences.

5.4.4 Discussion

From the questionnaire, although 17 participants were first-time using VR HMD, just 10 participants felt a little uncomfortable, and 1 participant felt uncomfortable in the user study experiment. It can contribute to the reasons that instead of switching viewpoints suddenly, we showed the viewpoint moving process and stopped the video replay when the viewpoint was moving.

From the data training results, it can be found that the Preference AF and FE get higher accuracy among the three tasks than the control group None. The reason can be attributed to the fact that there are usually multiple viewpoints for clearly watching a punch action. Without specific preference, the spectators are confused about selecting an optimal viewpoint, leading to various selections and the lowest accuracy (Figure 5.16.(a)). The AF task gets the middle accuracy. For the reason of Preference AF, by observation, usually more than two directions that the spectator could watch the arms form clearly (Figure 5.16.(b)) was majority to be selected, it results in lower accuracy than FE. On the other hand, for the FE task, since the spectators prefer enjoying the expression of the punched boxer, they will keep their view on the more limited angles. Just these limited angles can ensure watching the punch actions clearly and observing the expression at the same time(Figure 5.16.(c)).

From the results of the user study, hypotheses **H1** and **H2** are verified. The lower average moving steps in the SmartVP group in both preferences means the users require less time to find their preferred viewpoint when they are using the system to watch a boxing punch video. The higher score of the initial viewpoint in the SmartVP group in both preferences can conclude that the optimal viewpoints that the SmartVP system recommended are more satisfying to the users than those without SmartVP.



(a) None

(b) AF

(c) FE

Figure 5.16: The typical cases of different preferences in user study: (a) Preference None usually has multiple directions of optimal viewpoints to see the punch hitting clear; (b) Preference AF normally has two directions of optimal viewpoints to see the punch hitting and arm form clear; (c) Preference FE commonly has just only one direction of optimal viewpoint to see the punch hitting and face expression clear.

5.5 Chapter Conclusion

This work proposes a novel method to optimize the viewpoint for watching sports actions according to the different spectator preferences. For an implemented case of the boxing match in this work, a visibility model that utilizes eight collision bounding boxes is designed to calculate the visibility of upper body parts. An experiment is conducted to collect the spectator' optimal viewpoints from 24 experiment participants under three controlled preference conditions: seeing the punch Arm Form (AF), seeing the Facial Expression clearly (FE), and having no restriction (None). With the features of body parts visibility, punch side, punch offset, and the label of the participant-selected optimal viewpoints, a neural network classification model is trained. A method is introduced to evaluate this training model. The results show that this method has promising performance and can reproduce the optimal viewpoint with user preference (68.61% of accuracy for AF and 76.48% for FE). A user study is conducted to evaluate the empirical performance of SmartVP. The user study results also support that this method has promising performance in a practical application system.

5.6 Future Work

Considering the experiment time and other practical conditions, several limitations are set in this implementation, such as a 2D view ring and a specific angle of viewpoint. In the future, it is worth extending the view ring to 3D space and testing the influence of different viewpoint heights. In addition, exploring the influence of the time factor between frames in viewpoint selection also will be the main work in the next step.

Finally, the author hopes this research encourages future preference-based viewpoint selection techniques and the development of portable MR devices for use in watching sporting events both by broadcasting and live in the stadium.

Chapter 6
Conclusion

This dissertation presents an interactive visualization theory for watching rapid and occluded sports actions in VR. This theory is a primary visualization study for the case in future VR sports spectating applications. It aims to improve the user experience in spectating VR sports without temporal and spatial limitations. Two interactive visualization frameworks are proposed to solve the issues of spectating the rapid and occluded actions, respectively. They are: 1) an interactive visualization framework for rapid actions, which allows the user to control the time of VR sports video; 2) a viewpoint optimization framework depending on the individual preference for the occluded actions.

6.1 Summary

This section summarizes the contributions and findings of this work. The details are shown as follows:

Interactive Visualization Framework for Rapid Actions This work proposes a novel interactive visualization framework called MomentViz, which allows the user to freely control the time in any area of VR sports videos. Two levels of implementation are applied in this work. In the first level, a standard RGBD camera with 30FPS is utilized for the 3D data recording. Two different play speed frames of the same video are set in the same position. With the stencil technology, users can control the time in any area they like. A pilot experiment with Eight participants is conducted to verify the performance of this framework. Depending on the feedback of the pilot, the second level of MomentViz is implemented. In this implementation, a standard RGB camera and a high-speed camera are used to record the RGB and depth data. After the camera calibration, two data types can be fused to reconstruct the 3D videos. The user also can control the time with the same stencil technology. Twelve participants are recruited to join a user study evaluation. The results provide evidence that the second-level implementation with a high-speed camera improves the performance of this framework. In summary, the contributions of this work are as follows:

- This work designs a systematical framework for interactively visualizing the rapid actions in VR.

- Propose a novel method with the stencil technology to control the time in any area of VR sports videos.
- Implement two levels of this framework, including no high-speed camera and using a high-speed camera.
- A pilot is conducted to verify the performance of this method, and the results show that MomentViz outperforms the other conventional visualizations.
- A user study experiment is conducted to evaluate the implementation with a high-speed camera, and the results verify that it can improve the performance of MomentViz.

Viewpoint Optimization for Occluded Actions This work proposes a novel method called SmartVP for spectating the occluded actions in VR sports videos. This method can handle the individual preference of every spectator. A neural network classification model is designed for training the optimal viewpoint with different preferences in this method. In this model, the visibility of the boxer's upper body parts, punch side, and punch offset is extracted as features, while the spectator-selected preferred viewpoints are regarded as the label value. A boundary box visibility model is designed to account for the visibility of boxers' upper body parts. A VR experiment collects the spectators' preferred viewpoints from 24 participants under three controlled preference conditions, including seeing the punch arm form clearly (AF), seeing the facial expression clearly (FE), and no additional restriction (None). The least visibility difference (LVD) is utilized for converting the continuous values of user-selected preferred viewpoints to the discrete candidate viewpoint classes. A user study is conducted to evaluate the performance of this method, and the results show that this method can provide the optimal viewpoint with spectator preference and improve the user experience. In summary, the contributions of this work include the following:

- Design a model for training optimal viewpoints and propose an evaluation method to evaluate the training results.
- Extract the visibility of body parts, punch side and punch offset as the features, and propose a method to calculate the visibility.
- Experiment to collect the user-preferred viewpoints, including two different preferences.
- Propose the method to label the continuous user optimal viewpoint selections to discrete candidate viewpoint classes.

- Conduct a user study experiment to evaluate the performance, and the results show SmartVP can provide optimal viewpoints corresponding with individual preferences and improve the user experience.

Overall Though watching sports matches in VR provides the spectator with more immersive, the issues of rapid and occluded still exist with conventional visualization. From the pilot results discussed in chapter 3, since the sickness caused by the low FPS frames, and the difficulty of finding a viewpoint in 3D, these issues even become more severe in VR. This work which includes two parts augments the human vision for spectating the rapid and occluded sports actions in VR. The results show it has promising potential to improve the user experience of spectating sports matches in VR. It provides a new way to spectate a sports match without temporal and spatial limitations.

While several limitations of current VR devices limit the popularization speed of VR sports spectating, this work provides a fundamental and prospective study of interactive visualization for rapid and occluded actions in VR. The author hopes the findings can contribute to future VR sports spectating applications.

6.2 Future Outlook

This work focuses on the field of spectating sports in VR and provides a fundamental study of vision augmented for spectating rapid and occluded sports videos. However, this topic need not be limited to sports spectating. The author believes it also has promising potential in other fields, such as VR sports training. Compared with VR sports spectating, in VR sports training, getting and responding to users' (students) feedback is highly important. With pose estimation technology, it is possible to get the users' poses in real-time.

In the case of sports form learning, because of the different levels of students, even for one student, a beginner student who is not familiar with that form may have a different speed in the process of the entire form. It is difficult and inconvenient for the students to change the time speed of demonstration forms in learning manually. Therefore, the applications are desired to control and adjust the time speed automatically depending on individual needs. It is expected to get the users' poses and actions in real time and change the demonstration forms on time. At the same time speed between demonstration forms and users' forms can help the users understand the correct form and find their problems faster. Viewpoint Optimization is also highly required in sports form learning. Sometimes, students need help seeing the back directions of the form. This issue causes that, for that case, a teacher is necessary to observe the students' forms, which also limits the students' time and place. VR technology gives a possibility to address this problem. Similarly, with pose estimation technology, the VR system can find the users' pose error in real-time and visualize it from a good viewpoint to the students immediately. This method lets the students find the pose error faster out of sight.

Bibliography

Arthur, K. W. (2000). *Effects of field of view on performance with head-mounted displays.* The University of North Carolina at Chapel Hill.

Arzati, M. A. and Arzanpour, S. (2021). Viewpoint selection for dermdrone using deep reinforcement learning. In *2021 21st International Conference on Control, Automation and Systems (ICCAS)*, pages 544–553. IEEE.

Assa, J., Cohen-Or, D., Yeh, I.-C., and Lee, T.-Y. (2008). Motion overview of human actions. In *ACM SIGGRAPH Asia 2008 Papers*. Association for Computing Machinery.

Bao, Y., Wu, H., Zhang, T., Ramli, A. A., and Liu, X. (2016). Shooting a moving target: Motion-prediction-based transmission for 360-degree videos. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1161–1170. IEEE.

Bhatnagar, B. L., Sminchisescu, C., Theobalt, C., and Pons-Moll, G. (2020). Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision*, pages 311–329. Springer.

Brigante, C. M., Abbate, N., Basile, A., Faulisi, A. C., and Sessa, S. (2011). Towards miniaturization of a mems-based wearable motion capture system. *IEEE Transactions on industrial electronics*, 58(8):3234–3241.

Burdea, G. C. and Coiffet, P. (2003). *Virtual reality technology.* John Wiley & Sons.

Burns, E. (2022). How many people are expected to watch the world cup? `https://www.90min.com/posts/how-many-people-are-expected-to-watch-the-world-cup-2022`. Last accessed on November 28, 2022.

Cheng, W., Xu, L., Han, L., Guo, Y., and Fang, L. (2018). ihuman3d: Intelligent human body 3d reconstruction using a single flying camera. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1733–1741.

Corbillon, X., De Simone, F., Simon, G., and Frossard, P. (2018). Dynamic adaptive streaming for multi-viewpoint omnidirectional videos. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 237–249.

Córdova-Esparza, D.-M., Terven, J. R., Jiménez-Hernández, H., Vázquez-Cervantes, A., Herrera-Navarro, A.-M., and Ramírez-Pedraza, A. (2016). Multiple kinect v2 calibration. *Automatika*, 57(3):810–821.

Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S. R., Kowdle, A., Escolano, S. O., Rhemann, C., Kim, D., Taylor, J., et al. (2016). Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4):1–13.

Doubek, P. (2005). Multi-view tracking and viewpoint selection. volume 39. ETH Zurich.

Fourati, H., Manamanni, N., Afilal, L., and Handrich, Y. (2012). Complementary observer for body segments motion capturing by inertial and magnetic sensors. *IEEE/ASME transactions on Mechatronics*, 19(1):149–157.

Fu, J., Chen, Z., Chen, X., and Li, W. (2020). Sequential reinforced 360-degree video adaptive streaming with cross-user attentive network. *IEEE Transactions on Broadcasting*, 67(2):383–394.

Geršak, G., Lu, H., and Guna, J. (2020). Effect of vr technology matureness on vr sickness. *Multimedia Tools and Applications*, 79(21):14491–14507.

Goldman, A. I. (2015). *Theory of human action.* Princeton University Press.

Guerra-Filho, G. (2005). Optical motion capture: Theory and implementation. *RITA*, 12(2):61–90.

Guo, K., Xu, F., Yu, T., Liu, X., Dai, Q., and Liu, Y. (2017). Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (ToG)*, 36(4):1.

Han, S., Achar, M., Lee, S., and Peña-Mora, F. (2013). Empirical assessment of a rgb-d sensor on motion capture and action recognition for construction worker monitoring. *Visualization in Engineering*, 1(1):1–13.

Hardy, A. C. (1920). A study of the persistence of vision. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 221–224.

Hasler, N., Rosenhahn, B., Thormahlen, T., Wand, M., Gall, J., and Seidel, H.-P. (2009a). Markerless motion capture with unsynchronized moving cameras. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 224–231. IEEE.

Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., and Seidel, H.-P. (2009b). A statistical model of human pose and body shape. In *Computer graphics forum*, volume 28, pages 337–346. Wiley Online Library.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Hu, H.-N., Lin, Y.-C., Liu, M.-Y., Cheng, H.-T., Chang, Y.-J., and Sun, M. (2017). Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1396–1405. IEEE.

Huang, C., Gao, F., Pan, J., Yang, Z., Qiu, W., Chen, P., Yang, X., Shen, S., and Cheng, K.-T. (2018). Act: An autonomous drone cinematography system for action scenes. In *2018 ieee international conference on robotics and automation (icra)*, pages 7039–7046. IEEE.

Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., and Stamminger, M. (2016). Volumedeform: Real-time volumetric non-rigid reconstruction. In *European conference on computer vision*, pages 362–379. Springer.

Jarvie, G. (2013). *Sport, culture and society: an introduction*. Routledge.

Kiciroglu, S., Rhodin, H., Sinha, S. N., Salzmann, M., and Fua, P. (2020). Activemocap: Optimized viewpoint selection for active human motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 103–112.

Kim, H. G., Lim, H.-T., Lee, S., and Ro, Y. M. (2018). Vrsa net: Vr sickness assessment considering exceptional motion for 360 vr video. *IEEE transactions on image processing*, 28(4):1646–1660.

Kim, Y., Baek, S., and Bae, B.-C. (2017). Motion capture of the human body using multiple depth sensors. *ETRI Journal*, 39.

Kimm, D. and Thiel, D. V. (2015). Hand speed measurements in boxing. *Procedia Engineering*, 112:502–506.

Kirk, A., O'Brien, J. F., and Forsyth, D. A. (2004). Skeletal parameter estimation from optical motion capture data. In *ACM SIGGRAPH 2004 Sketches*, page 29.

Lee, Y. and Yoo, H. (2017). Low-cost 3d motion capture system using passive optical markers and monocular vision. *Optik*, 130:1397–1407.

Lewis, J. P., Cordner, M., and Fong, N. (2000). Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172.

Lin, J.-W., Duh, H. B.-L., Parker, D. E., Abi-Rached, H., and Furness, T. A. (2002). Effects of field of view on presence, enjoyment, memory, and simulator sickness in a virtual environment. In *Proceedings ieee virtual reality 2002*, pages 164–171. IEEE.

Liu, S., Zhang, J., Zhang, Y., and Zhu, R. (2020). A wearable motion capture device able to detect dynamic motion of human limbs. *Nature communications*, 11(1):1–12.

Mazuryk, T. and Gervautz, M. (1999). Virtual reality - history, applications, technology and future.

Meyer, J., Kuderer, M., Müller, J., and Burgard, W. (2014). Online marker labeling for fully automatic skeleton tracking in optical motion capture. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5652–5657. IEEE.

Miyata, N., Kouchi, M., Kurihara, T., and Mochimaru, M. (2004). Modeling of human hand link structure from optical motion capture data. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pages 2129–2135. IEEE.

Nakano, N., Sakura, T., Ueda, K., Omura, L., Kimura, A., Iino, Y., Fukashiro, S., and Yoshioka, S. (2020). Evaluation of 3d markerless motion capture accuracy using openpose with multiple video cameras. *Frontiers in sports and active living*, 2:50.

Napoli, A., Glass, S., Ward, C., Tucker, C., and Obeid, I. (2017). Performance analysis of a generalized motion capture system using microsoft kinect 2.0. *Biomedical Signal Processing and Control*, 38:265–280.

Newcombe, R. A., Fox, D., and Seitz, S. M. (2015). Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352.

O'Brien, J. F., Bodenheimer Jr, R. E., Brostow, G. J., and Hodgins, J. K. (1999). Automatic joint parameter estimation from magnetic motion capture data. Technical report, Georgia Institute of Technology.

Placht, S., Fürsattel, P., Mengue, E. A., Hofmann, H., Schaller, C., Balda, M., and Angelopoulou, E. (2014). Rochade: Robust checkerboard advanced detection for camera calibration. In *European conference on computer vision*, pages 766–779. Springer.

Pohlert, T. (2014). The pairwise multiple comparison of mean ranks package (pmcmr). *R package*, 27(2019):9.

Potthast, L. (2013). The decelerator helmet - a slow motion for real life. `https://www.designboom.com/technology/the-decelerator-helmet/`. Last accessed on November 28, 2022.

Rice, M. E. and Harris, G. T. (2005). Comparing effect sizes in follow-up studies: Roc area, cohen's d, and r. *Law and human behavior*, 29(5):615–620.

Rosenhahn, B., Brox, T., Kersting, U., Smith, A., Gurney, J., and Klette, R. (2006). A system for marker-less motion capture. *Künstliche Intelligenz*, 1(2006):45–51.

Rudoy, D. and Zelnik-Manor, L. (2010). Posing to the camera: Automatic viewpoint selection for human actions. In *ACCV*.

Slavcheva, M., Baust, M., Cremers, D., and Ilic, S. (2017). Killingfusion: Non-rigid 3d reconstruction without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1395.

Stanislaw, H. and Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1):137–149.

Strauß, T., Ziegler, J., and Beck, J. (2014). Calibrating multiple cameras with non-overlapping views using coded checkerboard targets. In *17th international IEEE conference on intelligent transportation systems (ITSC)*, pages 2623–2628. IEEE.

Tabachnick, B. G. and Fidell, L. S. (2007). *Experimental designs using ANOVA*, volume 724. Thomson/Brooks/Cole Belmont, CA.

Tognetti, A., Lorussi, F., Mura, G. D., Carbonaro, N., Pacelli, M., Paradiso, R., and Rossi, D. D. (2014). New generation of wearable goniometers for motion capture systems. *Journal of neuroengineering and rehabilitation*, 11(1):1–17.

Van der Kruk, E. and Reijne, M. M. (2018). Accuracy of human motion capture systems for sport applications; state-of-the-art review. *European journal of sport science*, 18(6):806–819.

Vollmer, M. and Möllmann, K.-P. (2012a). Oscillating droplets and incompressible liquids: slow-motion visualization of experiments with fluids. *Physics Education*, 47(6):664.

Vollmer, M. and Möllmann, K.-P. (2012b). Vapour pressure and adiabatic cooling from champagne: Slow-motion visualization of gas thermodynamics. *Physics Education*, 47(5):608.

Vollmer, M. and Möllmann, K.-P. (2018). Slow speed-fast motion: time-lapse recordings in physics education. *Physics Education*, 53(3):035019.

Wang, Z., Wu, W., Xu, X., and Xue, D. (2007). Recognition and location of the internal corners of planar checkerboard calibration pattern image. *Applied mathematics and computation*, 185(2):894–906.

Watanabe, Y., Ohno, H., Komuro, T., and Ishikawa, M. (2009). Synchronized video: An interface for harmonizing video with body movements. In *22nd Symposium on User Interface Software and Technology (UIST2009)(Victoria, 2009.10. 5)/Adjunct Proceedings*, pages 75–76.

Xie, L., Zhang, X., and Guo, Z. (2018). Cls: A cross-user learning based system for improving qoe in 360-degree video adaptive streaming. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 564–572.

Xu, D., Zhang, H., Wang, Q., and Bao, H. (2005). Poisson shape interpolation. In *Proceedings of the 2005 ACM symposium on Solid and physical modeling*, pages 267–274.

Yang, R. S., Chan, Y. H., Gong, R., Nguyen, M., Strozzi, A. G., Delmas, P., Gimel'farb, G., and Ababou, R. (2013). Multi-kinect scene reconstruction: Calibration and depth inconsistencies. In *2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013)*, pages 47–52. IEEE.

Yang, S., He, Y., and Zheng, X. (2019). Fovr: Attention-based vr streaming through bandwidth-limited wireless networks. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9. IEEE.

Yu, T., Guo, K., Xu, F., Dong, Y., Su, Z., Zhao, J., Li, J., Dai, Q., and Liu, Y. (2017). Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 910–919.

Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., and Liu, Y. (2018). Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7287–7296.

Zeltzer, D. (1992). Autonomy, interaction, and presence. *Presence: Teleoperators & Virtual Environments*, 1(1):127–132.

Zhang, Y., Zhao, P., Bian, K., Liu, Y., Song, L., and Li, X. (2019). Drl360: 360-degree video streaming with deep reinforcement learning. In *IEEE*

INFOCOM 2019-IEEE Conference on Computer Communications, pages 1252–1260. IEEE.

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334.

Zhao, P., Zhang, Y., Bian, K., Tuo, H., and Song, L. (2019). Laddernet: Knowledge transfer based viewpoint prediction in 360â¦ video. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1657–1661. IEEE.

Zhu, Y., Zhai, G., Yang, Y., Duan, H., Min, X., and Yang, X. (2021). Viewing behavior supported visual saliency predictor for 360 degree videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4188–4201.