Context-aware User-dependent Viewpoint Recommendation for Multi-view Field Ball Sports Videos

Xueting Wang

Abstract

As an extension of the widely used single-view video, multi-view videos by multiple cameras offer more viewpoint options. They are playing an important role in representing events in areas of education, surveillance, communication, and entertainment. To make the multi-view video technology more feasible in practice, many research works have been conducted on multi-view video capturing, compression, streaming, and delivery technologies. When multi-view video capturing and streaming are developed in practice, the representation technology on how to show the multi-view contents to users becomes an important issue. Therefore, this thesis focuses on viewpoint recommendation of multi-view videos among the multi-view video processing research targets.

In this thesis, the field ball sports content is focused as the target application, especially the soccer game, which is a typical sport game containing the common components of field ball sports.

Multi-view videos can be customized according to users' preferences, so that users can enjoy interesting content based on their own choices, instead of the fixed contents provided by professional TV broadcasters. However, there exists the problem of continual appropriate viewpoint selection, which leads to the annoyance and stress of a completely manual process if only existing multi-view video interfaces are provided to users. Thus, automatic user-dependent viewpoint recommendation is important for enhancing the advantage of multi-view video viewing.

The goal of the works presented in this thesis is to achieve automatic user-dependent viewpoint recommendation adapted to diverse video contexts and user contexts. Here, the video contexts are defined as representations of what is happening in the view, consisting of scene contexts and production contexts. The scene contexts include the size and number of visible objects, objects' spatial distribution, and scene events. The production contexts include camera location and optical axis, viewpoint switching angle and frequency. Meanwhile, users' viewpoint selection tendencies are considered as the user contexts, which are related to user personalities, interests, and experience level on the viewing content, and favorite objects. The users' viewpoint selection tendencies could be not only the common viewpoint selection tendency of most users, but also the unique tendency of a specific user or of a group of users with similar tendencies.

To effectively recommend viewpoints with high similarity to users' viewpoint selection tendencies, the recommendation task is considered as a supervised learning problem. The following factors are considered in the recommendation framework: context-dependent learning scheme, spatio-temporal feature representation, and user modelling for group-based recommendation, respectively. The effectiveness of recommendation is evaluated by the degree of similarity between the recommendation and actual users' viewpoint selection records.

This thesis is organized with the following chapters.

Chapter 1 introduces the background, motivation, challenges, and the purpose of the work presented in this thesis. Contributions of the work are also summarized in this chapter.

Chapter 2 reviews related works. An overview of related approaches on automatic video editing and recommendation, and viewpoint selection methods for multi-view videos are provided, and the differences with the proposed methods are discussed.

Chapter 3 describes the creation of a multi-view video dataset used in this thesis, together with user viewpoint selection records acquired through a multi-view video editing experiment.

Chapter 4 proposes a context-dependent learning scheme to adapt to various video contexts and users' viewpoint selection tendencies. Viewpoint evaluation and transition processes are used to represent the video contexts including scene contexts and production contexts by appearance features and transition costs of viewpoints. The proposed method uses different weight parameters to combine the appearance features and the transition costs, and optimizes them for each scene context by minimizing the difference between the generated viewpoint sequences and users' viewpoint selection records. This context-dependent learning scheme can provide both common and personal recommendation by using common or personal viewpoint selection records of users. The effectiveness of this method is confirmed by comparing the generated context-dependent and the independent viewpoint recommendations with actual selections made by users.

Chapter 5 proposes a method to improve the video context representation using spatio-temporal feature representation. It was found that different trajectory distri-

butions of focused objects cause a difference in the viewpoint selection for different users by analyzing users' viewpoint selection tendencies. Thus, the recommendation scheme is trained by learning the relationship between the users' viewpoint selection tendencies and the spatio-temporal scene context represented by object trajectories. Three methods were compared to find the best one to represent this relationship, and the Gaussian Mixture Models (GMM) based method achieved the best performance by assessing the degree of similarity between the viewpoint recommendation and user's selection records. These methods uses users' personal viewpoint selection records as the learning label, but can also be applied to common recommendation by considering common viewpoint selection records of multiple users.

Chapter 6 describes a method that focuses on the cold-start problem of user-dependent recommendation based on a user model for group recommendation. The clod-start problem here is the lack of sufficient user records for model construction. The relationship between viewpoint selection tendency and user attributes is analyzed and applied to solve the problem. A group-based recommendation framework consisting of a user grouping approach based on the similarity with existing users' viewpoint selection records, and a member group estimation approach based on the classification by user attributes are proposed in this chapter. The generated group-based recommendation yielded similar recommendation effectiveness to the personal recommendation, so it can be used when user's personal records are insufficient. The users with high emotional stability in personality and high-level experience in the target sport trend to have stable viewing pattern and receive better user-dependent viewpoint recommendations.

Finally, Chapter 7 concludes this thesis and presents directions for future work.

Acknowledgement

First and foremost I would like to express my gratitude to my supervisor, Professor Kenji Mase. His insightful comments and suggestions guided me in developing this study. In addition, his great research vision always inspired me.

I also would like to thank my thesis committee, Professor Yoshiharu Ishikawa, Professor Ichiro Ide, and Professor Takatsugu Hirayama for taking time to review my thesis and giving important comments. Their comments improved the quality of this thesis significantly.

Professor Takatsugu Hirayama was always eager to discuss this work, and these discussions were essential for writing this thesis. I also wish to express my gratitude to Dr. Yu Enokibori, Professor Junya Morita and Professor Albert Ali Salah. Their comments gave me fresh ideas because their areas of research are slightly different from mine.

I would like to offer my special thanks to all the members of the Mase laboratory, especially Dr. Kensho Hara and Mr. Fumiharu Tomiyasu. Daily discussions with them advanced my research and motivated me.

Last but not the least, I would like to thank to my family, especially my parents, for their constant support and unconditional love.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Video services have grown rapidly and become more and more important in our daily life thanks to the development of electronic and computing technology. Users can enjoy better quality and more convenient video viewing experiences, such as more than 100 TV programs, on-line live broadcasting, Youtube 360° video, etc. However, with the rapid development of the Internet streaming around the World and corresponding on-line services and applications, the traditional form of pre-prepared single-view video can not satisfy different user needs for diverse viewing contents.

Multiple-viewpoint (multi-view) videos taken by multiple cameras from different angles as shown in Figure 1.1, contain more diverse information and viewing options. Thus, multi-view videos are expected to change our daily video viewing experiences from just accepting a pre-prepared service to enjoying a more creative and interactive service. This kind of new viewing form could lead to novel and exciting applications in areas of education, surveillance, communication, and entertainment, such as better broadcasting for Web lectures, surgeon training, high quality surveillance systems [Ahmad, 2007, Wang, 2013]. Especially, for large-scale events viewing, including sports, concerts, and so on, fans could enjoy more diverse contents according to their own preferred viewing angles for preferred viewing targets.

To make the multi-view video technology more feasible in practice, many works have been conducted on multi-view video capturing, compression, streaming, and delivery technologies [Collins et al., 2002, Mase et al., 2006, Ho and Oh, 2007, Pan et al., 2011, Cheung et al., 2011, Marutani et al., 2012, Maugey and Frossard, 2013]. For example, a multi-camera capturing system called Eye Vision was used

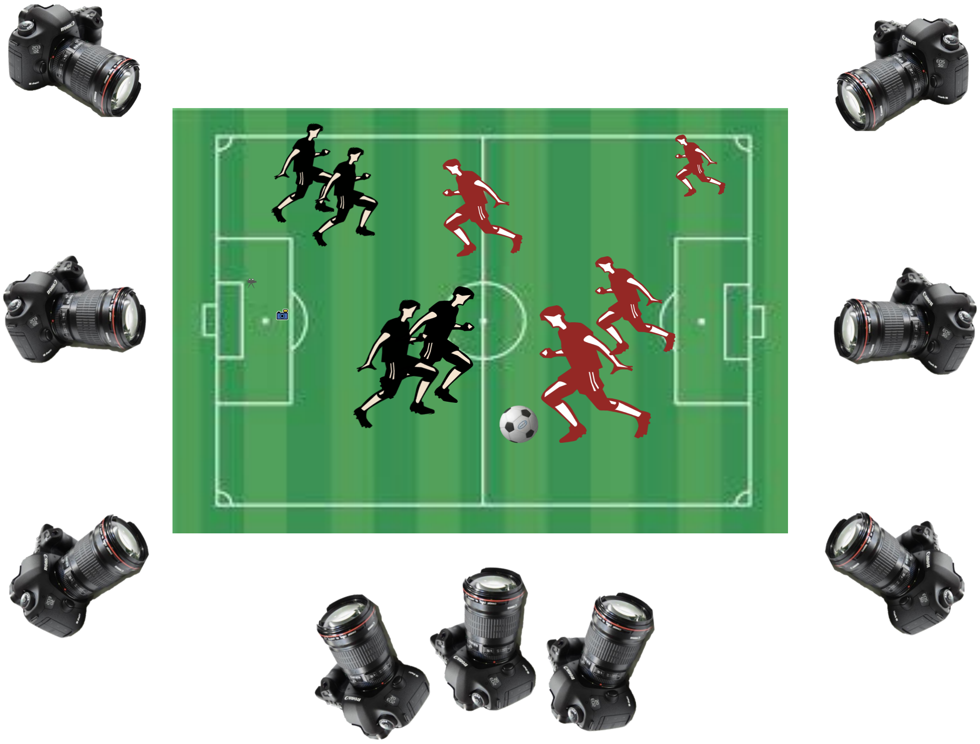


Figure 1.1: Multi-view video capturing for a soccer game.

to shoot and broadcast Super-bowl 2001 [Kanade et al., 1997]. Kurutepe et al. proposed a standards-based, flexible, end-to-end framework for the streaming of 3D representations in the form of multi-view videos [Kurutepe et al., 2007]. Furthermore, free-view videos can be generated to provide more viewpoint options by interpolating scenes, modeling 3D scenes and depth-based methods. [Debevec et al., 1998, Ohm and Müller, 1999, Kameda et al., 2004, Kim et al., 2006, Kilner et al., 2006, Horiuchi et al., 2012, Matsuyama et al., 2012]. When multi-view video capturing and streaming are developed in practice, the representation technology on how to show the multi-view contents to users becomes an important issue for multi-view services and applications, which is the focus of this thesis.

In this thesis, field ball sports including soccer, basketball, football, hockey, handball, and volleyball are considered as the target application. Sport is a hot field, which has attracted people's attention for a long time. Users' requirements also increase fast and diversely along with the development of digital media and the Internet. Many researchers have developed applications on sports video processing such as object tracking, action, or pattern analysis [Perše et al., 2009, Liu et al., 2013, Schlipsing et al., 2017, Memmert et al., 2017, Yao et al., 2017], event and highlight extraction [Rui et al., 2000, Wang et al., 2014b, Godi et al., 2017, Wang et al., 2017b], and video understanding and summarization [Ekin et al., 2003, Chen et al., 2011, Sigari et al., 2015, Chauhan et al., 2016], etc. Sports video, especially the field ball sports video, also attracts large attention from the research field of multi-view video, such as multi-camera recording for baseball games [Rui et al., 2000], multi-view 3D pose estimation of football players [Kazemi et al., 2013], and multi-view ball and player tracking [Du et al., 2006, Ren et al., 2010, Zhang et al., 2017], etc. This is because multi-view videos have an advantage on representing diverse information for the field ball sports taking place in wide fields.

There are some common components among field ball sports. A ball usually indicates the location of the current important scene, which is very often the focus of the viewer. The players' spatial-temporal information, such as their trajectories are also an important part of the game scene, which also attracts a viewer's attention. Moreover, different events usually occur in different areas of the field, so the field area information is also a factor that affects a viewer's attention. Since soccer is a typical ball sport containing these components, and is targeted in many works in the field of video processing, the proposed method is applied to multi-view soccer game videos in this thesis.

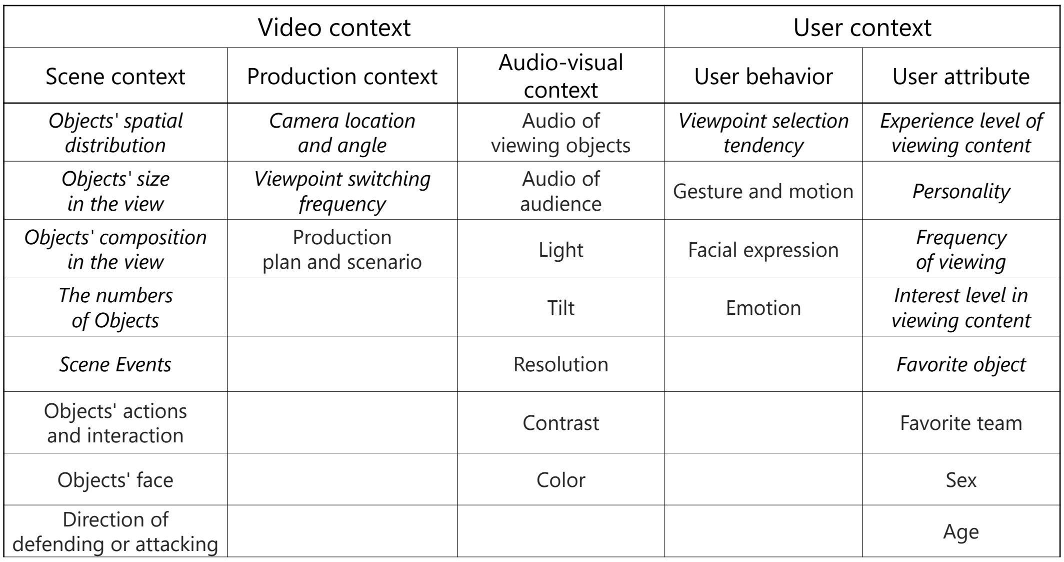
1.2 Motivation and Challenges

The research focus of this thesis is to provide a user-dependent multi-view video representation technology.

Several viewing interfaces exist, which allow users to freely choose their preferred viewpoints. Lou et al. provided a system, which allows users to select their desired viewing directions and exciting visual experiences interactively in real time [Lou et al., 2005]. Tokai et al. developed a target-centered viewing method across different viewpoints named Pegged Scope [Tokai et al., 2008]. These interfaces enhance users' viewing experiences as the enhanced viewing interfaces are more flexible and interactive than a single-view video. However, when using these existing multi-view video interfaces, completely manual and continual selection of appropriate viewpoints for a long time could be stressful and annoying, especially when the number of selection options increases. Thus, viewing support such as automatic viewpoint sequence recommendation is necessary to enhance the user's viewing experience and convenience for multi-view video viewing. Especially, viewpoint recommendation with a high-quality representation of the viewing event and high adaptation to different users is important to extend the advantage of the multi-view video viewing.

Some works on automatic viewpoint selection have been developed based on predefined rules devised from common sense or professional knowledge. For example, in [Shen et al., 2007], best-view selection involves a content analysis stage that assigns a score to each viewpoint based on Quality of View (QoV). The QoV of each viewpoint is evaluated by fixed rules based on common sense, such as the larger the focused object in the view is, the higher the score will be. Zhang et al. applied cinematography rules for on-line lecture video capturing such as the minimum and maximum duration of a shot [Zhang et al., 2008]. They also applied a rule that a shot should be switched to the person who asks the question. Saini et al. proposed a framework for the automatic mashup of dance performance videos taken by mobile phones by choosing the best angle based on professional editing rules, through learning, for example, the shot duration from the directors' viewpoint transition [Saini et al., 2012]. These works can provide recommendation to the users who are used to common sense or professional knowledge. However, when users have the opportunity to select the viewpoints themselves, they tend to select the best viewpoint to enjoy the event with their own viewing pattern. In this thesis, such user-dependent viewpoint recommendation is considered.

A user's viewing pattern for multi-view videos varies according to diverse video contexts and user contexts. Table 1.1 summarizes different kinds of contextual information in video and user contexts. The video contexts are the representations of what is happening in the view, including scene contexts, production contexts, and audio/visual contexts. The scene contexts include the size, composition and the number of visible objects, objects' spatial distribution, objects' actions and interaction, the direction of defending or attacking, and events occurring during the game etc. The production contexts include video production information such as the camera location and angle, the viewpoint switching frequency, and the video production plan and scenario. Moreover, the light condition, image quality (resolution, contrast, color), and audio information are also parts of video contexts. For user contexts, a user's behavior while viewing and the user's attribute information are included. User behavior includes a user's viewpoint selection tendency, gesture, motion, facial expression, and emotion while viewing. User attributes include basic information of users such as sex, age, and personality. Furthermore, a user's interests, experience and professional level on the viewing target, and favorite objects are also factors, which can influence their viewpoint selection tendencies.



Video context			User context	
Scene context	Production context	Audio-visual context	User behavior	User attribute
Objects' spatial distribution	*Camera location and angle*	Audio of viewing objects	*Viewpoint selection tendency*	*Experience level of viewing content*
Objects' size in the view	*Viewpoint switching frequency*	Audio of audience	Gesture and motion	*Personality*
Objects' composition in the view	Production plan and scenario	Light	Facial expression	*Frequency of viewing*
The numbers of Objects		Tilt	Emotion	*Interest level in viewing content*
Scene Events		Resolution		*Favorite object*
Objects' actions and interaction		Contrast		Favorite team
Objects' face		Color		Sex
Direction of defending or attacking				Age

Table 1.1: Video contexts and user contexts in multi-view video viewing. The contextual information considered in this thesis are indicated in red, focusing on video contexts including scene and production contexts, and user contexts including user's viewpoint selection tendency and user attributes.

The video and user contexts related to field ball sports indicated in italic in Table 1.1 are focused in this thesis. Since the ball is the focus of a game and players move accordingly, a large amount of contextual information is implicitly encoded in ball and player locations on the game field. Thus, the works presented in this thesis make use of video contexts, including object-based scene contexts and production contexts such as viewpoint switching angle and frequency. For the user contexts, the works presented in this thesis make use of users' viewpoint selection tendencies, which can be represented by their viewpoint selection records, and is considered related to user personalities, interests, and experience or professional level of the viewing sport.

The goal of the work presented in this thesis is to achieve automatic user-dependent viewpoint recommendation adapted to diverse video contexts and user contexts, which is important for the user-dependent multi-view viewpoint recommendation. The main challenge is how to represent the various video contexts, and adaptively relate them to corresponding user's viewpoint selection tendency. Firstly, the recommendation scheme needs a high quality representation of video contexts to understand and evaluate what is happening in the view of viewpoints. Simultaneously, it should be able to acquire and predict users' tendency on viewpoint selection. The users' viewpoint selection tendency here could be not only the common viewpoint selection tendency of most users, but also the unique tendency of a specific user. Furthermore, we should consider the situation that some users have a similar viewpoint selection tendency while some do not. The group of users with similar viewpoint selection tendencies are also an important recommending target. This kind of recommendation can be used in the situations that users are not satisfied with common recommendation, and personal recommendation is not available due to insufficient personal data.

In the following, the challenges focused on to realize context-aware user-dependent viewpoint recommendation are given.

Context Adaption: Fixed evaluation rules of viewpoint quality are inadequate to represent different viewpoint selection tendencies on diverse video contexts and different users. For example, the viewpoint selection tendencies of users of a soccer game depend on the location where a game event occurs, e.g. the ball. If it is a scene where the ball is located in the penalty area, users might choose viewpoints that zoom into the ball and switch viewpoints to view it from various directions. In another case, they may prefer a wide and stable shot if the ball is in the mid-field. Thus, an adaptation scheme to various video and user contexts is necessary.

Spatial Feature Representation: The spatial information of the focused object, such as the locations of a ball and players are the important part to represent a game event. The recommendation result could be limited if only the scale of objects, such as the number of humans visible in the view, is considered. For example, the camera view will most likely focus on the center-field even when the ball and the front player approach the goal, but are far away from most other players.

Temporal Feature Representation: Users select their preferred viewpoints not only considering the current contextual information, but also the understanding of what occurred in the past and predicting what will possibly occur in the short future. Thus, recommendations that only consider the current contextual information will cause inappropriate selection or frequent viewpoint switching.

Cold-Start: In the area of recommendation systems, user-dependent recommendation, especially personal recommendation is usually learned or optimized from a user's own data or history. However, in practice, it is difficult to acquire sufficient amount of personal data for model construction. In general, this is called the "*cold-start problem*", and the user-dependent multi-view viewpoint recommendation is not an exception.

User Analysis: The analysis on "what kind of user has what kind of viewpoint selection tendency" is important for the user-dependent recommendation. The quality of experience of an application or a service is with respect to the utility and/or enjoyment in light of the user's personality and current state [Le Callet et al., 2012]. Thus, here, a user's viewpoint selection tendency and the satisfaction on the recommendation are assumed to be related to their attributes including personality, interest, and experience level of the viewing content. Understanding and applying the relationship between such user attributes and viewpoint selection tendency might lead to better and practical viewpoint recommendation.

Reflecting the focus and challenges of the task, the following research issues are considered:

1. Is an adaptive recommendation scheme more effective than a fixed scheme to reflect different users' viewpoint selection tendencies for different video contexts?

2. Is the spatio-temporal video context representation effective to enhance the recommendation performance?

3. Is it possible to find a similar user group to apply collaborative filtering to overcome the *cold-start* problem in user-dependent recommendation?

4. What kinds of user attributes are related to viewpoint selection tendencies and can be used for the viewpoint recommendation?

1.3 Contributions

Here, to realize context-aware user-dependent viewpoint recommendation, the recommendation is considered as a supervised learning problem, as shown in Figure 1.2. Features from the spatio-temporal contextual information of the video are extracted and used for its representation, and an appropriate viewpoint is predicted through a context-dependent learning scheme. Users' viewpoint selection tendency is reflected in the recommendation by using their viewpoint selection records as the learning labels. The kinds of user records used as the learning labels can lead to different types of recommendations, including common recommendation for most of the users, personal recommendation for a specific user, or recommendations for different groups of users.

To generate a viewpoint recommendation with high similarity to users' viewpoint selection tendency, several factors are considered in the recommendation framework:

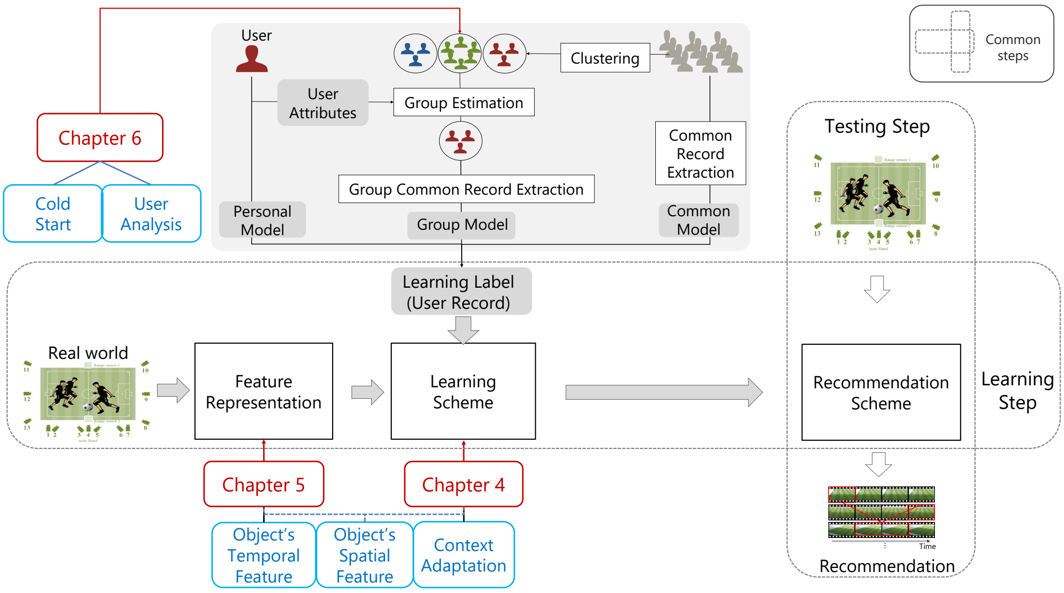


Figure 1.2: Outline of the recommendation framework including the learning step and the testing step. The red boxes indicate the relation between the structure of this thesis and the corresponding focused components in the flow. The blue boxes indicate the relation between the structure of this thesis and the target challenges.

context-dependent learning scheme, *spatio-temporal feature representation*, and *user model for group-based recommendation*. The effectiveness of recommendation is evaluated by the degree of similarity between the recommendation and actual user's viewpoint selection records.

This proposed framework makes the following contributions:

- The proposed context-dependent recommendation scheme provides a more effective and user-dependent viewpoint recommendation than a fixed scheme for multi-view videos.

- The proposed spatio-temporal video context representation improves the concordance between generated viewpoint recommendations and viewpoint selections made by humans.

- The proposed group-based recommendation provides practical and effective user-dependent recommendation to overcome the *cold-start* problem even for a new user with no viewpoint selection record.

- For the group-based recommendation, the relationship between personality and profile attributes of users and their viewpoint selection tendencies are investigated and modeled.

1.4 Structure of the Thesis

The structure of this thesis is as follows. Figure 1.2 shows the relation between the thesis structure and the challenges discussed.

Chapter 2 reviews related works. An overview of related approaches on automatic video editing, summarization, recommendation, and viewpoint selection methods for multi-view videos are provided.

Chapter 3 describes the creation of a multi-view video dataset used in this thesis, and users' viewpoint selection records acquired through mutli-view video editing experiments.

Chapter 4 presents an adaptive approach to increase the correspondence between users' viewpoint selection tendencies and various video contexts by a *context-dependent learning scheme*. Viewpoint evaluation and transition processes are used to represent video contexts including the scene contexts and the production contexts. First, in the evaluation process, a camera agent is designed to evaluate the view quality of each viewpoint. A producer agent determines the most appropriate sequence of viewpoint transition with the minimum transition cost. In these two processes, the proposed method combines the appearance features and calculates the transition cost using different weight parameters. The weight parameters are optimized to represent different viewpoint selection tendencies corresponding to various video contexts by minimizing the difference between the generated viewpoint sequences and the user's viewpoint selection records. This framework can provide a common recommendation when the optimization is performed based on the common viewpoint selection records extracted from a majority of users. Meanwhile, it can also be performed based on personal viewpoint selection tendency, so the possibility of personal recommendation is also investigated. Furthermore, this parameter-based scheme provides a possibility for users to control the representation of their preference by changing the pattern of parameters adapted to various video contexts. The generated context-dependent recommendation outperformed the context-independent recommendations compared with actual selections made by users, which shows the effectiveness of this method.

Chapter 5 introduces a method to improve the video context representation by *spatio-temporal feature representation*. Whether different trajectory distributions of focused objects cause a difference in the viewpoint selection for different users is investigated by analyzing a user's viewpoint selection tendency. Thus, the recommendation scheme is trained by learning the relationship between the user's viewpoint selection tendency and the spatio-temporal scene context represented by object trajectories. Three methods based on Gaussian Mixture Models (GMM), Support Vector Machines (SVM) with a 2D histogram, and SVM with a Bag-of-Words (BoW) are compared to seek for the best one to represent this relationship. The experimental results showed that the GMM-based method outperforms other methods by assessing the degree of similarity between the viewpoint recommendation and user's selection records. This method focuses on a user's personal viewpoint selection tendency, but can also be applied to recommendation for common tendency when referring to common viewpoint selection records of many users as the learning label.

Chapter 6 presents a method to overcome the *cold-start* problem of user-dependent recommendation by generating and applying a *user model for group-based recommendation*. Existing user-dependent viewpoint recommendation methods are developed by learning the user viewpoint selection records, which have difficulty in acquiring sufficient personal viewing records in practice. Moreover, they neglect the importance of the user's attribute information, such as the user's personality, interest, and experience level in the viewing content. Thus, a group-based recommendation framework consisting of a user grouping approach based on the similarity in existing users' viewpoint selection records, and a member group estimation approach based on the classification by user attributes are proposed. The generated group-based recommendation from user's member group yields better recommendation effectiveness than the one from other groups, which shows the effectiveness of group-based recommendation. The users with high emotional stability in personality and high-level experience in the target sport, have a high possibility of having consistent viewing patterns and receive better user-dependent viewpoint recommendations.

Chapter 7 concludes this thesis and presents future directions of this work.

Chapter 2

Related Work

In this chapter, first, existing works on automatic video editing and recommendation are introduced. Next, existing works on automatic viewpoint selection for multi-view videos are introduced and the differences with the proposed methods are discussed.

2.1 Automatic Video Editing

In this section, two types of related works on automatic video editing are separately introduced: automatic editing and summarization of existing videos, and automatic camera control.

2.1.1 Automatic Editing and Summarization of Existing Videos

Video editing is the work to find optimal shots from multiple video sources and concatenate them to generate a video sequence. Some works approach this task by referring to metadata. Davenport et al. created a home-movie editing assistance system using metadata such as time code, camera position, and voice annotation, to select shots, sounds, and text trunks, and concatenate them for cinematic storytelling [Davenport et al., 1991]. Kumano et al. proposed an intelligent support system by automatically extracting metadata such as shot size, camera work, human face direction, walking direction, and eye direction, and applied video grammars to concatenate appropriate shots into final videos to introduce a restaurant for cooking shows in [Kumano et al., 2002]. Here, video grammars are special rules for video editing. Hua et al. presented a system that automates home video editing [Hua et al., 2003]. This system extracts a set of highlight segments from a set of raw home videos. It aligns the selected segments based on editing rules and matches them to the content of the incidental music.

Video summarization is also an important technology. There are two basic forms on generating a short summary of a video: the key-frame sequence [DeMenthon et al., 1998, Orriols and Binefa, 2001, Li et al., 2005] and the video skim [Smith and Kanade, 1998, Babaguchi, 2000, Sundaram et al., 2002, Gong and Liu, 2003, Xu et al., 2005, Shao et al., 2006]. The key-frame sequence collects salient frames from the video source. The video skim extracts the most representative video segments from the video source. For more detailed survey on summarization methods, (please) refer to [Truong and Venkatesh, 2007].

Multi-view video summarization is conducted to explore the content correlations among multi-view video sources and select the most representative shots for generating a summary video. Bocconi et al. proposed an automatic short video documentaries maker, called Vox Populi, to generate video sequences from news following a logical line of argumentation and drawing a conclusion [Bocconi et al., 2005]. The proposed framework enhances video summarization by allowing the combination of different news stories into one coherent explanation about a topic of the current news. Fu et al. formulated the multi-view video summarization problem as a spatio-temporal shot graph labeling task by emphasizing useful information and precluding redundancy among video features [Fu et al., 2010]. The summarization is generated through solving a multi-objective optimization problem using shot importance evaluated based on brightness difference and conspicuous noises. Nack and Ide developed a framework to facilitate the journalist with a summarised audio-visual video by combining different news stories into one coherent explanation about a topic of the current news [Nack and Ide, 2011]. The framework exploits data in form of polls for story outline development; retrieves relevant materials by a combination of event templates and summarization over topic threads; and generates the final video by applying a set of trimming rules. Mase et al. proposed a ubiquitous experience media system to record interactions among multiple presenters and visitors by multiple video cameras and wearable sensors in an exhibition room [Mase et al., 2006]. They extracted appropriate scenes from the viewpoints of individual users by clustering events having spatial and temporal relationships and provided a summary video for a quick overview of the events that the users experienced.

2.1.2 Automatic Camera Control

There are also works on digital camera control including pan, tilt, and zoom of one fixed camera and automatic camera control on active cameras.

For digital camera work, Ariki et al. generated soccer videos by automatically recognizing soccer game events based on player and ball tracking [Ariki et al., 2006, Ariki et al., 2008]. They cut out a specific region from a fixed camera view and moved it according to the position of the ball and players to synthesize zoom and pan from existing videos. Gandhi et al. simulated Pan-Tilt-Zoom (PTZ) camera movements within the frame of a single static camera to generate multiple clips suitable for video editing [Gandhi et al., 2014]. Hayashi et al. proposed an automatic cooking video authoring method to explain and describe the cooking operations by reducing both temporal and spatial redundancies from a video stream recorded by a fixed camera according to the description of cooking instructions in a cooking recipe [Hayashi et al., 2013].

There are also many works on multi-camera control and coordination of PTZ cameras and smart cameras [Kim and Wolf, 2010, Esterle et al., 2014]. For tasks in surveillance and machine vision applications, these active cameras need to be controlled and coordinated optimally to perform the desired tasks including object tracking, detection, recognition and activity analysis. Collins et al. proposed an active camera system for acquiring multi-view video of a person by a real-time tracking algorithm that adjusts the pan, tilt, zoom, and focus parameters of multiple active cameras to keep the moving person centered in each view [Collins et al., 2002]. Bramberger et al. approached multi-camera tracking tasks by applying a multi-agent system for traffic surveillance, which uses a distributed embedded system with limited resources that consists of smart cameras [Bramberger et al., 2005]. Lai et al. controlled PTZ cameras to maximize the video quality of the tracking targets by a game-theoretic solution based on a linear production game [Lai et al., 2010]. Li et al. applied the auction mechanism from economics to achieve multi-camera tracking and the active camera control (pan and tilt) by modeling the camera bids with prior knowledge of the camera homographies [Li and Bhanu, 2012].

For explanation and story-telling video generation, Onishi and Fukunaga et al. proposed a computer-controlled camera work to control camera pan/tilt/zoom and find the best camera from multiple camera images to shoot a suitable lecture video [Onishi and Fukunaga, 2004]. The shooting area is directed based on lecturer's face, eye, and hand features, and blackboard's features extracted from the image. Ozeki et al. built a prototype pan-tilt camera control system to record an explanation video from multiple cameras for desktop manipulation [Ozeki et al., 2001]. They conducted the camerawork to explain the manipulation by tracking different targets according to different showing purposes.

2.1.3 Discussion

For the video editing and summarization of multiple existing videos, the task of video processing is to search or summarize the important scenes for event extraction or storytelling. Thus, the main challenge of these works is to detect highlights from the video source or analyze the location and action of an object, and to concatenate different videos from various sources into a summary video. Meanwhile, multi-view video viewpoint recommendation focuses on finding the best representation from different angles at the same timing and concatenate them into a smoothly switched video along the time line, which is different from finding the highlights across the video duration.

On the other hand, automatic camera control focuses on the tasks of object tracking, activity recognition, and behavior understanding. It consider camera parameter optimization or camera coordination strategy to generate a recommended view for tasks such as finding the best camera scheduling to track one focused target. The proposed method in this thesis focuses on fixed multi-camera systems, which can provide more options of different views with different shooting areas for various contents and targets than just the views tracking the focused object. Besides, methods using common sense or professional editing rules could be insufficient to be applied for user-dependent recommendation, since they do not consider various user preferences for different video contexts.

2.2 Automatic Video Recommendation

In this section, general recommendation systems and automatic video recommendation are introduced.

2.2.1 Recommendation Systems

There are typically two schemes for recommendation systems: collaborative filtering and content-based filtering. The collaborative filtering approaches are used to predict or rate items with users' interests by analyzing their behaviors, activities, or preferences and their similarity to other users [Hill et al., 1995, Resnick and Varian, 1997]. Content-based filtering approaches consider item characteristics to recommend additional items with similar item properties [Mooney and Roy, 2000]. Both collaborative and content-based filtering can be cast as learning problems to learn a function to describe the characteristics of a user and an artifact, and predict the user's preferences on the artifact. Basu et al. formalized the movie recommendation problem as a learning problem that takes a user and a movie as input and produces a liked or disliked label of user as output [Basu et al., 1998].

However, so-called a *cold-start* problem imposes these approaches to acquire large amount of user information to make accurate recommendations. The *cold-start* problem can be overcome by adopting a hybrid approach combining content-based filtering and collaborative filtering; New items would be assigned a rating based on the rating assigned by the users to other similar items, where the similarity is determined according to the items' content-based characteristics [Schein et al., 2002]. Moreover, the personality characteristics of a user also become important to generate personalized recommendation and overcome the *cold-start* problem, because they contain valuable information that enables systems to better understand users' preferences [Tkalcic and Chen, 2015]. The personality of a user has been used to improve user-similarity calculation in the *cold-start* problem [Hu and Pu, 2010]. Furthermore, recommendations for a group of users can be enhanced by personality-based group modeling [Quijano-Sanchez et al., 2010].

2.2.2 Video Recommendation

Video recommendation is one important application of recommendation systems since it is not an easy task for users to find a video of interest from a huge number of videos in line with the development of Internet services. Since on-line video recommendation applications usually target on an individual user, the quality of experience of the user on the video is important to evaluate its performance. Therefore, personalized video recommendation attracts much attention to correspond to diverse needs and interests of users.

There are many works on enhancing the correspondence between videos and users. Chen et al. proposed a user-video tripartite graph for personalized video search. They combined the clicks and queries information by integrating the click-through information with a user-query graph indicating if a user ever issues a query, and a query-video graph indicating if a video appears in the search result of a query [Chen et al., 2012]. Cui et al. improved the performance of recommendation in social network by converting the video recommendation problem into a similarity matching problem in a common space, where both users and videos are represented with social attributes and content attributes [Cui et al., 2014]. They learned such a common representation through a REgularized Dual-fActor Regression (REDAR) method based on matrix factorization to flexibly integrate social attributes and content attributes. Yang et al. presented an on-line video recommendation system to find a list of the most relevant videos based on multimodal fusion and relevance feedback consisting of video content and related information such as query, title, tags, and surroundings [Yang et al., 2007].

2.2.3 Discussion

The multi-view viewpoint recommendation problem can also be regarded as one kind of application of recommendation systems; The viewpoint is the recommendation item, the rating can be reflected by the viewpoint selection tendency of users. For user-dependent viewpoint recommendation, a machine learning method is proposed to enhance the correspondence between the videos and users in the thesis. It learns a user's viewpoint selection tendency based on context analysis on viewing content and their viewpoint selection records as well as content-based filtering. Moreover, to overcome the *cold-start* problem in the machine learning framework, the proposed method combines with collaborative filtering that learns from similar users for a group-based recommendation even when sufficient personal records are not available. Furthermore, considering the importance and effect of a user's personality and profile information on recommendation, the user similarity is estimated by relating the users' viewpoint selection tendency with user profile and personality.

2.3 Multi-view Viewpoint Recommendation

Table 2.1 summarizes the approaches of related works on multi-view viewpoint selection and recommendation. Here, they are grouped according to two factors: video context and user context. For the video context, those using object based feature for context representation are introduced first, followed by other kinds of multi-modal context information. For the user context, they are divided into two groups by whether they are user-dependent or not. Note that the proposed methods focus on user-dependent recommendation taking an object-based video context representation. The characteristics of each approach are discussed in the following

			User context				
			User-independent		User-dependent		
			Common sense	Professional editing rules	Common tendency	Group tendency	Personal tendency
Video context	Object based feature	Object spatial feature	[Cutler et al., 2002] [Shen et al., 2007] [Jiang et al., 2008] [Daniyal et al., 2010] [Daniyal et al., 2011] [Wang et al., 2013]	[Cai et al. 1999] [Morioka et al. 2008] [Zhang et al. 2008] [Ranjan et al.2010] [li et al., 2011] [Wang et al., 2013] [Chen et al., 2013]	[Mase et al.,2011] [Chen et al., 2013] *[Wang et al., 2015]*	*[Wang et al., 2017a]*	[Muramatsu et al., 2014] *[Wang et al., 2015]* *[Wang et al., 2016]* [Tomiyasu et al., 2016]
		Object temporal feature	[Jiang et al., 2008] [Daniyal et al., 2011]			*[Wang et al., 2017a]*	[Muramatsu et al., 2014] *[Wang et al., 2016]*
		Single object	[Cutler et al., 2002] [Shen et al., 2007] [Wang et al., 2013]	[Cai et al. 1999] [Morioka et al. 2008] [Ranjan et al.2010] [li et al., 2011] [Wang et al., 2013]			
		Multiple objects	[Jiang et al., 2008] [Daniyal et al., 2010] [Daniyal et al., 2011]	[Zhang et al. 2008] [Chen et al., 2013]	[Mase et al.,2011] [Hirayama et al. 2013] [Chen et al., 2013] *[Wang et al., 2015]*	*[Wang et al., 2017a]*	*[Wang et al., 2015]* [Tomiyasu et al., 2016] [Wang et al., 2016]
	Multi-modal feature	Audio feature	[Cutler et al., 2002]	[Zhang et al. 2008] [Ranjan et al.2010] [Saini et al., 2012] [Wang et al., 2013]			
		Visual feature		[Saini et al., 2012]			

Table 2.1: Related works on viewpoint recommendation for multi-view videos. The proposed methods in this thesis are indicated in italic, focusing on user-dependent recommendation taking the object spatio-temporal context adaptation representation.

sections.

2.3.1 Works that Consider Video Context

Object-based Features

First, let us review works that select viewpoints mainly by considering objects' spatio-temporal features.

Shen et al. proposed a best-view selection method by detailed content analysis based on Quality of View (QoV) [Shen et al., 2007] considering object location, face orientation, and the size of an object in the view for a surveillance system. Jiang et al. optimized multiple objects tracking and formulated the best-view video synthesis as a recursive decision problem according to the trajectories of objects. [Jiang et al., 2008]. Daniyal et al. presented a novel evaluation algorithm based on individual spatial features of the object, such as the location of each object and combined the evaluation scores of multiple objects to select the best viewpoint [Daniyal et al., 2010]. An extensive work by Daniyal and Cavallaro presented an algorithm to minimize the number of inter-camera switches [Daniyal and Cavallaro, 2011]. Chen et al. proposed a method for automatic viewpoint suggestion by analyzing features of a group of objects such as the visible number of objects [Chen et al., 2013]. Muramatsu et al. selected viewpoints by using the average of object features, such as location, distance to the camera and the size in the view during a short time [Muramatsu et al., 2014].

Multi-modal Features

Next, let us review works that select viewpoints by considering multi-modal (audio-visual) information.

Some works focus on audio-visual object tracking. Cutler et al. used multiple devices including a 360° camera, a whiteboard camera, an overview camera, and a microphone array to provide a rich experience for an on-line remote meeting service by audio-based speaker detection and vision-based human tracking [Cutler et al., 2002]. Zhang et al. presented an end-to-end on-line lecture capturing and broadcasting system including audio and video information based on automated speaker/audience capturing and tracking [Zhang et al., 2008]. Ranjan et al. also conducted video recording from multiple cameras for on-line lectures and meetings, which used speaker's face detection and audio-visual human tracking [Ranjan et al., 2010]. The audio information is difficult to use for viewpoint selection due to smaller auditory differences among camera positions and also the noise of the crowd for some sports played on a wide-scale field.

There are also some works that focus on the visual feature of a video including lighting, resolution, contrast, color, and tilt information. Saini et al. proposed a framework for the automatic mashup of dance performance videos taken by mobile phones. They chose the best angle based on video quality factors such as illumination, shakiness, and tilt information [Saini et al., 2012]. Wang et al. proposed an approach towards real-time control, selection, and transmission of the best view of human faces in a Skype video conferencing [Wang et al., 2013]. They applied a proportional, integral, and derivative feedback-based centralized multi-camera control mechanism to track and select the best view of a person by static cameras.

Discussion

For viewpoint selection from multi-view field ball sports videos, we can assume that the object based contextual information, such as locations of the ball and players on the field, is more reliable and effective to explain what is happening in a game.

Besides, the viewpoint evaluation capability and transition mechanism in the works introduced above are insufficient to adapt to various scene contexts and user contexts, which is quite important for user-dependent viewpoint sequence recommendation.

In addition, these approaches process at the frame-level or without sufficient representation on past and future object dynamics. We can assume that the video context can be described better by the object's trajectory information, and that machine-learning representations are more effective for user-dependent representation.

Thus, the proposed methods focus on object-based video context representation including a context-dependent learning scheme and a spatio-temporal feature representation adaptive to various user contexts in the recommendation framework.

2.3.2 Works that Consider User Context

User-independent Recommendation

Most of the related works on multi-view viewpoint selection focus mainly on rules based on common sense or professional knowledge.

There are many works that focus on common sense. In the work of Cutler et al., camera selection was conducted based on face trajectory for on-line lectures and meetings with fixed rules according to common sense such as a better front face to show the viewing target [Cutler et al., 2002]. Daniyal et al. presented an algorithm for viewpoint-quality ranking based on object scoring by common sense such that it gives better score in proportion to the size of the players in a basketball game [Daniyal et al., 2010].

Meanwhile, some other works considered professional knowledge and rules. Ranjan et al. applied professional rules from television production to improve the production quality of a meeting video, such as maintaining consistent left/right orientation via the 180° axis [Ranjan et al., 2010]. Zhang et al. used cinematography rules such as the system should give a shot to the person who asks the question, and the minimum and maximum duration of a shot for on-line lecture video capturing

[Zhang et al., 2008]. Saini et al. chose the best angle based on professional editing rules, such as learning the shot duration from the directors' viewpoint transition for dance performance videos taken by mobile phones [Saini et al., 2012]. Chen et al. generated candidate viewpoints recommendation by learning the director's styles to apply professional editing knowledge [Chen et al., 2013]. There are also works on multi-view camera selection for object tracking, which consider the knowledge and rules for tracking such as "hand-off rules". The hand-off rules are used to select a camera for a certain object and hand-off this object from one camera to another. Cai and Aggarwal proposed an automatic camera selection mechanism consisting of a prediction phase of determining when to switch the camera and a selection phase based on the view of the human subject to be tracked with robust spatial matching [Cai and Aggarwal, 1999]. Morioka et al. presented a fuzzy-based camera selection strategy for tracking objects by selecting the appropriate camera for hand-off [Morioka et al., 2008]. In their work, the state transitions of fuzzy automaton are driven by transition rules that are based on the previous camera state and tracking level in a monitoring area. Li and Bhanu presented a comparison of geometry, statistics, and game-theory based approaches for centralized and distributed camera selection and hand-off mechanisms [Li and Bhanu, 2011].

User-dependent Recommendation

There are only a few works on user-dependent viewpoint recommendation.

First, as for works that focus on the selection tendency of most users, Mase et al., proposed a viewpoint recommendation method that selects the most popular viewpoint among all users based on the viewing history of users [Mase et al., 2011]. Hirayama et al. proposed an agent-assisted multi-view video viewer that incorporates target-centered viewpoint switching, which helps to fix the user's gaze on a target object [Hirayama et al., 2013]. They also introduced a social viewpoint recommendation method based on higher view-visit frequencies of multiple users. In addition, for the personal recommendation, object's features such as position, distance to the camera, and object size in the view are learned and weight parameters for feature combination are optimized adaptive to each user's viewpoint selection records [Muramatsu et al., 2014, Tomiyasu and Mase, 2015].

Discussion

The works introduced as user-independent recommendation generate generic recommendations related to the viewing target or professional recommendations like TV programs, while they neglect users' diversity and personality. For the user-dependent recommendation, they generate common recommendations according to the common tendency of the majority of users and also personal recommendations. However, it might be difficult to employ in practice or their performance could be limited when the users' viewpoint selection records are not sufficient for training the model. Even in these user-dependent works, they seldom conduct user analysis such as detailed analysis on the relation between users' viewpoint selection tendencies and their profile or personal attributes, which is highly related to the quality of experience of user-dependent recommendations.

Therefore, this thesis focuses on the user-dependent viewpoint recommendation using user-adaptive approaches. It not only considers the common viewpoint selection tendency of most users and personal tendency discussed in Chapters 4 and 5 [Wang et al., 2015, Wang et al., 2016], respectively, but also considers the possibility to apply collaborative filtering to solve the *cold-start* problem by extracting a user group with similar viewpoint selection tendency in Chapter 6 [Wang et al., 2017a].

Chapter 3

Data Acquisition

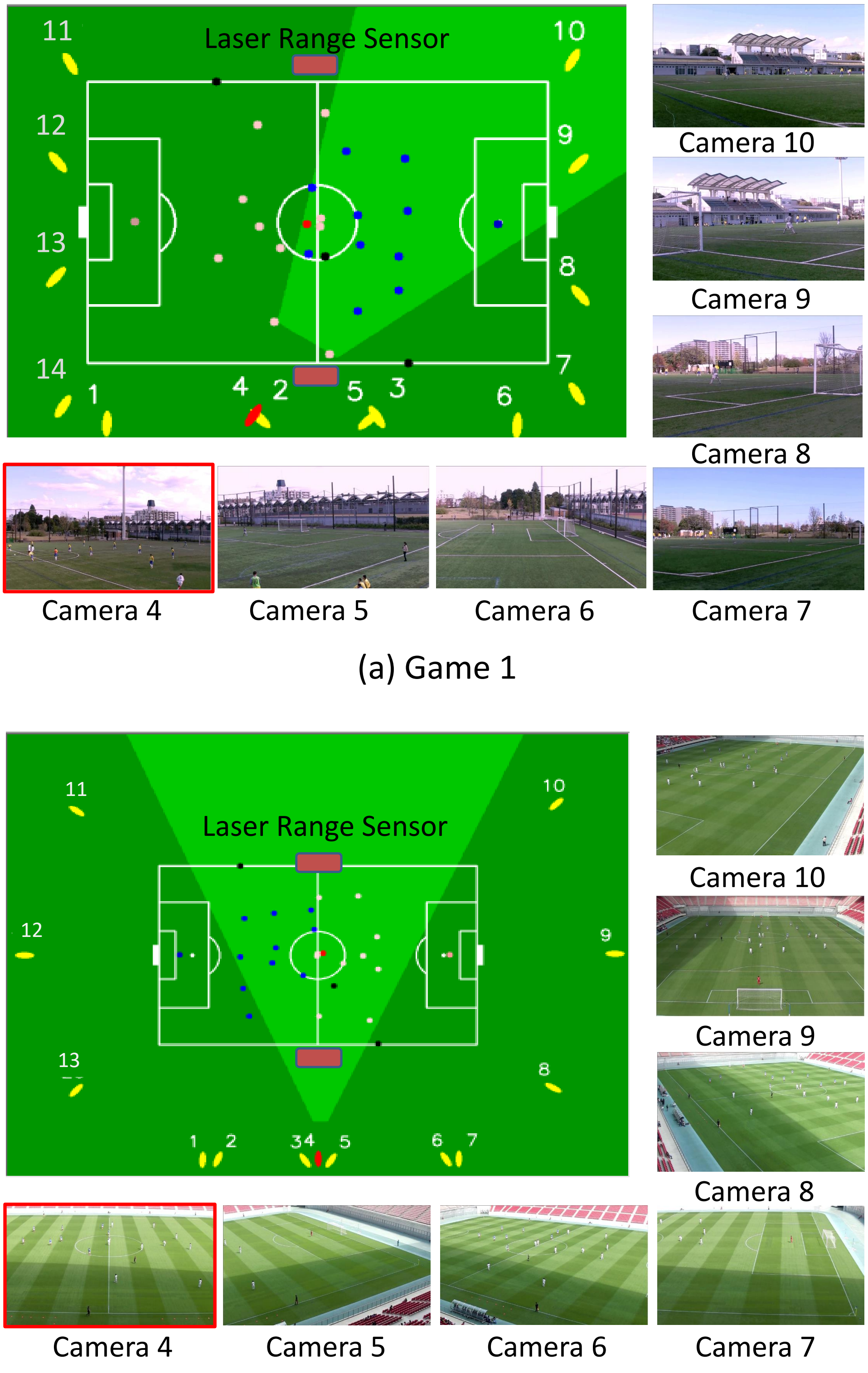
To achieve and verify the effectiveness of the user-dependent viewpoint recommendation methods proposed in this thesis, users' viewpoint selection records were collected through a video editing experiment using actual multi-view video datasets recording soccer games. This chapter introduces the details of the common components used in the following chapters including the dataset of multi-view videos and object tracking data in Section 3.1, and the details of multi-view video editing experiments to acquire users' viewpoint selection records in Section 3.2.

3.1 Dataset Creation

3.1.1 Multi-view Video Filming

Multi-view videos of two soccer games were filmed as source materials. They were two soccer games held in different venues with different camera settings. One game was played in the Akabane Sports Stadium in 2011 (Game 1). Another was played at Toyota Stadium in 2013 (Game 2). The former was a practice match within a top class high school soccer club. The latter was an annual friendship match between two university teams.

In Game 1, cameras 1 to 6 were set in the main stand, and the others were set outside the edge of the field. In Game 2, all the camera were placed on the audience stand. Both games were filmed using digital cameras (CASIO EX-F1, at 30 fps and 1,920 × 1,080 pixels) with no pan, tilt, nor zoom functions. Only the cameras near the main stand were used as shown in Figure 3.1 because radical changes that occur when the viewpoint transfers from one side of the field to another can cause



(a) Game 1

(b) Game 2

Figure 3.1: Camera positions and sample viewpoint images of two soccer games.

Table 3.1: Length of the presented video sequences.

Game 1			Game 2		
Sequence Number	Length		Sequence Number	Length	
	[frame]	[s]		[frame]	[s]
1	811	27.0	1	901	30.0
2	526	17.5	2	901	30.0
3	401	13.3	3	1,201	40.0
4	461	15.4	4	901	30.0
5	1,051	35.0	5	901	30.0
6	1,051	35.0	6	1,021	34.0
7	781	26.0	7	1,081	36.0
8	1,081	36.0	8	1,021	34.0
9	961	32.0	9	1,061	35.0
10	1,261	42.0	10	961	32.0
11	1,171	39.0			
Total	9,556	318.5	Total	9,950	331.7

cognitive discomfort. Figure 3.1 includes sample images of cameras placed on the right half of the field. The camera IDs are given counterclockwise.

Eleven short video sequences[1] from Game 1 and ten from Game 2 that contained typical soccer scenes were excerpted and presented to subject users of an editing experiment, as shown in Table 3.1. The cameras were manually synchronized after filmed. The average lengths of the video sequences were 28.9 and 33.1 seconds, and their standard deviations were 9.8 and 3.3 seconds, respectively.

These video sequences contained typical play scenes to attract users' interests as shown in Table 3.2, for example, scenes of dribble (players sparse/dense), passing (short/long passing), sliding, shooting, cross, throw-in, heading, body check, goal-kick, and free-kick. Note that since corner-kick and penalty-kick were not included in the two games, the performance of the proposed method might be limited in such situations.

[1] In this thesis, the following terms are used to refer to video content;
Sequence: a short-length video excerpted from the entire footage.
Cut: several frames segmented from the video sequence. The continuous sequence between two viewpoint switching by a user in a video sequence is one kind of a cut. Another kind of a cut with fixed length by a sliding window method is introduced in Chapter 5.
Frame: the minimum unit of a video sequence.

Sequence No.	dribble		passing		sliding	shooting	cross	throw-in	heading	body check	free kick	goal kick
	players sparse	players dense	short passing	long passing								
Game 1												
1		✓	✓	✓					✓	✓		
2		✓	✓		✓					✓		
3	✓	✓	✓	✓	✓					✓		
4	✓	✓	✓		✓	✓						
5	✓	✓	✓	✓								✓
6	✓	✓	✓	✓	✓							
7	✓	✓	✓	✓		✓	✓					
8		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
9	✓	✓	✓		✓		✓		✓	✓		
10	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	
11		✓	✓	✓	✓		✓	✓	✓			
Game 2												
1		✓	✓	✓	✓			✓		✓		
2			✓		✓			✓	✓	✓		
3	✓	✓	✓	✓	✓		✓	✓	✓	✓		
4		✓	✓		✓			✓		✓		
5			✓	✓				✓		✓		
6	✓		✓	✓	✓			✓	✓	✓		
7		✓	✓	✓	✓					✓	✓	
8	✓		✓		✓				✓	✓		✓
9		✓	✓		✓			✓	✓	✓	✓	
10	✓		✓	✓	✓				✓	✓		

Table 3.2: Details of the soccer scenes/events in the sequences.

3.1.2 Object Trajectories

The 2D locations of the players on the field were obtained by two laser range sensors (SICK LMS511, Operating range: 65 m, Field of view: 185°, Angular resolution: 0.17°, Scanning frequency: 25 Hz). They were set on both sides of the field at the height of 90 cm when filming the multi-view videos, as shown in Figure 3.2. The range sensor emits an infrared beam in a semicircular area and acquires the location of an object on the measurement plane from the returned signal based on the time of light. Here, the data of two range sensors were combined to obtain the precise location data of players in wide field of the soccer games [Kabeya et al., 2016].

However, we cannot acquire the correspondence between the location data and the players. Therefore, here, the correspondence between the location data and the players were manually given in the initial time, and each player was tracked by selecting the points closest to the previous timing. If the location was not acquired by the sensors or a mis-tracking occurred, the data were complemented by linear interpolation.

Since the ball is too small, the ball location cannot be acquired by the range sensor. Thus, the ball location was acquired through semi-automatic processing.

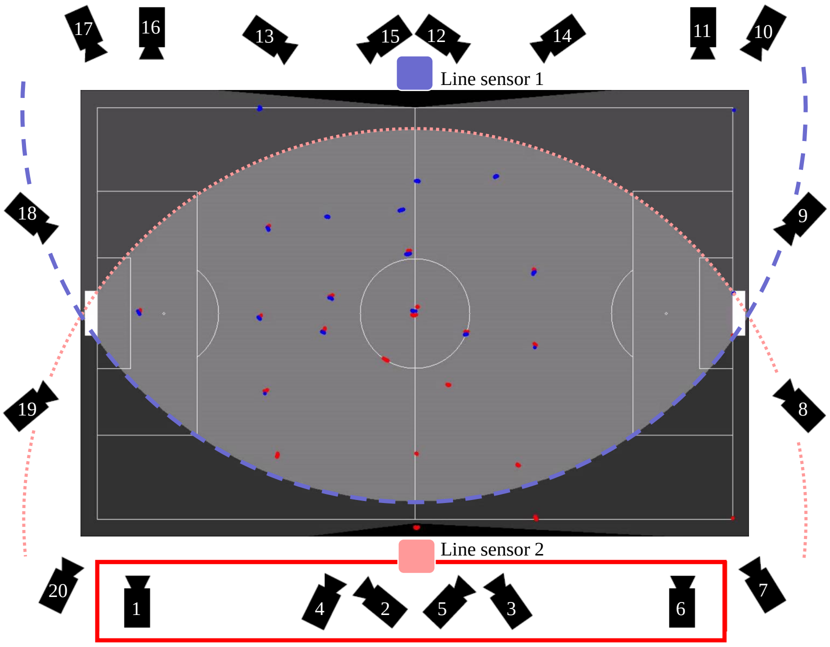


Figure 3.2: Location of the range sensors set on both sides of the field.

Concretely, frames when the ball bounces from the ground or is touched by the players were manually labeled. Then the location of the player involved in the play was considered as the current ball position. Considering a case when a player touches the ball in the air, the position of the feet of the player was used. Hence, the ball positions were given in a 2D coordinate system. For the ball position among player touch, some vision-based and sensor-based tracking techniques are being researched separately for this purpose [D'Orazio and Leo, 2010], but a basic interpolation procedure between key-frames was used since the main focus here is the viewpoint navigation but not automatic ball tracking.

3.1.3 Camera Information

In the method proposed in Chapter 4, the position (location and angle) of each camera is necessary to calculate measures such as the distance between the camera and the ball. Thus, the correspondence between the field coordinate system and the image coordinate system needs to be determined. For this, the homography matrix which can represent the correspondence is used. Homography is to transform a certain projective space to a different space with a projective transformation. Here, it is calculated manually using four or more corresponding points (e.g. line intersections

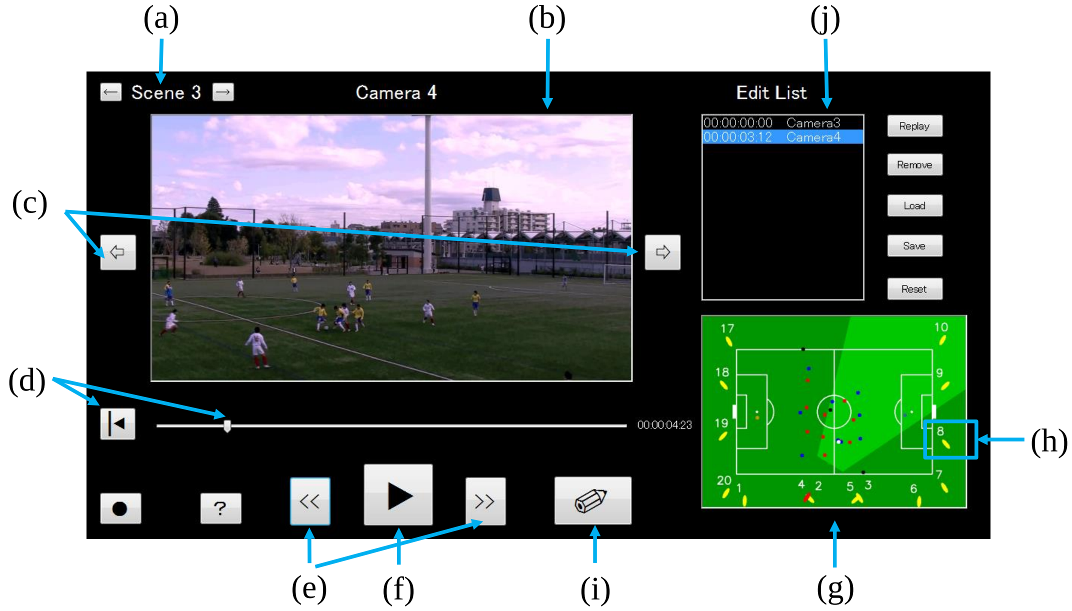


Figure 3.3: Interface for the Multi-view video editing experiment.

on the field) on the soccer field plane and the image plane of each viewpoint video. Thus, a homography matrix is acquired for each viewpoint. A point on the image coordinate system can be calculated using the corresponding homography matrix with the point on the field coordinate system. Accordingly, camera positions on the field coordinate system can be calculated.

3.2 Viewpoint Selection Records Acquisition

This section introduces a multi-view video editing experiment to collect user viewpoint selection records using multi-view video dataset described in Section 3.1.

3.2.1 Multi-view Video Editing Interface

For the multi-view video editing experiment, a user interface shown in Figure 3.3 was implemented. The functions of this interface indicated in the figure are listed as follows:

(a) Sequence ID number: Randomly chosen from the game video in the experiment

(b) Video pane: Shows the selected viewpoint

(c) Viewpoint switch (Left/Right): Switch buttons to left/right adjacent viewpoints

(d) Time-line show control: Back to start

(e) Non-linear time-line control: Step forward/backward (1 frame)

(f) Time-line control: Play/Pause

(g) Bird's-eye view of the location of players/ball/cameras/shooting areas of cameras

(h) Direct viewpoint switch: Clickable map of cameras (viewpoints)

(i) Recording button of the chosen viewpoint and timing

(j) List of recorded viewpoints and timings

3.2.2 Editing Experiment Procedure

In the video editing experiment, the participants were instructed to watch the multiview sequences and to edit viewpoint sequences following their interests by choosing the most suitable viewpoint at every time frame.

After the introduction of the experiment and the interface, they were given 5 minutes to practice. Since it appeared that the participants spent approximately 10 minutes to edit a sequence lasting 30 seconds in a preliminary experiment, short video sequences that contain typical soccer scenes were chosen as described in Section 3.1.1 to facilitate the task. The video sequences were presented randomly in the experiment. Participants were allowed to replay forward/backward the sequences arbitrarily while choosing various viewpoints. When they confirmed a viewpoint, it was recorded with the timing. The participants were instructed to write down their intention after editing each sequence. Additionally, after the experiment, they were asked to fill out a post-questionnaire that asked the important issues when choosing the appropriate viewpoints in multiple-choice questions and free format.

3.2.3 Viewpoint Selection Records of Users

Multi-view video editing experiments were conducted twice. At the first time, ten subjects participated, which are used in Chapters 4 and 5. An additional nineteen

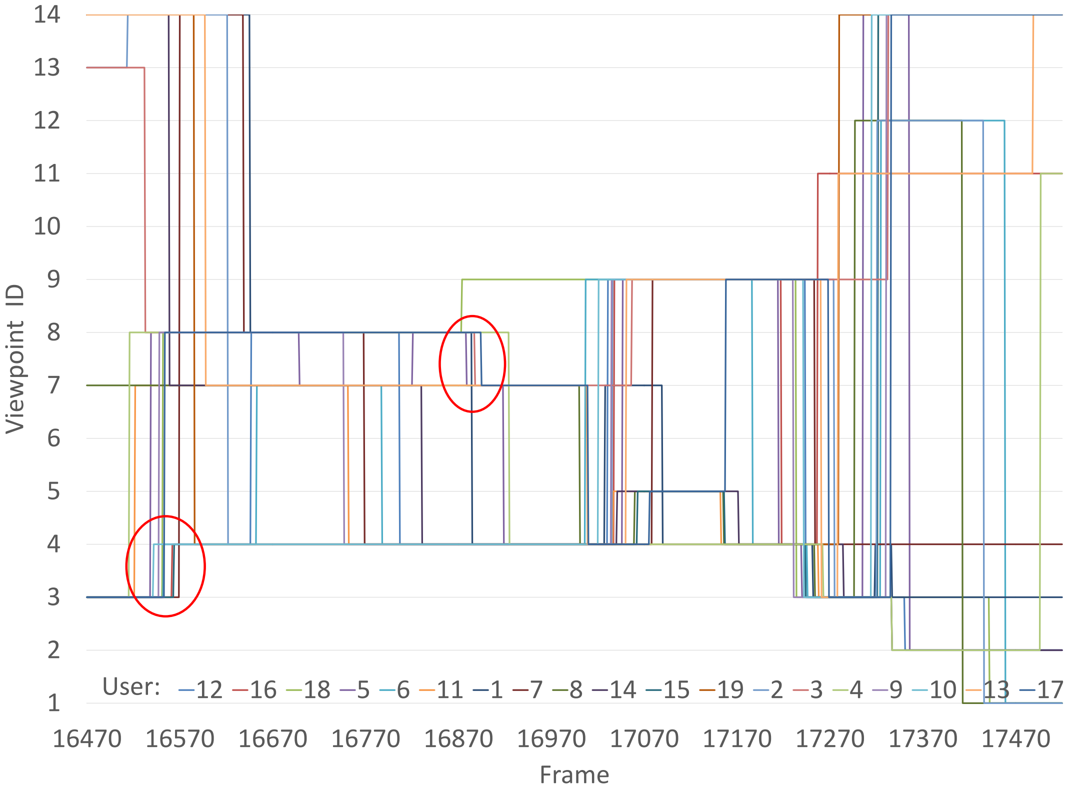


Figure 3.4: Example of selected viewpoint sequences by nineteen participants. Sometimes, most of the participants made a common selection, The difference on the timing of viewpoint switches was often confined to a short time-range as indicated by red circles. Besides, the choices were diverse from 17,270 to 17,570 frames, which showed different personal preferences.

participants and their personality and profile information were collected to conduct the group-based recommendation in Chapter 6 at the second time. Through the comparison of the records of all participants, as shown in Figure 3.4, we can see that the record of each participant reflects his/her personal preference, which can be used for personal preference modeling. Furthermore, we can see that sometimes most participants made a common selection. Although the timing of viewpoint switches differed for each participant, they were often confined to a short time-range of approximately 1 second. Thus, it seems that there is a common selection preference among participants, by connecting the viewpoint most frequently chosen by most participants for each experimental sequence, it could be used for common preference modeling.

Chapter 4

Context-dependent Learning Scheme

4.1 Introduction

As introduced in Chapter 1, the viewpoint recommendation problem is converted to a learning problem as shown in the outline of the thesis (Figure 1.2). In the common steps of the outline, the acquired multi-view and object tracking data are represented into features to learn collected users' viewpoint selection records through a learning scheme. This chapter focuses on the part of learning scheme that it proposes a context-dependent learning scheme to adapt to various video contexts and user contexts for multi-view viewpoint recommendation.

The generation of an optimal viewpoint sequence would require us to first consider contextual information, including scene information such as the location and size of focused objects in the view, and events occurring during the game. Other contextual information includes video production information such as the camera location and angle, and camera switching frequency. In this chapter, contextual information of this nature is referred as the scene context and the production context, respectively. As a more detailed example of a scene context, the location of the focus of the game, e.g., the position of the ball, affects the viewpoint selection tendencies of users. If the scene involves a ball located in the penalty area, most users might choose viewpoints that provide close-views of the ball from different angles. On the other hand, a user might prefer a stable shot with a wide angle if the ball is in the mid field. The proposed method uses this information for parameter optimization.

Therefore, context-dependent viewpoint recommendation corresponding to users' viewpoint selection tendency is important. The proposed method focuses on the contextual information of scene and production to carry out viewpoint evaluation and transition processes. First, the viewpoint evaluation process consists of designing a camera agent to evaluate the view quality of each viewpoint. This is performed by considering the contextual information of the scene as represented by the appearance features of objects, e.g., position or distance from the camera, in the given production context, depending on the user's object preference. Next, during a viewpoint transition process, such as when switching between camera views, the producer agent determines the most appropriate sequence of viewpoint transitions with the minimum transition cost, e.g. smoothness of transition, by taking into account the view quality and production contextual information such as multi-video editing rules. These two processes combine the appearance features and calculate the transition cost using different weight parameters, which are optimized to represent different viewpoint selection tendencies corresponding to both scene and production contexts.

In this chapter, several kinds of contextual divisions are tested, which are then compared with a recommendation by context-independent parameter optimization to analyze the effectiveness of the proposed context-dependent recommendation. First, optimization is performed based on the common viewpoint sequences extracted from multiple-users' viewpoint selection records, as the goal. Next, the possibility of using personalized recommendation based on personalized optimization is examined by experimenting with individual viewpoint selection records.

The remainder of this chapter is organized as follows. Section 4.2 reviews related works. Section 4.3 presents the structure of the proposed method for viewpoint sequence recommendation using scene and production contexts, which are detailed and interpreted in Section 4.4. In Section 4.5 the weight parameter optimization procedure and different context divisions are introduced. Section 4.6 introduces the experiment. Experimental results are analyzed in Section 4.7. Section 4.8 summarizes this chapter.

4.2 Related Work

There are several works on video context representation. Jiang et al. optimized multiple object tracking and formulated best-view video synthesis as a recursive decision problem [Jiang et al., 2008]. In this work, viewpoint switching frequently occurs since it does not consider the production contextual information on viewpoint transition, such as the relative positions of cameras. Such a system would produce short and uncomfortable camera switching. Shen et al. proposed a best-view selection method using a detailed content analysis based on Quality of View (QoV), which is a confidence measure for view performance evaluation by considering the view angle and distance from subjects [Shen et al., 2007]. Daniyal and Cavallaro presented an algorithm for viewpoint sequence generation focusing on minimizing the number of inter-camera switches [Daniyal and Cavallaro, 2011]. However, they used the same criteria for content analysis and viewpoint transition arrangement of all durations regardless of users' different viewpoint selection tendencies for scene contexts of different durations. Hence, such a method may select a viewpoint presenting a more detailed view, whereas the user may prefer an overview as a mid field game scene.

Therefore, the viewpoint evaluation capabilities and transition mechanisms in the studies mentioned above are not sufficiently adaptive to various scene contexts and user viewpoint selection tendencies, that are actually quite important for user-dependent viewpoint sequence recommendation for different video contexts.

4.3 Procedure of Viewpoint Recommendation

This section explains the procedure for the proposed video sequence recommendation method [Wang et al., 2014a]. The proposed method consists of two processes: evaluation and transition. The outline of the method is shown in Figure 4.1.

4.3.1 Viewpoint Evaluation Process

For each viewpoint, a camera agent is assigned, which is responsible for assessing the view quality in the evaluation process by considering the scene contextual information of each viewpoint within the production context and the user's viewpoint selection tendency. According to the QoV measure [Shen et al., 2007], view quality is strongly related to, and can be evaluated based on, the appearance of focused objects in a scene. Here, the following four appearance measures are used to evaluate every object in the video frame: in-frame measure f_e, proximity measure f_p, composition measure f_c, and context-related location measure f_l. The viewpoint evaluation process involves the combined evaluation of all objects by considering the number and importance of objects. Weight parameters for the combination are defined so as to represent the relative importance of each of these measures, and the combined set of parameters is optimized using production contexts including user's viewpoint selection tendency.

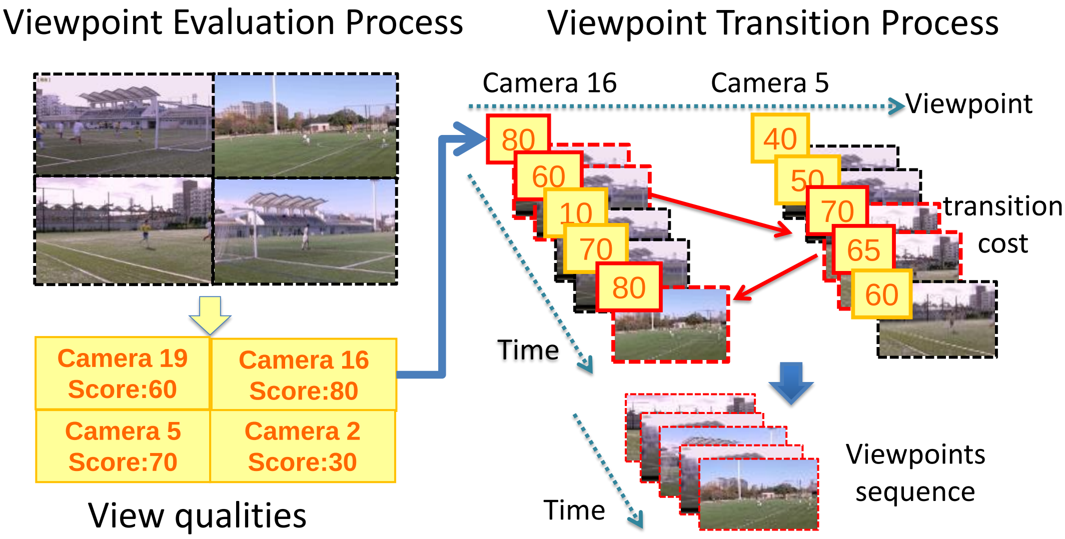


Figure 4.1: Outline of the proposed viewpoint sequence recommendation method through viewpoint evaluation and transition processes. The transition cost is minimized in terms of the duration of each viewpoint, the temporal view quality change, and the visual angle difference.

In the following functions, $p_i^w = (x_i, y_i, z_i)$ is used to represent the field position of the i-th camera (or the i-th viewpoint p_i) in the world coordinate system and $o^w = (x, y, z)$ for the field position of object O. $o_i = (u_i, v_i)$ is the position of the object in the image coordinate system in the i-th camera view. The evaluation steps are as follows:

1. Evaluation of a single object O, in a video frame: The view quality of a single object is calculated by the appearance characteristics including visibility, apparent size, composition, and location.

   The appearance measures are defined as follows,

   - In-frame existence measure f_e: The visibility of the object in the Field Of View (FOV) is defined as follows:

$$f_e(o^w, p_i^w) = \begin{cases} 1 & \text{(object in view)}, \\ 0 & \text{(otherwise)}. \end{cases} \tag{4.1}$$

   - Proximity measure f_p: The inverse distance from camera p_i to object O in an event space which is promotional to the apparent size of the object in the image as follows:

$$f_p(o^w, p_i^w) = \begin{cases} 1 - \frac{d_w(o^w, p_i^w)}{D_f} & (d_w(o^w, p_i^w) < D_f), \\ 0 & \text{(otherwise)}, \end{cases} \tag{4.2}$$

where $d_w(o^w, p_i^w)$ is a 2D Euclidean distance function between object O and the i-th camera on the field. This measure is normalized by a constant D_f, which represents the length of the diagonal of the field.

- Composition measure f_c: The composition of the object in the view.

  The visual attention that an object attracts depends on its position in an image, with objects in the center of an image known to attract more attention. Moreover, according to "the rule of thirds" [Peterson, 2011], which is widely used in visual arts, important compositional elements should be placed at intersections of lines dividing the image into nine (= 3 × 3) equal parts. Therefore, here, the average distance between the object and the center point of the image e_c, and the lower two intersections, e_l and e_r, which are normalized by the length of the diagonal line D of the image, is calculated as follows,

$$f_c(o^w, p_i^w) = 1 - \frac{d(o_i, e_c) + d(o_i, e_l) + d(o_i, e_r)}{3D}. \tag{4.3}$$

  where $d(o_i, e)$ is a 2D Euclidean distance function of object O, and point e in the i-th camera image.

- Scene context-related location measure f_l: The relevant importance between the location of the object and the scene context.

  The field is divided into several non-overlapping areas and each area is assigned a score $f_l \in [0, 1]$ that varies according to the scene context. In case of regarding the ball as an important scene descriptor, a detailed example of the measure and scene context in Figure 4.2 shows the score distribution of the location measure corresponding to the scene context of the left penalty area containing the ball. The f_l of a player O at the far side of the field is assigned a low value to represent his/her low relevance to the current penalty area scene context.

In the end, the view quality of object O in camera p_i is calculated by combining

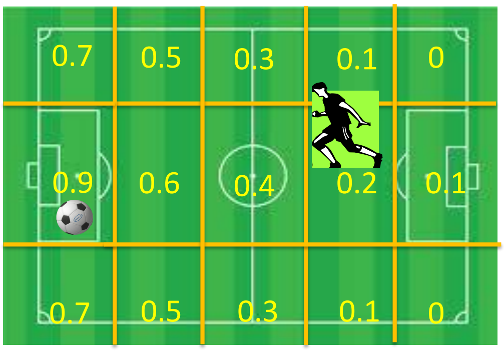


Figure 4.2: Example of the score distribution of the location measure f_l for the left penalty area scene context. The location measure scores are decided empirically.

these measures with weights $\{\omega_p, \omega_c, \omega_l\}$ as follows,

$$v_s(o^w, p_i^w) = f_e(o^w, p_i^w)(\omega_p f_p(o^w, p_i^w) + \omega_c f_c(o^w, p_i^w) + \omega_l f_l(o^w, p_i^w)), \tag{4.4}$$

where $\omega_p + \omega_c + \omega_l = 1$.

2. Evaluation of multiple objects:

   The camera agent evaluates multiple objects by considering both the average view quality evaluation and their count number using weight ω_n as follows,

   $$v(O, p_i^w) = \omega_n |O| + (1 - \omega_n)\frac{1}{|O|}\sum_{o_k^w \in O} v_s(o_k^w, p_i^w), \tag{4.5}$$

   where O is the set of targeted objects $o_k (1 \leq k \leq |O|)$, and $|O|$ is the number objects in O.

3. Overall value of the viewpoint:

   The camera agent generates an overall evaluation for viewpoint p_i by assigning different weights to the relevant main objects O_m and other sub-objects O_s, where weight ω_m represents the preference level of O_m against O_s as follows:

   $$V(O, p_i^w) = \omega_m v(O_m, p_i^w) + (1 - \omega_m) v(O_s, p_i^w). \tag{4.6}$$

   Note that the main object is chosen by the user.

Through the above steps, the camera agents determine the overall view quality of a video frame image from the i-th viewpoint and submit them to the producer agent for further processing.

4.3.2 Viewpoint Transition Process

The producer agent determines the optimal sequence of viewpoints in this process. Here, this problem is solved by calculating the transition cost based on the view quality evaluated by camera agents and production rules, e.g, the duration of the viewpoint and the camera transition loads with different weights. The weights vary with different viewpoint selection tendencies according to scene contexts and production contexts based on a user's viewpoint selection tendency. The viewpoint transition is restricted if the cost is too high. $V(O, p_i^w)$ is a function of time frame, t. In the following, the i-th viewpoint at time t is represented by $p_i(t)$.

- Duration cost of previous viewpoint: In order to avoid frequent switching, the agent calculates the contiguous frame number $T(p_i(t-1))$ of the same viewpoint $p_i(t-1)$ until the previous frame $t-1$ as follows,

  $$C_d(p_i(t-1)) = 1 - \frac{T(p_i(t-1))}{T_{\text{ave}}}, \tag{4.7}$$

  where T_{ave} is a term for normalization and represents the average duration of frames for which all users in the experiment contiguously select viewpoints. The shorter the duration preceding the viewpoint, the higher the cost of switching to another viewpoint becomes.

- Change cost of temporal view quality: The cost is defined as the difference in evaluation scores between the previous viewpoint $p_i(t-1)$ and the candidate viewpoint $p_c(t)$ of the current frame as follows,

  $$C_q(O, p_i(t-1), p_c(t)) \quad = \quad -(V(O, p_c^w(t)) - V(O, p_i^w(t-1))). \tag{4.8}$$

  In contrast to the duration cost, the higher the change cost is, the lower the switching cost to candidate $p_c(t)$ becomes.

- Difference cost of visual angles: To avoid the dizziness caused by overwhelming change in position of the camera, two factors are taken into account: the change in the view angles of the cameras, f_a, and the change in the location,

f_n, between the previous and the candidate viewpoint as follows,

$$C_a(p_i(t-1), p_c(t)) \quad = \quad f_a(p_i(t-1), p_c(t)) + f_n(p_i(t-1), p_c(t)). \tag{4.9}$$

The larger the difference cost is, the higher the switching cost to candidate $p_c(t)$ becomes.

In the end, the cost function for the current viewpoint is determined by combining these three factors as follows,

$$\begin{aligned} \mathrm{Cost}(O, p_i(t-1), p_c(t)) &= \omega_d C_d(p_i(t-1)) \\ &+ \omega_q C_q(O, p_i(t-1), p_c(t)) \\ &+ \omega_a C_a(p_i(t-1), p_c(t)), \end{aligned} \tag{4.10}$$

where $\omega_d + \omega_q + \omega_a = 1$.

To find the best viewpoint for object set O, the producer agent calculates the cost functions of each candidate viewpoint and chooses one $p_i(t)$ with the minimum cost and lower than a pre-defined threshold as defined below. Note that when the minimum cost is above the threshold, the previous viewpoint remains selected.

$$\mathrm{Cost}_{\mathrm{min}}(O, p_i(t-1), p_c(t)) \quad = \quad \min\{\mathrm{Cost}(O, p_i(t-1), p_c(t))\}, \tag{4.11}$$

$$p_i(t) \quad = \quad \begin{cases} p_c(t) & (\mathrm{Cost}_{\mathrm{min}} < \mathrm{threshold}), \\ p_i(t-1) & (\mathrm{otherwise}), \end{cases} \tag{4.12}$$

where the notation of object is omitted for simplification. The producer agent repeats this process for each frame and connects the viewpoints across frames to generate the recommended sequence.

4.4 Context Adaptation

It is considered that users' viewpoint selection tendencies differ based on their preference for various scene contexts including the location of the main object, and production contexts including the switching frequency. Therefore, let's assume that there are differently-weighted sets for the measures and costs defined in the previous sections adapted to various video contexts and user contexts.

First, the adaption to video contexts including scene and production contexts are introduced. The user survey conducted in the experiment detailed in Section 4.6, indicated that viewers of a soccer game considered the ball as the most important object in it. Hence, here, the ball is considered as an important scene context descriptor and its influence is quantified based on its position to model the weight set for viewpoint evaluation and transition processes. The field is divided into several regions, and they are grouped into several areas representing different characteristic contexts. Thus, a video sequence can be represented as a series of scene contexts in terms of ball positions and its influential regions. Moreover, we can generate various patterns of viewpoint by assigning corresponding weight parameters $\{\omega_p,\omega_c,\omega_l,\omega_n,\omega_m\}$ for viewpoint evaluation and weight parameters $\{\omega_d, \omega_q, \omega_a\}$ for viewpoint transition to adapt to various scene and production contexts.

For the user contexts, users' viewpoint selection tendencies can be learned by analyzing viewpoint selection records of actual users and generate various viewpoint patterns for different contexts by optimizing the corresponding weight parameters. The optimal parameter set capable of generating the preferable viewpoint sequence for the corresponding contexts was similar to users' viewpoint selections. In particular, a user's favorite player can be specified as the main object of the scene by assigning the player a higher weight value ω_m when the parameter optimization is not conducted. Thus, the proposed method generates a recommended viewpoint sequence featuring the favorite player more often.

4.5 Context-dependent Weight Parameter Optimization

This section describes context-dependent weight parameter optimization in detail.

4.5.1 Outline of Optimization

In the optimization process, a brute-force attack algorithm is applied to conduct the searching task for the optimal set of the weight parameter vector as follows,

$$\omega^j = (\omega_p^j, \omega_c^j, \omega_l^j, \omega_n^j, \omega_m^j, \omega_d^j, \omega_q^j, \omega_a^j), (j = 1, 2, ..., R), \tag{4.13}$$

where R is the number of context defined. Parameters within the range $[0, 1]$ are sought over the scene contexts, to achieve the highest similarity of the selected viewpoints by the proposed method against the users' selections. The step size is 0.1. For the five weight parameters in the viewpoint evaluation process, the similarity between the best-ranked viewpoints recommended by camera agents and user selections are calculated in the optimization. Furthermore, the three parameters in the transition process are optimized by comparing with the selected viewpoints by producer agent and user selections.

The inputs of the optimization are the user selected viewpoint records p^u and the ball and player data, which are used to generate the viewpoint sequence p^s by the proposed method. To quantify the similarity between them, concordance rates related to the sameness of the camera positions at each frame are defined with the parameter vector ω^j as follows:

$$E(t;\omega^j) = \begin{cases} 1 & (\text{if } p^s(t;\omega^j) = p^u(t)), \\ 0 & (\text{otherwise}). \end{cases} \tag{4.14}$$

$$\text{Rate} = \frac{\sum_t E(t;\omega^j)}{L}, \tag{4.15}$$

An optimal weight vector is given as the vector which maximizes the sum of the similarity as follows,

$$\omega^j = \arg\max_{\omega^j} \text{Rate}, \tag{4.16}$$

where L is the length of the sequence.

Parameter optimization is first performed based on the common viewpoint sequences, extracted from multiple-user viewpoint selection records, as the optimization goal. Here, the parameter performance was evaluated with leave-one-sequence-out cross-validation, by using one sequence of the viewpoint selection records as testing data, and the other sequences as training data. This step is repeated until each sequence has been set as the test data. The performance is evaluated using the average accuracy of all the test sequences. Furthermore, individual viewpoint selection records were used to experiment with personalized optimization to investigate the possibility of personalized recommendation, which is detailed in Section 4.7.4.

4.5.2 Scene Context Division

Several kinds of scene context division are defined. As mentioned above, a video sequence can be represented as a series of scene contexts in terms of ball positions and

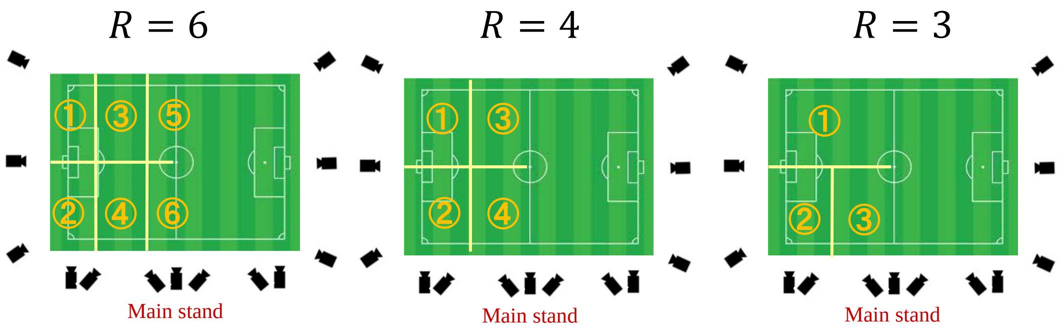


Figure 4.3: Scene context divisions.

its influential regions. It was assumed that the contexts were symmetrical between the left and right sides of the field so half of the field was divided into R regions representing R characteristic contexts, as shown in Figure 4.3. Context-dependent optimization is performed for each division. On the other hand, context-independent optimization is performed as a baseline method using $R = 1$ when there is only one division, which means the entire field, thereby allowing us to optimize one set of parameters no matter where the ball is. The performance of the proposed method using the optimized parameter vector by concordance rate is evaluated as in Eq. 4.15.

4.6 Experiment

Viewpoint selection records of ten participants acquired in the multi-view video editing experiment using the dataset and steps described in Chapter 3 were used to optimize the weight parameters, and to verify the effectiveness of the proposed method.

The ten participants comprised six males and four females, all in the age group of 20 to 39. They were asked to state their profile information via a questionnaire. In the questionnaire, the interest level in soccer was assessed on a four-level selection from “almost none” to “very much”. 10% of the participants was strongly interested in soccer (Level 4), 40% had a general interest, and 50% agreed to the statement “a little” (Level 2). For the viewing frequency of a soccer game, 50% of the participants were occasional viewers with a frequency of a few times a year, and 40% viewed a few times a month or even more than once a week. Moreover, half of the participants had soccer playing experience, but with no particular expertise. Furthermore, two of them had amateur video photography experience with a video camera, or video editing experience using an editing software.

The editing record of each participant reflected their personal viewpoint selection tendency, and can be used for personal tendency modeling. The selected viewpoint sequence of common viewpoint selection tendency was generated by connecting the viewpoint the participants chose most frequently for each experimental sequence. The common viewpoint sequence was then used for common tendency modeling and for the performance test.

To evaluate the effectiveness of the context-dependent recommendation, let's compare the concordance rates of the context-independent baseline method $R = 1$ with the scores using context-dependent optimized weight parameters of each scene context division. The baseline method was conducted for the entire field without context division.

4.7 Results and Analysis

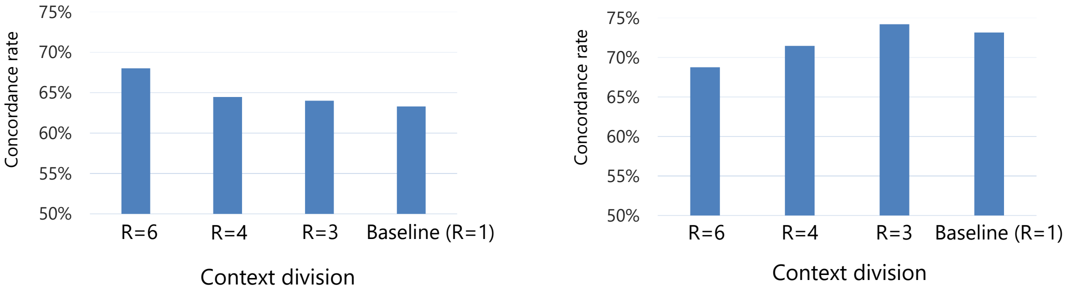
According to the method explained in Section 4.3, a viewpoint recommendation could be generated as designed using the optimized parameters. Even though the computational cost of parameter optimization is high, once the parameters are specified, the computation of the viewpoint selection requires very little time. The processing time for 15 minutes of video data on a PC with 2.67 GHz Intel Core i7 CPU took only 483 milliseconds.

This section presents the verification of the effectiveness of context-dependent parameter optimization and an analysis on the effect of different context divisions on performance changes. Additionally, a detailed evaluation of the proposed method was performed by separately analyzing the performance of the camera and producer agents using the optimized parameters and then users' viewpoint selection tendencies were analyzed through the optimized parameters.

4.7.1 Effectiveness of the Context-dependent Recommendation

The concordance rate of the context-independent baseline method $R = 1$ was compared with the scores using context-dependent optimized weight parameters of each scene context division. The concordance rates of the best-ranked viewpoints recommended by camera agents for the common viewpoint selection tendency are shown in Figure 4.4.

The graphs show that, for Game 1, the results obtained with the proposed method were similar to the user's viewpoint selection tendencies when the appropriate



(a) Game 1

(b) Game 2

Figure 4.4: Concordance rates of the best-ranked viewpoints generated through different context division and baseline.

context-dependent weight parameters of all the context divisions ($R = 3, 4$, and 6) were used, contrary to the baseline method, and achieved the highest similarity for $R = 6$. Therefore, we can say that it is effective to conduct context-dependent optimization. For Game 2, the score of $R = 3$ was higher than the others, on the contrary. The scores of $R = 4$ and 6 were slightly lower than that of the baseline. This may be due to excessive divisions which caused a scarcity in learning data for each context. To verify the assumption, an attempt to optimize parameters using the resubstitution estimate method was made, which achieved 79% for $R = 6$ and 80% for $R = 4$, surpassing 75% of the baseline. This indicates that the scarcity of learning data may cause over-fitting.

4.7.2 Effectiveness of the Camera Agents and Producer Agent

To evaluate the performance of the proposed method in a detailed manner, let's focus on the result of context division $R = 3$ for the common viewpoint selection tendency.

Performance Test of the Camera Agents

First, let's verify the performance of the camera agent by calculating the coverage rate, which is the rate indicating the similarity between users' selections and the best-ranked viewpoints the camera agents suggested for the duration of the sequence. In Game 1, the coverage rate of the best-ranked viewpoints was approximately 64%. The rate became approximately 81% if we allowed the viewpoints with the top two ranks. In Game 2, the coverage rates increased from 74% to 87%, respectively. The high coverage of users' selections shows that the performance of the camera agent is fairly reliable.

Table 4.1: Comparison of the switching frequency (common viewpoint selection tendency and excerpted personal viewpoint selection tendency samples)

	Game 1 [times]			Game 2 [times]		
	CamS	ProS	UserS	CamS	ProS	UserS
Common	16.0	7.3	6.4	10.3	6.1	5.4
User #4	15.7	5.7	2.5	11.9	5.0	2.0
User #6	15.5	10.3	4.8	9.7	5.6	3.6
User #7	17.7	6.4	3.2	13.4	4.7	2.1
Average	16.4	7.0	3.9	10.9	5.7	2.9

Performance Test of the Producer Agent

Next, let's evaluate the producer agent's results generated based on the outputs of camera agents through both the concordance rate (Eq. 4.15) and the switching frequency.

The concordance rate of Game 1 was approximately 62%, whereas that of Game 2 was 75% indicating that the proposed method is effective across videos from different games.

For verifying the ability of the producer agent to reduce the frequency of viewpoint switching, the number of switching occurrences of the sequences the users edited (UserS) was compared with the top-ranked viewpoint sequences of the camera agents (CamS) and the final output generated by the producer agent (ProS). Table 4.1 lists the comparison for the common viewpoint selection tendency. It shows (see "common" row) that the producer agent decreased the number of switching recommended by camera agents from 16.0 and 10.3 to 7.3 and 6.1 times for the two games, respectively, which makes it closer to the users' decisions.

4.7.3 Analysis of the Optimized Parameters

This section presents the analysis on the behavior of the optimized parameters dependent on three contexts for $R = 3$, namely upside (r_1), bottom-corner (r_2), and bottom-mid-field (r_3).

For the common recommendation, a sample of optimized parameters for each scene context of Game 2 are shown in Figure 4.5. It shows that each view-related context yielded corresponding weight parameters. For instance, in the upside area

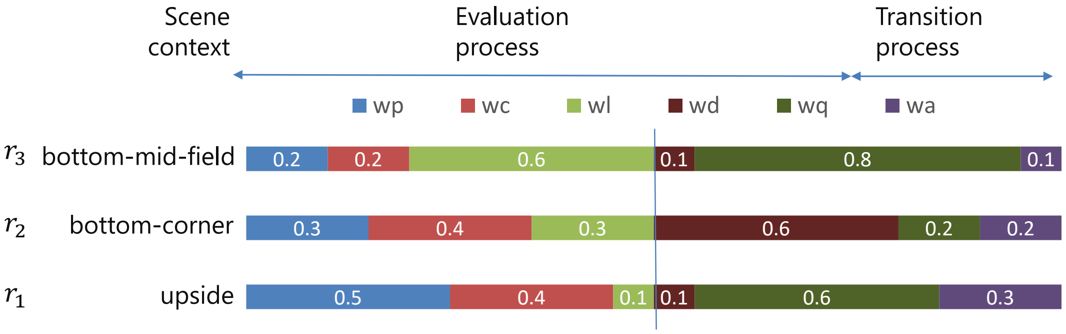


Figure 4.5: Sample of optimized weight parameters corresponding to different scene context areas for the common viewpoint selection tendency in Game 2 with R = 3.

context r_1, the value of the proximity measure w_p (= 0.5) was the largest. Thus we can infer that the proximity between the objects and the camera was the most important measure for viewpoint evaluation, indicating that many users might tend to watch views of the upside area more closely. On the other hand, the value of the location measure w_l (= 0.6) was the largest in the bottom-mid-field context r_3. Thus, users may be more interested in the location distribution of the ball and players, because the context-related location measure was the most important.

4.7.4 Analysis of Personal Recommendation

The performance of the proposed method for personal recommendation was also evaluated.

The average concordance rate of the ten users in the performance test of the producer agent by personal optimization was approximately 53% for both two Games. This is much lower than the common optimization (refer to Section 4.7.2). It is considered due to the insufficient learning data for personal optimization comparing to common optimization using the extracted common records of most users with more stability. Still the scores are relatively high against random choice out of 13 and 14 cameras of each of the two games, which would be approximately 8%.

For the ability on controlling the switching frequency, Table 4.1 lists examples of typical user results and the averages of all users'. The table shows that the producer agent decreased the number of switching suggested by camera agents from 16.4 and 10.9 to 7.0 and 5.7 times for the two games, respectively, which makes it closer to the user's decision.

Moreover, it is possible to analyze the personal viewpoint selection tendency

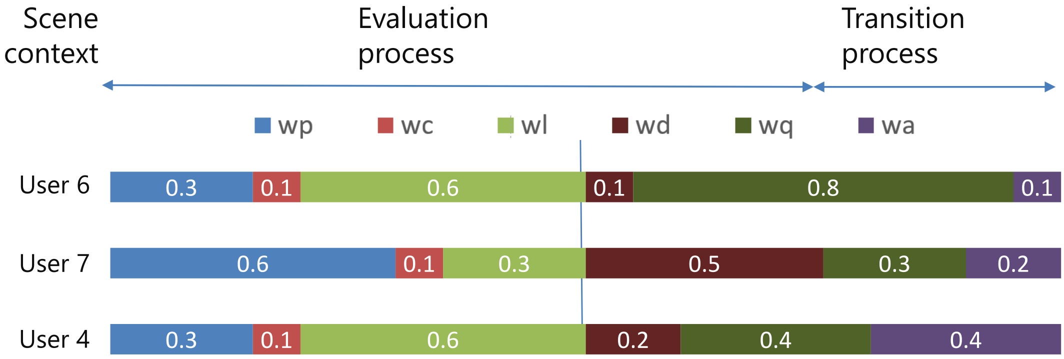


Figure 4.6: Sample of optimized weight parameters for different users in Game 1 in upside area scene context with R = 3.

of each user through the optimized parameters. For instance, Figure 4.6 shows optimized parameters for users #4, #6, and #7 in the upside scene context r_1. The graph shows that the weight parameters for transition process of user #4 included a view angle weight w_a (= 0.4) the same as the value of view quality change. Apart from the angle, user #7 paid attention to the duration of the previous viewpoint selection w_t (= 0.5). On the other hand, for user #6, the value of the view quality change weight w_q (= 0.8) was the largest, disregarding the duration and angle changes. We can confirm these observations through the switching frequencies listed in Table 4.1. User #6 changed the viewpoint more frequently than users #4 and #7, which corresponds to the viewpoint selection tendency indicated by the weight parameters. This enables us to use the optimized parameters to analyze user viewpoint selection tendencies for further research.

4.8 Summary

This chapter proposed a context-dependent learning scheme to generate viewpoint recommendation adapted to various video contexts and user contexts for multi-view videos. The concordance rate with users' viewpoint selection records of the context-dependent recommendation outperformed the context-independent recommendation, which shows the effectiveness of the context-dependent learning scheme. The proposed method selects context-dependent optimal sets of viewpoints for both common and personal viewpoint selection tendencies. In addition to automatic recommendations, the proposed method permits users to change their preferences during a recommendation by changing the parameters or the main object of their interest. The experimental results were used to analyze user behavior and viewpoint selection tendency in more detail.

Chapter 5

Spatio-temporal Video Context Representation

5.1 Introduction

As shown in the outline of the thesis (Figure 1.2), in the common step, the acquired multi-view and object tracking data are represented as features to learn a user's viewpoint selection tendency. The feature representation influences the quality of the viewpoint recommendation. Thus, this chapter proposes a video context representation based on spatio-temporal feature to generate a viewpoint recommendation with better similarity to a user's viewpoint selection tendency.

The scene context of a sport game related to objects is known to be effective to select the viewpoint for multi-view sports videos [Shen et al., 2007, Daniyal et al., 2010, Chen et al., 2013, Wang et al., 2015]. It is often represented by frame-level object features, for instance, the size of a player visible in the view. However, these works do not sufficiently consider the representation of past and future object dynamics. Since the users would naturally select their preferred viewpoints considering the temporal information from past to future, spatial-temporal object dynamics is considered for the viewpoint selection in this chapter. Besides, user's personal viewpoint selection tendency on spatio-temporal object dynamics is also considered.

Personal viewpoint recommendation is realized as shown in Figure 5.1 considering the spatio-temporal scene context represented by the trajectories of the main focused objects, i.e., the ball and players. Regarding the trajectory processing, some works used time-series models, such as Markov Models, for representing player's

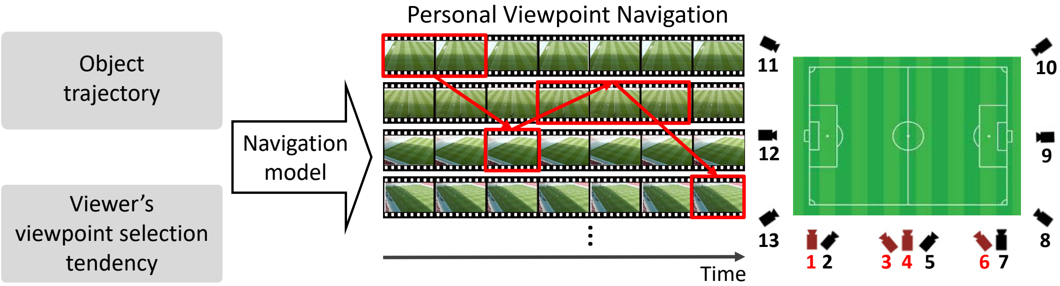
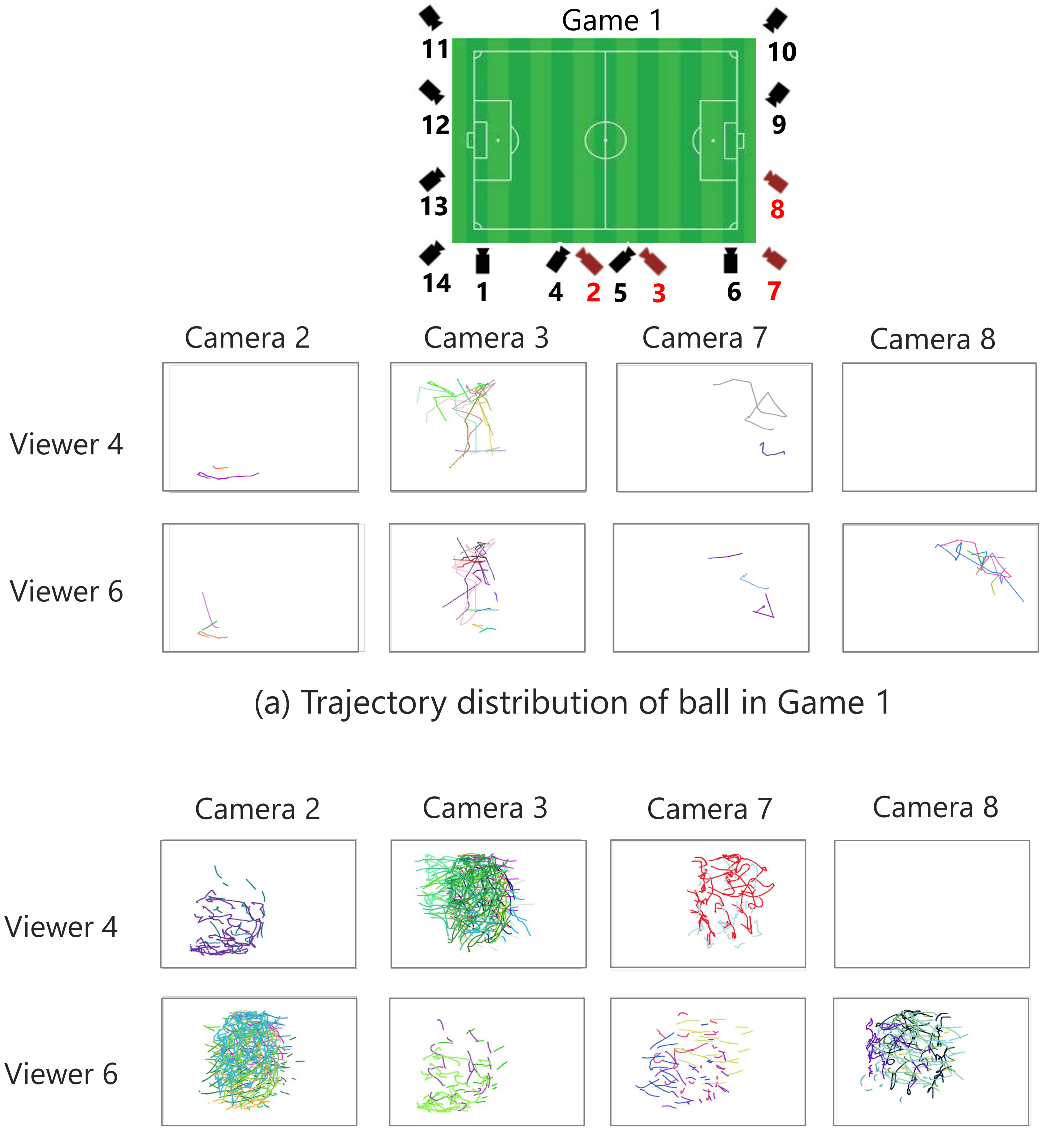

Object trajectory
Viewer's viewpoint selection tendency
Navigation model
Personal Viewpoint Navigation
Time
11
12
13
10
9
8
1 2
3 4 5
6 7

Figure 5.1: Outline of personal viewpoint recommendation by learning the relationship between the trajectory distribution of the focused objects and each user's viewpoint selection tendency.

dynamics or action recognition [Sun et al., 2009, Itoda et al., 2015]. Observation of trajectory distributions of soccer games introduced in detail in Section 5.3, reveal that different users show different viewpoint selection tendencies for different trajectory distributions. The spatial distribution of objects' trajectories include dynamic information and can represent the objects' actions. Thus, referring to the spatial distribution of objects' trajectories to represent the spatio-temporal scene context is proposed in this chapter to improve the recommendation quality for the viewpoint selection.

Machine learning is applied to learn the relationship between the trajectory distribution of the focused objects and each user's viewpoint selection tendency for personal viewpoint recommendation. The video sequences are first segmented into cuts. The object trajectory distribution in a cut is extracted as a feature to learn each user's viewpoint selection tendency. Personal recommendation is generated by selecting appropriate viewpoints through the proposed learning methods. Three methods based on Gaussian Mixture Models (GMM), Support Vector Machines (SVM) with a 2D histogram, and SVM with a Bag-of-Words (BoW) are evaluated. In addition, the effectiveness of using the combination of focused objects is compared with just using each focused object to find the user's interest in multiple objects. Moreover, the fact that most users focus on the main object in a sport game while some of them show unique interest in specific objects is presented.

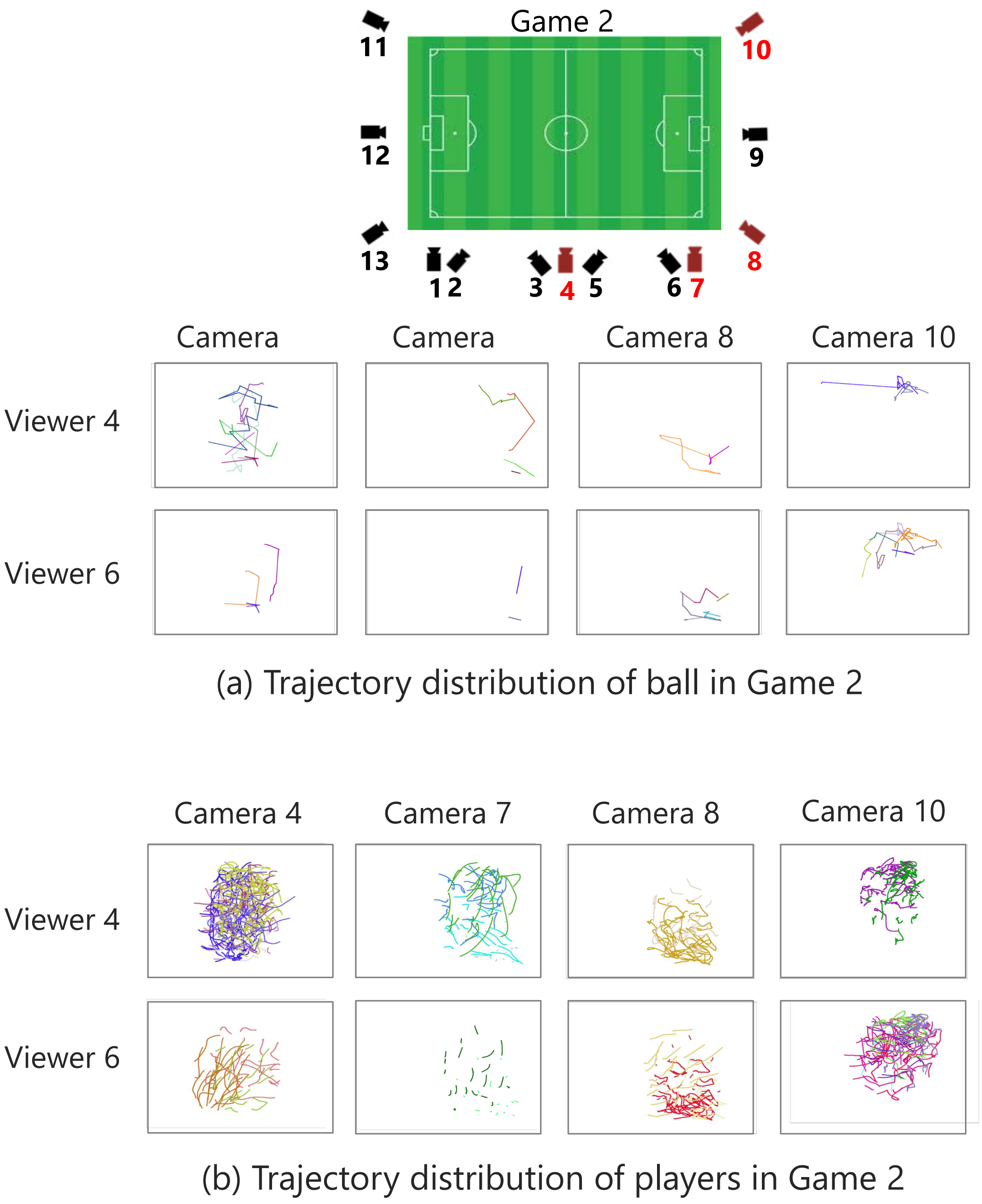
The remainder of this chapter is organized as follows. Section 5.2 reviews related works. In Section 5.3, the relationship between scene context and user's viewpoint selection tendency is analyzed. In Section 5.4, the proposed framework of object trajectory based viewpoint recommendation is introduced. In Section 5.6 experimental results are analyzed. Section 5.7 summarizes the chapter.



(a) Trajectory distribution of ball in Game 1

(b) Trajectory distribution of players in Game 1

Figure 5.2: Ball (a) and players (b) trajectory distribution for each selected viewpoint (extracted samples) of two users in Game 1. Each of different colored lines shows the trajectory during a cut in the entire soccer field. Camera settings of Game 1 are shown on the top. The cameras in red are the extracted ones for showing the trajectory distribution below.



(a) Trajectory distribution of ball in Game 2

(b) Trajectory distribution of players in Game 2

Figure 5.3: Ball (a) and players (b) trajectory distribution for each selected viewpoint (extracted samples) of two users in Game 2. Each of different colored lines shows the trajectory during a cut in the entire soccer field. Camera settings of Game 2 are shown on the top. The cameras in red are the extracted ones for showing the trajectory distribution below.

5.2 Related Work

Several researchers have focused on the scene context of a game. Daniyal et al. presented an algorithm for viewpoint-quality ranking based on frame-level features, including size and location of the players in a basketball game [Daniyal et al., 2010, Daniyal and Cavallaro, 2011]. Chen et al. focused on features of a group of objects such as the number of players who are visible from a viewpoint [Chen et al., 2013]. Their approach optimizes the viewpoint transition by viewpoint-quality evaluation with dynamic features corresponding to the scene context represented by the object position [Wang et al., 2015]. Muramatsu et al. selected viewpoints referring to the average of object features, such as position, distance to the camera and size in the view during a short time, to learn from the user's viewpoint selection records [Muramatsu et al., 2014].

However, these approaches processed at the frame-level or without sufficient representation on past and future object dynamics. We can expect that the scene context of a game can be described better by the object's trajectory information, and that the recent machine-learning representations are more effective than such simple statistical representations.

5.3 Analysis of Trajectory Distribution and Viewpoint-selection Tendency

We can assume that the users select appropriate viewpoints based on the scene context, which can be represented by the spatio-temporal movement of focused objects, i.e., the ball and players in the case of a field ball sports. Thus, in this section, the relationship is analyzed between the trajectory distribution of focused objects and each user's viewpoint selection records, which were acquired from the multi-view editing experiment described in Chapter 3.

5.3.1 Ball Trajectory Distribution

It is known that most users of soccer games have a tendency to follow the ball [Iwatsuki et al., 2013]. Thus, let's first analyze the relationship between the trajectory of a ball and personal viewpoint selection tendency. Figure 5.2 (a) and Figure 5.3 (a) show the ball trajectories when each camera was selected by two users for Games 1 and 2. The rectangles represent the ground plane of the soccer field. Each of differently colored lines in the rectangle indicates the trajectory during a cut in the entire soccer field. In both games, when a user selected a viewpoint, the ball trajectories tended to center around an arbitrary location on the soccer field. Besides, the trend was different between users. For example, in Game 1, Viewer 6 preferred Camera 8 when the ball moved in the corner area on the right-side of the field while Viewer 4 did not select Camera 8. The difference of users also existed when Cameras 4 and 10 were selected in Game 2. It can be inferred that users' preference on the location, angle, and other conditions of the focused objects in a viewpoint differed according to scene contexts. As a result, since their viewpoint selection tendencies are different, the ball trajectory distributions appear differently when different viewpoints are selected by different users.

Therefore, using the ball trajectory distribution is considered as effective to learn different viewpoint selection tendencies.

5.3.2 Players Trajectory Distribution

Players are also important objects in soccer games. Thus, let's also analyze their trajectories to represent the scene context that has an impact on personal viewpoint recommendation.

First, trajectories of all players except for the two keepers were collected and the same analysis as the ball was performed. Figure 5.2 (b) and Figure 5.3 (b) show the players' trajectory distributions when each camera was selected for Games 1 and 2. From these distributions, we can see that players' trajectories also have different trends among not only the viewpoints but also the users. However, the difference is not as distinct as the case of ball trajectory.

In addition, the player who will receive the ball at the next moment was analyzed since the users seemed to pay their attention to that player. In this case, the players' trajectories showed similar trends with the ball trajectory case for each selected viewpoint.

As a result, we can say that using players' trajectory distribution also can be effective to learn different viewpoint selection tendencies for some users who pay more attention to the players.

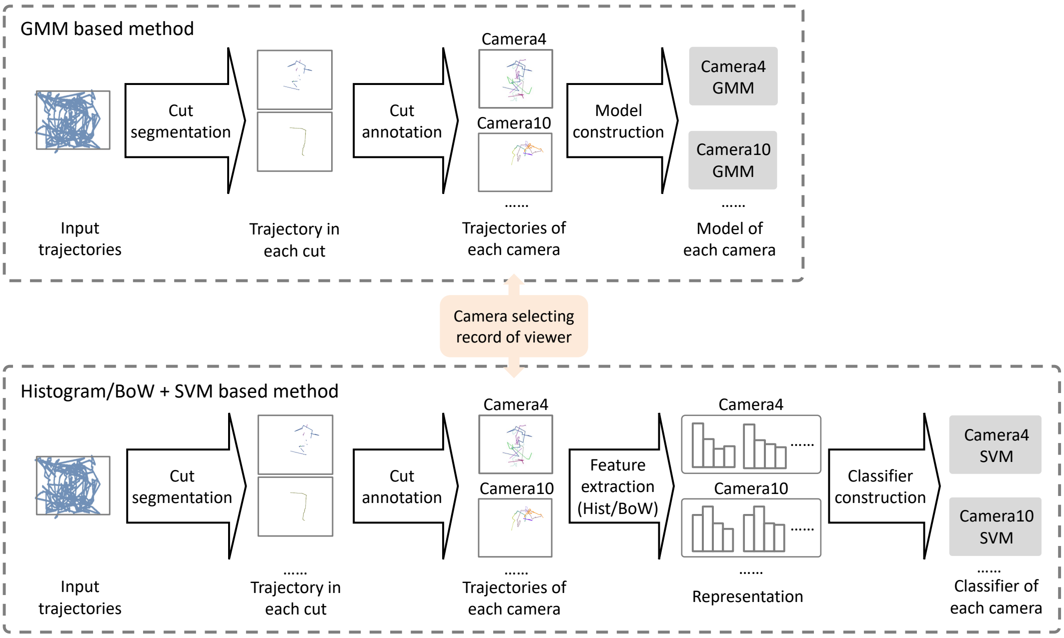


Figure 5.4: Outline of the schemes for learning the relationship between personal viewpoint selection tendency and trajectory distribution of focused objects in cuts. *Cut annotation* means to gather and label the cuts when the same viewpoint is selected.

5.4 Object Trajectory based Recommendation

Based on the result of the analysis obtained in Section 5.3, the following three kinds of objects were chosen as focused objects:

- B: the ball of a soccer game,
- P_n: the player who will receive the ball at the next moment,
- P_{all}: all players of a soccer game except for the keepers.

The proposed recommendation framework learns the relationship between the personal viewpoint selection tendency and the trajectory distribution of the focused object to recommend a viewpoint for each cut.

In this section, first, three methods for learning the relationship by using each focused object and a combination method of using trajectory data in cuts of multiple focused objects are discussed. Then, two definitions of a "cut" including several frames segmented from the video sequence are introduced. All the combinations;

three learning schemes times five kinds of focused object combinations times two kinds of cut definitions are compared to find the best one for viewpoint recommendation.

In the following discussion, a sub-trajectory of each focused object is represented by $T_{C_i} = \{\mathbf{x}_{t,j} | t = 1, 2, \ldots, F_i; j = 1, 2, \ldots, J\}$, where i is the cut index and $\mathbf{x}_{t,j} \in \mathbb{R}^2$ is the point on the field coordinate system at frame t of object j. F_i is the length of cut i. J is the number of focused objects, where J of B or P_n is 1, J of P_{all} is 20. $\mathbf{x}_t$ is used to represent $\mathbf{x}_{t,j}$ for short including all the objects as the focused objects at frame t. The v-th viewpoint is represented by v $(1 \leq v \leq V)$, where V is the number of viewpoints.

5.4.1 Machine Learning Scheme

Three methods were compared to find the best learning scheme for the relationship. They were first applied to a single kind of focused object trajectory. The outline of these methods is shown in Figure 5.4. GMM was considered as an appropriate descriptor of trajectory distribution for each selected viewpoint since it is widely used to express object-position distribution. Two baseline representations were tested to describe the trajectory in a cut, one used a simple 2D histogram, another used a BoW with soft assignment considering the softer boundary than a simple 2D histogram.

Gaussian Mixture Model based Method

GMM is a linear combination of several Gaussian components as follows,

$$p(\mathbf{x}_t) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}_t | \mu_k, \Sigma_k), \tag{5.1}$$

where K is the number of the Gaussian components, π_k is the weight of the k-th Gaussian component with $\sum_{k=1}^{K} \pi_k = 1$, $N(\mathbf{x}_t | \mu_k, \Sigma_k)$ is the Gaussian component density with parameters μ_k and Σ_k. The number of components is based on experimental results, as discussed later. In this chapter, it is used to represent the focused object trajectory distributions of each viewpoint for each user. The focused object trajectories in cuts are gathered while viewpoint v is selected, and are represented as T_v. Thus, to generate GMM (p_v) of viewpoint v, $\mathbf{x}_t$ is sampled from T_v in a training dataset. EM algorithm is applied to estimate the parameters (π_k, μ_k, and Σ_k).

For each video sequence in a test dataset, the video sequence is first divided into cuts and a focused object trajectory T_{C_i} is extracted from cut i. Next, the total log-likelihood for the points on trajectory T_{C_i} under the generated GMM of each viewpoint for the user is calculated. Thus, given a trajectory T_{C_i}, a viewpoint R with the largest log-likelihood is output as follows,

$$R(T_{C_i}) = \underset{1 \leq v \leq V}{\arg\max} \sum_{\mathbf{x}_t \in T_{C_i}} \log p_v(\mathbf{x}_t). \tag{5.2}$$

Histogram and SVM based Method

A method based on SVM with a 2D histogram called *Hist-SVM* for short is considered as one baseline representation. The field is spatially divided equally into $M \times N$ bins. A 2D histogram of points of focused object trajectory T_{C_i} in cut i is calculated. Histogram normalization is applied considering the differences in the lengths of the cuts.

In the training step, the normalized histogram is used as a feature vector and the learning step is performed by SVM with a Radial Basis Function (RBF) kernel. The supervising signals are the viewpoint selection records of each user. Thus, a one-vs-all classifier corresponding to each viewpoint is built.

In the testing step, the learned classifiers output viewpoints with the maximum score for each input trajectory T_{C_i} in cut i. The output viewpoints of cuts compose the recommended sequence. Here, M and N are set to optimal values based on the experimental results.

BoW and SVM based Method

A method based on SVM with a Bag-of-Words (BoW) representation [Csurka et al., 2004] called *BoW-SVM* for short is also considered as a baseline representation. All the points of the focused object trajectories in the training data are clustered to codewords by an unsupervised GMM shown in Equation 5.1. The number of Gaussian components is empirically set. A codeword is defined for each Gaussian component in the GMM. Thus, the number of codewords is the same as K, i.e., the number of components. Soft assignment is applied to generate the histogram of codewords in cut i. Under the k-th Gaussian components, the value of the bin of the

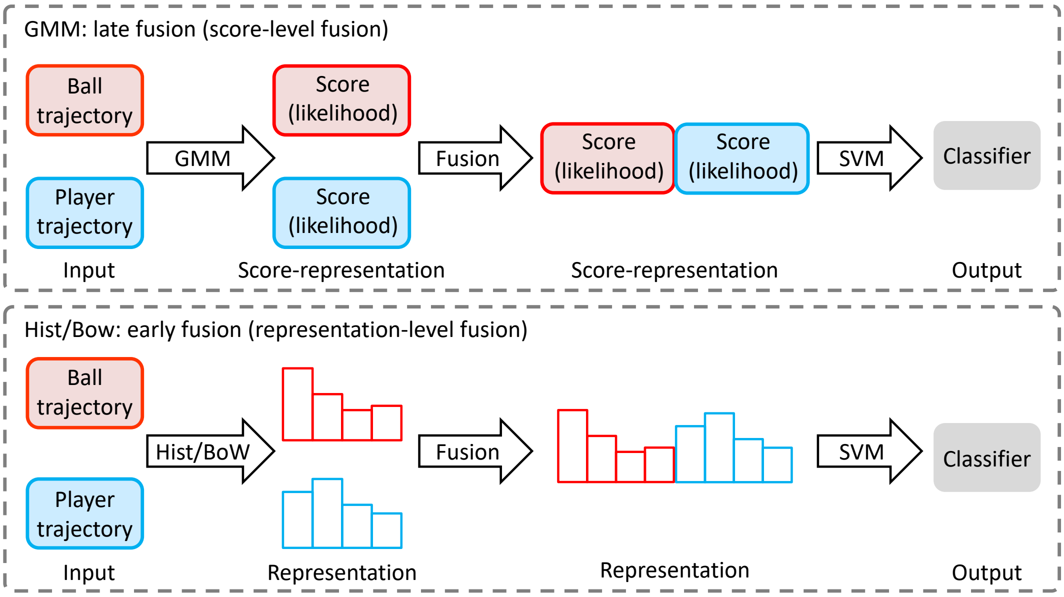


Figure 5.5: Fusion methods of combining focused object trajectories.

k-th codeword histogram is calculated as responsibility $p(k|\mathbf{x}_t)$ as follows:

$$p(k|\mathbf{x}_t) = \frac{\pi_k N(\mathbf{x}_t|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(\mathbf{x}_t|\mu_k, \Sigma_k)}. \tag{5.3}$$

Thus, the responsibility vector of K codewords of $\mathbf{x}_t$ will be $\mathbf{a}(\mathbf{x}_t) = [p(1|\mathbf{x}_t), \ldots, p(K|\mathbf{x}_t)]$. Then the normalized codewords histogram can be generated as a feature vector $\mathbf{A}_i$ of cut i as follows,

$$\mathbf{A}_i = \frac{\sum_{t=1}^{F_i} \mathbf{a}(\mathbf{x}_t)}{F_i}, \tag{5.4}$$

where $\mathbf{x}_t \in T_{C_i}$ and F_i is the length of cut i.

In the training step, the relationship between the feature vectors $\mathbf{A}_i$ ($T_{C_i} \in T_v$) and the selected viewpoint v by each user is learned with an RBF kernel-based SVM. Thus, a one-vs-all classifier is built as with *Hist-SVM*.

In the testing step, the learned classifiers output a viewpoint for each input trajectory of a cut as with *Hist-SVM*.

5.4.2 Focused Objects Fusion

Different users may focus different focused objects. Thus, each object or different combination of objects are used as feature vectors to find which object or combination is the most focused by each user. The proposed learning scheme introduced in Section 5.4.1 was applied to a single focused object trajectory. In this section, fusion methods corresponding to the proposed learning schemes for two kinds of combined focused objects are added: $B + P_n$ and $B + P_{\text{all}}$.

There are several methods used to fuse multiple features, typically including early fusion conducted at the representation level, or late fusion conducted at the score level [Snoek et al., 2005, Peng et al., 2016]. For the representation level fusion, multiple features are integrated into a single feature representation, which is fed into one supervised classifier. The integrated feature representation may reflect better multiple information and correlation, though the higher dimension will increase the difficulty of learning. For score level fusion, multiple features are separately fed into classifiers. The scores of the classifiers are combined afterwards to yield a multiple representation for the final learning stage. Late fusion focuses on the individual feature strengths but it may result in the loss of correlation of features.

As shown in Figure 5.5, different fusion methods are used depending on the learning scheme. For the GMM-based method, if early fusion is applied, we will have to integrate 2D trajectory data of ball and players included in the focused objects (21×2 dimensions for B combined with P_{all}) into a single feature vector before the GMM generation. Considering the difficulty of learning a high dimensional features, the late fusion is applied; First, GMMs of viewpoints for ball (B) trajectory and GMMs for players (P_n or P_{all}) trajectory are generated separately. Then the likelihood scores of each GMM of the ball and players are integrated into one feature vector and learned by an SVM afterwards instead of the maximum likelihood decision rule used for each focused object. For the *Hist/BoW-SVM* based method, early fusion is applied considering the correlation of ball and players trajectories; First, the data of different focused objects are input into histogram/BoW generator separately and the resulting representations are then fused into a single integrated feature vector for the supervised learning using SVM.

5.4.3 Video Cut Segmentation

Considering that users would naturally select viewpoints according to the scene context in past and future periods, not solely on the current frame, cuts which consists of multiple frames are used to represent the spatio-temporal scene context. Therefore, it is necessary to determine how the video sequence should be segmented. Here, two kinds of segmentation are considered.

Ideal Segmentation (*SegU*)

The ideal segmentation (*SegU* for short) is a result of the scene context being appropriately classified according to personal preferences, which is behind the user's viewpoint selection. It is unavailable for viewpoint recommendation in practice. Here, the user's viewpoint switching timing is recorded to ideally segment the video sequences to verify the upper bound of the recommendation accuracy. Thus, for *SegU* based experiments, the switching timing is applied for cut segmentation both in the training and testing steps, and the best viewpoint for each segment is selected.

Equal Segmentation (*SegS*)

Equal segmentation (*SegS* for short) is a result when the video sequence is segmented into cuts with a fixed length. The sliding-window method is applied to compensate for the overlap in each cut. Concretely, a cut with a window size is generated around each frame. The selected viewpoint based on the trajectory distribution in a cut is assigned to the center frame of the cut. Thus, for *SegS* based experiments, a sliding window processing is applied along the video sequence both in the training and testing steps, and the best viewpoint for each frame is selected. Unlike *SegU*, *SegS* can be performed without deciding the cut boundary by external information such as viewpoint switching timing in *SegU*. Therefore, *SegS* can be used in practice. The window size is experimentally selected.

5.5 Experiment

In this chapter, the proposed method is validated using the collected ten participants' viewpoint selection records described in Chapter 4. The editing record of each participant reflected their personal viewpoint selection tendency, and can be used for personal tendency modeling.

To evaluate the effectiveness of the proposed method, the concordance rates of the following two methods were compared.

AveragePos uses the centroid of the ball positions during a cut as the feature and trains an RBF kernel based SVM [Muramatsu et al., 2014].

WeightOptm uses context-dependent weights optimized by a brute-force method to combine the features including the distance between cameras and objects (ball and players), composition in the view, angle change between switching viewpoints in each frame [Wang et al., 2015].

The concordance rate between each participant's viewing record $R^u(t)$ and the output viewpoints $R^s(t; \mathbf{x}_t)$ of proposed methods at each frame of a video sequence was calculated the same as in Section 4.5.1 as follows,

$$E(t; \mathbf{x}_t) = \begin{cases} 1 & (\text{if } R^s(t; \mathbf{x}_t) = R^u(t)), \\ 0 & (\text{otherwise}). \end{cases} \tag{5.5}$$

$$\text{Rate} = \frac{\sum_t E(t; \mathbf{x}_t)}{L}, \tag{5.6}$$

where L is the length of a sequence.

A leave-one-sequence-out cross-validation was performed by using one sequence of each participant's viewing record as test data, and the rest as training data. The concordance rate of each test sequence and the average rate over all the test data was calculated.

5.6 Results and Analysis

The evaluation framework described above was used to evaluate the performance of the proposed methods using three learning schemes, five kinds of focused object combinations with two kinds of cut segmentation methods, to find the best combination for personal viewpoint recommendation.

5.6.1 Parameters

The proposed methods achieved the highest average concordance rate using the following parameters. The numbers of components in GMM for Games 1 and 2 were 4 and 1, respectively. The numbers of codewords in *BoW-SVM* using each focused object for Games 1 and 2 were 33 and 39, and using the focused object combinations were 12 and 13, respectively. With regard to *Hist-SVM*, $21 (= 7 \times 3)$ bins were the best for both games.

Table 5.1: Result of variance analysis (three-way ANOVA) for the concordance rates of ten participants using all the proposed methods (three learning schemes, five focused object combinations and two cut segmentations) for the two games.

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)	
Game 1						
Learning Scheme (LS)	2	0.019	0.009	0.598	0.551	
Objects Combination (OC)	4	0.257	0.064	4.148	0.003	**
Cut Segmentation (CS)	1	0.154	0.154	9.960	0.002	**
LS * TC	8	0.039	0.005	0.312	0.961	
LS * CS	2	0.024	0.012	0.762	0.468	
TC *CS	4	0.045	0.011	0.720	0.579	
LS * TC *CS	8	0.046	0.369	0.369	0.936	
Residuals	270	4.180	0.016			
Game 2						
Learning Scheme (LS)	2	0.048	0.024	2.434	0.090	+
Objects Combination (OC)	4	0.479	0.120	12.242	0.000	***
Cut Segmentation (CS)	1	0.143	0.143	14.663	0.000	***
LS * TC	8	0.225	0.028	2.877	0.004	**
LS * CS	2	0.007	0.004	0.352	0.703	
TC *CS	4	0.063	0.016	1.608	0.173	
LS * TC *CS	8	0.056	0.007	0.718	0.676	
Residuals	270	2.641	0.010			
Signif. codes: ‘***’ 0.001, ‘**’ 0.01 , ‘*’ 0.05, ‘+’ 0.1						

5.6.2 Comparison of Different Factors

Variance was analyzed (three-way ANOVA) including all the results of ten participants using the proposed methods (three learning schemes, five focused object combinations and two cut segmentations) to investigate the effects of different factors and their interaction. Table 5.1 summarizes the ANOVA results.

Each main effect for focused object combinations and cut segmentations was statistically significant ($p < .01$) in both games. Nonetheless, there was no significant difference in the learning schemes in Game 1 and marginal difference ($p < .1$) in Game 2. Moreover, any interaction of different factors was not significant in Game 1, while the interaction of the learning schemes and focused object combinations was statistically significant ($p < .01$) in Game 2. For Game 1, all factors were independent since there was no significant interaction of different factors. Details of each factor are discussed below.

Table 5.2: Concordance rates of the three learning schemes using the five focused object combinations with segmentation *SegU* of the two games. (*SegU*: segmentation according to user's viewpoint switching.)

	Fusion	B	P$_n$	P$_{\text{all}}$	B + P$_n$	B + P$_{\text{all}}$
Game 1						
Model	GMM	**66.64%**±9.79%	61.27%±10.02%	52.58%±15.28%	55.29%±12.27%	57.93%±12.95%
	Hist-SVM	59.90%±12.67%	57.78%±12.54%	47.88%±12.46%	58.47%±11.40%	57.63%±13.37%
	BoW-SVM	61.62%±12.12%	57.97%±11.12%	53.70%±11.23%	58.67%±11.01%	59.94%±12.33%
Game 2						
Model	GMM	**56.65%**±11.02%	48.51%±12.71%	38.20%±10.83%	42.70%±7.63%	42.51%±11.49%
	Hist-SVM	51.30%±10.76%	37.78% ±10.74%	41.10%±10.19%	48.65%±12.58%	52.93%±12.45%
	BoW-SVM	49.57%±10.88%	35.31%±9.81%	41.84%±5.60%	49.65%±8.14%	48.06%±9.50%

Comparison of Learning Schemes

Let's first discuss the recommendation performance of the proposed method under an ideal situation using the ideal segmentation, i.e, *SegU*. The average and the standard deviation of the concordance rates of ten participants using the three learning schemes and the five focused object combinations with ideal segmentation for the two games are shown in Table 5.2. From this table, the GMM-based method achieved the best concordance rates of 66.64% and 56.65% for the two games, respectively. Besides, considering the average concordance rate of all kinds of focused object combinations in the two games, the GMM-based method also achieved the best average concordance rate of 52.23%. Thus, with regard to the average concordance rates, the results of the GMM-based method were higher than those based on *Hist-SVM* and *BoW-SVM*, although there was no significant difference in the learning schemes by ANOVA analysis.

Next, the recommended sequences generated by different learning schemes using only the ball trajectory *B* were compared. The recommendation of the *Hist-SVM* based method was unstable when the trajectory distribution existed around the border lines of histogram division, while the *BoW-SVM* and the GMM-based methods selected the same viewpoint stably as the user record. The *BoW-SVM* based method made mistakes when the trajectory distribution existed at the center area especially on the far side of the field where multiple cameras (i.e., cameras 3, 4, 7, and 14 in Game 1) can cover the game. In contrast, the GMM-based method worked better in this situation. This is considered as a result from the difference between the construction methods of GMM. The *BoW-SVM* based method constructed GMM over all the ball trajectories around the field, which can capture global distributions whereas it is less sensitive to subtle difference at local areas. The GMM-based method constructed GMM from the gathered trajectories of each viewpoint (i.e. the

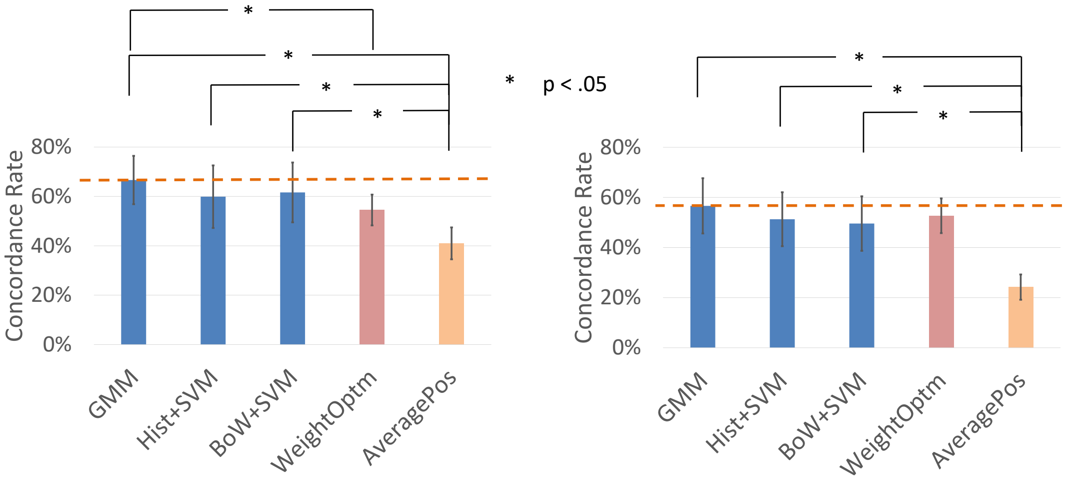


Figure 5.6: Average and standard deviation of the concordance rates of ten participants using only ball trajectory with *SegU* for the two games.

trajectories at local areas). The localized GMM acquired more discriminative ability and achieved a better performance for the overlapped trajectory distributions from multiple selected cuts.

Furthermore, the concordance rates of the comparative methods (i.e., AveragePos and WeightOptm mentioned in Section 5.5) were compared with the proposed methods using only the ball trajectory B with ideal segmentation ($SegU$). The average and the standard deviation of the concordance rates of these methods are shown in Figure 5.6. The pairwise comparisons using T-test with Bonferroni adjustment method revealed that all the proposed methods achieved significantly higher concordance rates than those of AveragePos (Game 1: mean = 40.99%, SD = 6.45% and Game 2: mean = 24.21%, SD = 5.05%), with $p < .05$ for both games. This result shows that the centroid of the ball position during a cut could not represent the scene context sufficiently. The GMM-based method performed significantly better than WeightOptm (Game 1: mean = 54.51%, SD = 6.24%), with p < .05 for Game 1. Although there was no significant difference between the GMM-based method and WeightOptm for Game 2, the former achieved a better average concordance rate. Since the WeightOptm method employed the distance information between cameras and focused objects besides scene context representation with focused objects information, using the trajectory distribution learned by the GMM-based method was probably more effective for scene context representation without camera information.

More detailed results are discussed using the GMM-based method below.

Comparison of Focused Object Combinations

To compare the difference among focused object combinations, pairwise comparisons using T-test with Bonferroni adjustment [Hochberg, 1988] were applied to the results of five combinations (B, P_n, P_{all}, $B + P_n$, $B + P_{\text{all}}$) using the three learning schemes with the ideal segmentation of ten participants for the two games. For Game 1, the comparisons showed that using only B (mean = 62.72%, SD = 12.15%) was significantly better than P_{all} (mean = 51.38%, SD = 13.57%) and $B + P_n$ (mean = 57.47%, SD = 11.87%), with p < .05. P_{all} was also significantly lower than P_n (mean = 59.01%, SD = 11.58%), $B + P_n$ and $B + P_{\text{all}}$ (mean = 58.50%, SD = 13.15%), with p < .05. For Game 2, only B (mean = 52.50%, SD = 11.49%) was significantly better than the P_n (mean = 39.82%, SD = 12.92%) and P_{all} (mean = 40.35%, SD = 9.47%), with p < .05. However, there was no significant difference between using only B and using $B + P_{\text{all}}$ for both games.

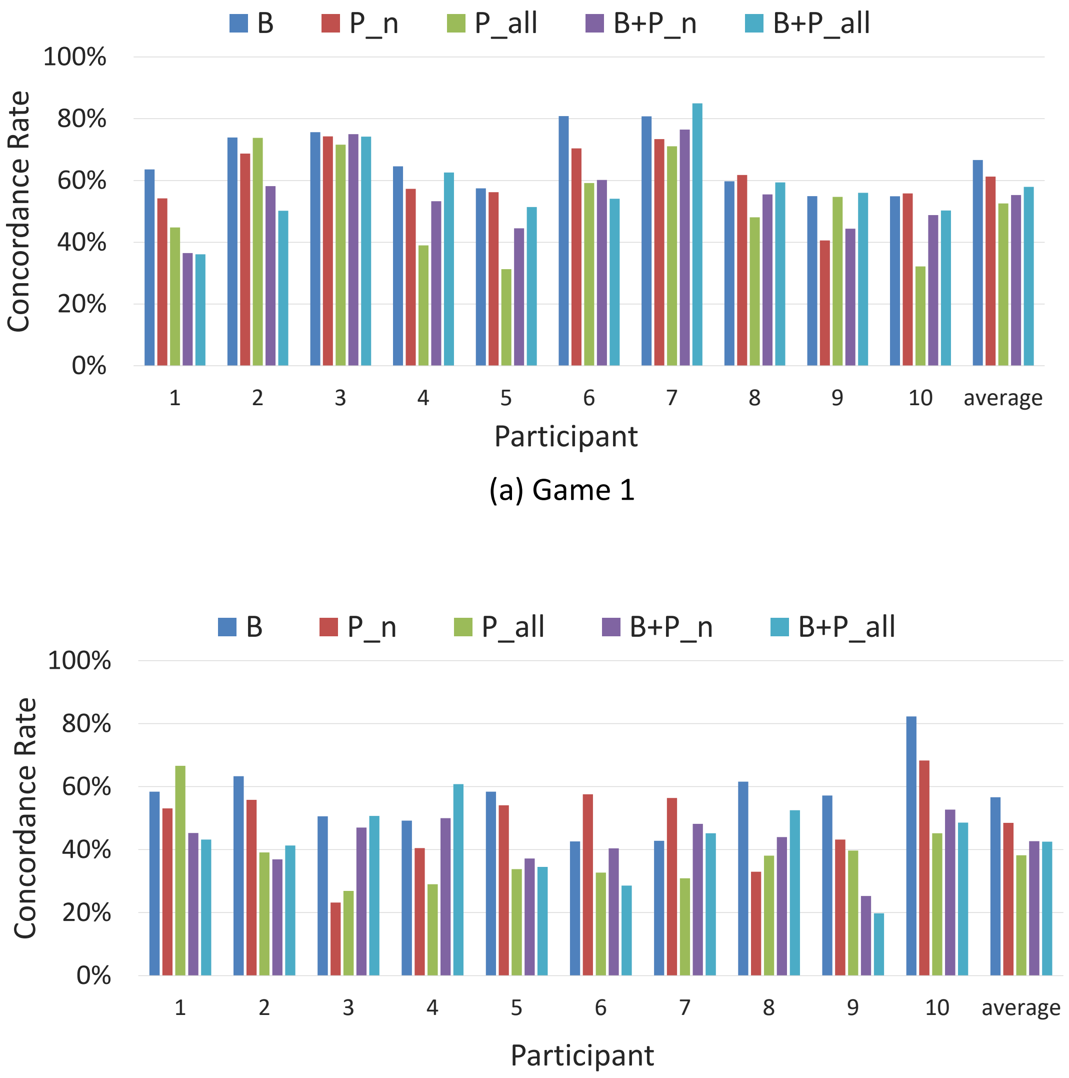
Regarding the interaction of the learning schemes and the focused object combinations from Table 5.2, the GMM-based method achieved better average concordance rate than the *Hist-SVM* and the *BoW-SVM* based methods when using only B, while there existed a contrary tendency among the learning schemes when using $B + P_n$ and $B + P_{\text{all}}$.

Furthermore, let's see the concordance rates of each participant using the five focused object combinations in Figure 5.7. We can see that most of the participants acquired the best concordance on the recommendation using only B. For Participants 7 and 9 in Game 1, the recommendation using $B + P_{\text{all}}$ achieved the best concordance rate. In Game 2, Participants 6 and 7 achieved the best concordance rate using P_n. Therefore, we can say that the ball was the main focused object of interest in the soccer games, which attracted more attention than players for viewpoint selection. However, some participants such as Participant 7 would pay more attention to players.

Therefore, for personal recommendation, the recommendation can achieve better effectiveness if appropriate objects were focused to reflect a user's viewpoint selection tendency.

Comparison of Cut Segmentation

The average concordance rates of the proposed methods using the two kinds of cut segmentation with only the ball trajectory B for the two games were compared. The GMM-based method with ideal segmentation (*SegU*) achieved 66.64% and 56.65%, while the one with equal segmentation (*SegS*) achieved 57.06% and 46.67% for the



(a) Game 1

(b) Game 2

Figure 5.7: Concordance rate of each participant using the five focused object combinations with the GMM-based method and *SegU* for the two games. B: the ball; P_n: the player who will receive the ball next; P_{all}: all players except goal keepers; $B + P_n$: the combination of B and P_n. $B + P_{\text{all}}$: the combination of B and P_{all}.

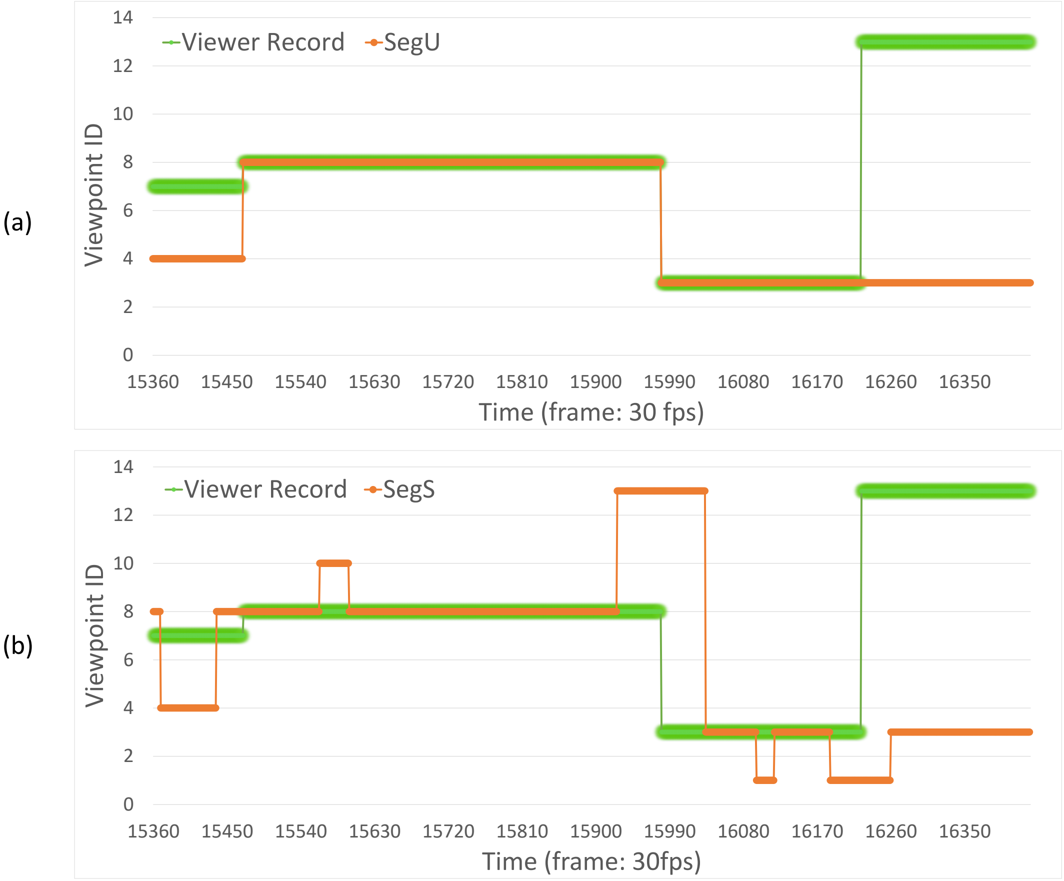


Figure 5.8: Sample of comparison on the generated recommended sequences using the equal segmentation with sliding window (*SegS*) and the ideal segmentation (*SegU*) for a user. *Viewer Record* represents the viewpoint selection record of the user. (a) Viewpoint sequence generated using *SegU* and user's record. (b) Viewpoint sequence generated using *SegS* and user's record.

two games, respectively. Thus, we can confirm that the ideal segmentation achieved better concordance rates than using the equal segmentation since the same tendency is also shown in the *Hist-SVM* and the *BoW-SVM* based methods.

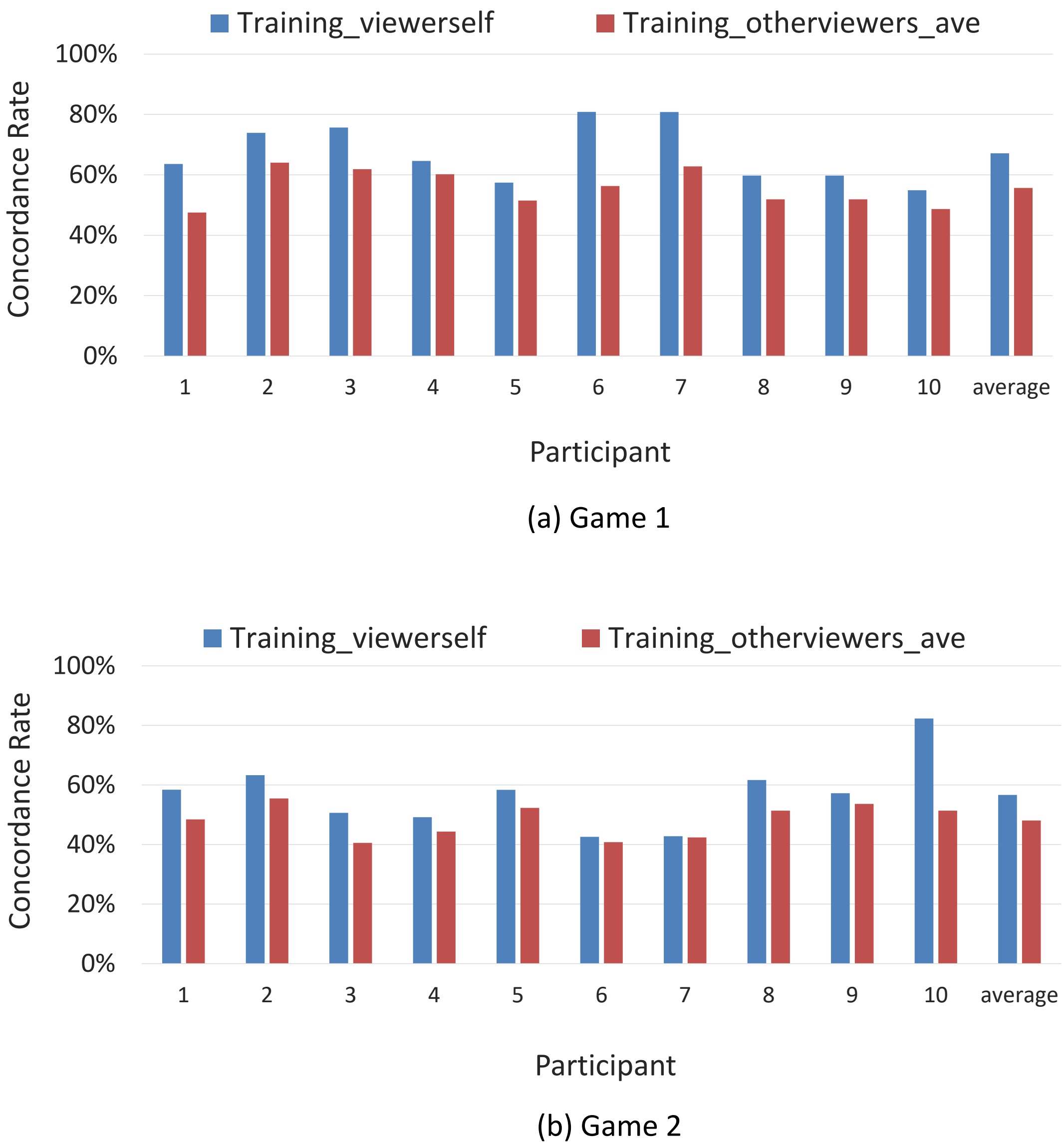
Let's compare the generated recommended sequences to investigate the difference between *SegS* and *SegU*. Figure 5.8 shows a sample of recommendations using *SegS* and *SegU*. We can see that the recommendation using *SegS* switched more frequently with shorter duration than *SegU*. Thus, we assume that smoothing the generated sequences or applying dynamic adaptation of window size according to the game context would lead to possible improvement on the performance.

5.6.3 Need of Personal Recommendation

The effectiveness of learning from each user's own record with learning from other users were compared to verify the need of personal recommendation. The results of learning the two kinds of record data using only the ball trajectory for each participant in the two games are shown in Figures 5.9 (a) and (b). In the figures, *Training_otherviewers_ave* represents the average performance using other participants' records by leave-one-participant-out cross validation. *Training_viewerself* represents the result of leave-one-sequence-out cross-validation using each participant's own record. The performance of learning from each participant was better than learning by other participants. This results show that each participant had a different viewing tendency against other participants and the proposed recommendation reflected their personal viewpoint selection tendency.

5.7 Summary

In this chapter, a viewpoint recommendation method focused on spatio-temporal context representation was proposed. Personal recommendation was generated by learning the relationship between the personal viewpoint selection tendency and the spatio-temporal scene context in the form of the trajectory distribution of focused objects. The experimental results showed that the GMM-based method outperforms other methods. For each method, the ball is the important focused object for viewpoint recommendation. Using only the ball trajectory or using combinations of the ball and the players' trajectories lead to better recommendation than only using the players' trajectories.



(a) Game 1

(b) Game 2

Figure 5.9: Comparing with results of learning using the other users' records for each user. *Training_viewerself* represents the result of learning from each participant's own record. *Training_otherviewers_ave* represents the average performance of learning from other participants' records.

Chapter 6

User Modelling for Group-based Viewpoint Recommendation

6.1 Introduction

As introduced in Chapter 1, in this thesis, the viewpoint recommendation problem is converted to a learning problem by using the viewpoint selection records of users as learning labels to generate user-dependent viewpoint recommendation (Figure 1.2). In this framework, there is a problem called the *cold-start* problem which is the lack of sufficient user records for model construction. Thus, this chapter focuses on overcoming the *cold-start* problem by acquiring a user model for group-based recommendation through relation with user attributes.

Some works select appropriate viewpoints by assessing the quality of the viewing experience in terms of video context (Figure 6.1.(a)) [Cai and Aggarwal, 1999, Daniyal et al., 2010]. These works mainly targeted generic recommendations related to the viewing content based on common sense or professional rules, instead of user diversity. Some existing works on personal viewpoint recommendation (Figure 6.1.(b)) learned or optimized the weight parameters for feature combinations from the user's viewpoint selection records [Muramatsu et al., 2014, Wang et al., 2015]. However, these works have difficulty in practice or limited performance owing to the difficulty in acquiring sufficient personal records for model construction.

Compared to these works, this chapter introduces user attribute information into viewpoint recommendation. The quality of experience of an application or a service based on the utility and/or enjoyment, which in turn depend on the user's personality and current state [Le Callet et al., 2012]. Thus, there is an assumption

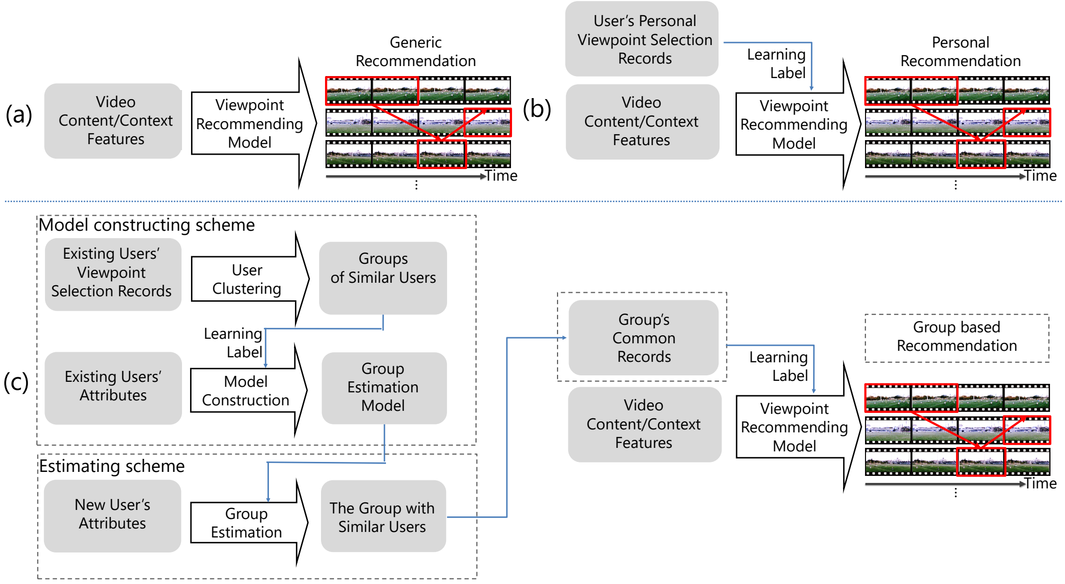

(a)
Video Content/Context Features
Viewpoint Recommending Model
Generic Recommendation
Time
(b)
User's Personal Viewpoint Selection Records
Video Content/Context Features
Learning Label
Viewpoint Recommending Model
Personal Recommendation
Time
(c)
Model constructing scheme
Existing Users' Viewpoint Selection Records
User Clustering
Groups of Similar Users
Learning Label
Existing Users' Attributes
Model Construction
Group Estimation Model
Estimating scheme
New User's Attributes
Group Estimation
The Group with Similar Users
Group's Common Records
Video Content/Context Features
Learning Label
Viewpoint Recommending Model
Group based Recommendation
Time

Figure 6.1: (a) Traditional video context based generic viewpoint recommendation approach. (b) Traditional user-record based personal viewpoint recommendation approach. (c) Proposed group-based viewpoint recommendation approach considering user attributes.

that the user's attributes, including personality, interest, and experience level of the viewing content, is related to the user's viewpoint selection tendency and influences the quality of the viewing experience of the recommendation.

To deal with the limitations of existing approaches, the proposed method considers constructing a recommendation model by learning from existing records of users with similar viewpoint selection tendencies and similar attributes. If existing users can be clustered into groups based on their viewpoint selection tendencies and a group can be estimated for a new user based on the user's attributes, it will be possible to generate a recommendation based on the existing group members' viewing records, instead of using personal records. In this chapter, this kind of recommendation is defined as a group-based recommendation, which is shown in Figure 6.1(c).

A group-based recommendation is proposed utilizing user attributes, enhancing the practicability even in the absence of sufficient personal records for training conventional recommendation models. For this group-based recommendation, the similarity of the users' viewpoint selection records is analyzed and the users are clustered into meaningful groups. This is not only useful for group-based recommendations, but also for user analysis in multi-view video viewing. Next, a group estimation model is constructed to classify an attribute vector of a user into one of the generated groups. Whether there are some discernible groups in users' attributes and viewpoint selection tendencies is also investigated.

The remainder of this chapter is organized as follows. Section 6.2 reviews related works. The proposed method is introduced in Section 6.3. Section 6.4 introduces the dataset used in this chapter. Experimental results are discussed in Section 6.5. Section 6.6 summarizes this chapter.

6.2 Related Work

Existing works that generate generic recommendations related to the viewing content [Cutler et al., 2002, Zhang et al., 2008, Ranjan et al., 2010], neglected user diversity and personality. There are only a few existing works on personal viewpoint recommendation [Muramatsu et al., 2014, Wang et al., 2015]. They learned the object's features, or optimized the weight parameters for feature combinations according to each user's viewpoint selection records. However it might be difficult to employ them in practice or they may have limited performance when the personal viewing records are not sufficient for training the model. Therefore, in this chapter, the proposed method extends the trajectory-based viewpoint recommendation model to a group-based recommendation model, instead of learning personal viewing records.

Existing viewpoint selection works have not investigated the relationship between a user's viewpoint selection tendency and his/her attribute information, such as personality, interest, and experience in the content. Because the experience quality of an application or a service is related to a user's personality [Le Callet et al., 2012], this work investigates whether the user personality can be used for viewpoint recommendation. The so called "Big Five personality traits" are commonly used to represent a user's personality. These five traits are Extraversion, Agreeableness, Conscientiousness, Emotional Stability and Openness to Experience [Hilgard et al., 1975, Barrick and Mount, 1991]. Regarding personality, Wang et al. found that Extraversion has an influence on the perceived enjoyment of a user's blogging activity, including video materials [Wang et al., 2012]. Zhu et al. proposed a model to predict a quality of experience assessment, combining video content assessment with individual user characteristics, such as interest and personality [Zhu et al., 2016].

There is an assumption that user attributes, such as personality and interest, have a relationship with viewpoint selection tendencies and can be used for viewpoint recommendation.

6.3 Group-based Viewpoint Recommendation by User Attributes

In this chapter, a novel viewpoint recommendation framework is proposed as illustrated in Figure 6.1(c).

A viewpoint recommendation model is constructed using video context features in a traditional approach, as illustrated in Figure 6.1(b). Considering the difficulty of acquiring a user's viewpoint selection records, the proposed method attempts to use the common records of a group consisting of existing users with similar viewpoint selection tendencies.

To achieve this group-based viewpoint recommendation, the proposed scheme consists of two parts: clustering existing users into groups based on similarity in viewpoint selection tendencies (Section 6.3.1), and constructing the group estimation model using user attributes as features (Section 6.3.2). Thus, for a new user, it becomes possible to estimate which group the user should belong to considering their attributes. Accordingly, a recommendation model can be constructed by learning the common records of users in the group, which is introduced in Section 6.3.3.

6.3.1 User Grouping based on Viewpoint Selection Tendency

In this section, the proposed method first investigates whether there exists a difference in viewpoint selection tendencies among different users, and calculates the similarities between them. Then the users are clustered into different groups based on the similarities.

Analysis of Viewpoint Selection Tendency

Users' viewpoint selection tendencies can be represented by their viewpoint selection records. A sample of the viewpoint selection records of different users acquired from a video-editing experiment with multi-view soccer game videos is shown in Figure 6.2. It can be observed that both similarities and differences exist among the viewpoint selection records of different users.

To quantify the similarity or difference in the viewpoint selection tendencies

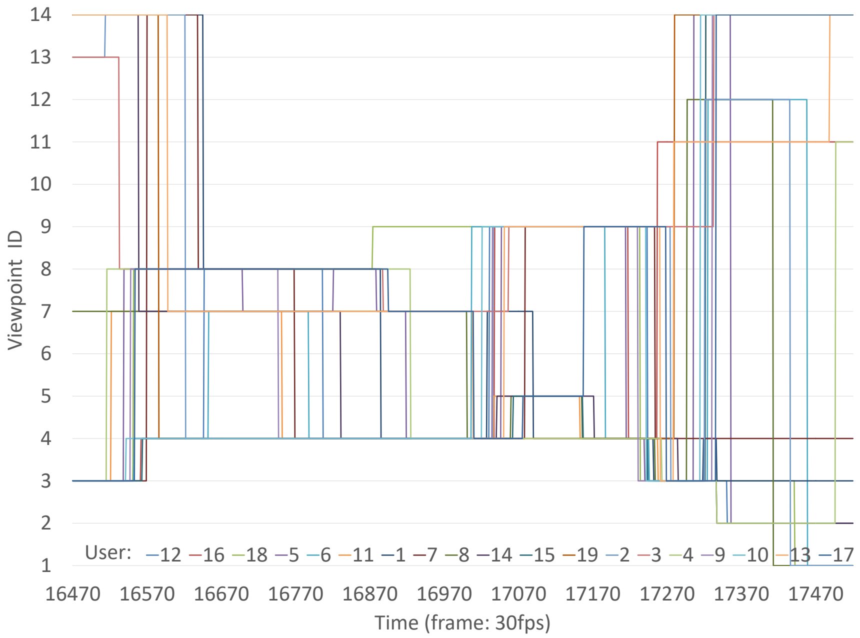


Figure 6.2: Viewpoint selection records of different users in a video sequence.

between users, the sameness of the camera positions at each frame t is calculated as a similarity score between the viewpoint sequences of each pair of users (u_a and u_b) as follows:

$$E(t) = \begin{cases} 1 & (\text{if } u_a(t) = u_b(t)), \\ 0 & (\text{otherwise}). \end{cases} \tag{6.1}$$

$$S(u_a, u_b) = \frac{\sum_{n=1}^{N} \sum_{t=1}^{L_n} E(t)}{N}, \tag{6.2}$$

where L_n is the length of the n-th video sequence and N is the number of video sequences. The confusion matrix of the similarity score between each pair of users is shown in Figure 6.3. Based on this figure, we can see that some users are highly similar with each other (e.g., users no. 6, 12, 15, and 16) while some users are distinctly different from other users (e.g., users no. 1, 7, and 13). Thus, it is possible to cluster users into different groups according to the similarity score.

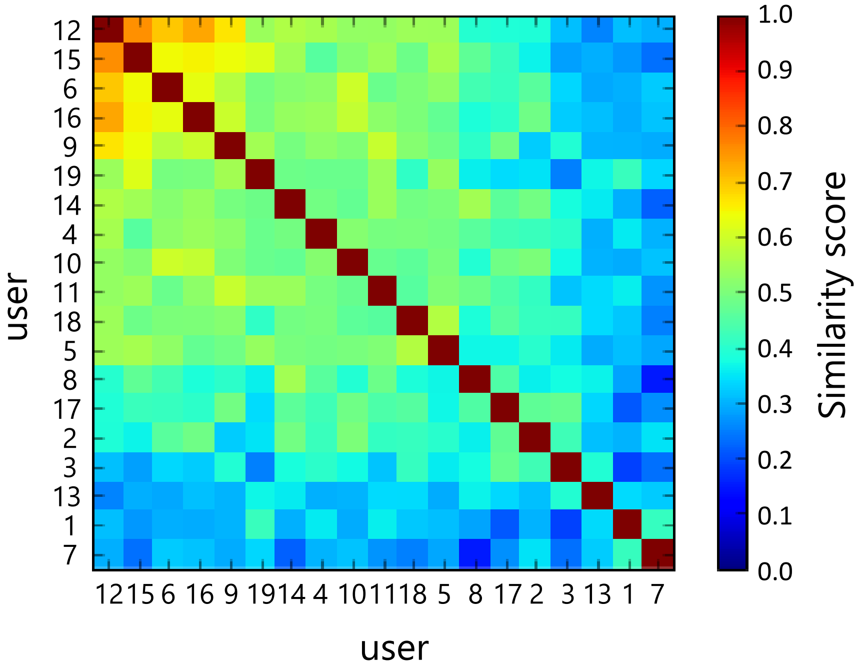


Figure 6.3: The confusion matrix of the similarity score between the viewpoint selection records of each pair of users.

User Clustering

User clustering is conducted according to the similarity scores of users' viewpoint selection records. For user u_i, the similarity scores calculated from a comparison with each user (including him/herself) are used as a feature vector $R_{u_i} = \{S(u_i, u_1), \ldots, S(u_i, u_I)\}$ of the user, where I is the number of users. An agglomerative hierarchical clustering is applied to the feature vector using the Euclidean distance as the measure of distance between the feature vectors with Ward's method as the linkage criterion.

6.3.2 User Group Estimation based on User Attributes

This section introduces the group estimation method based on user attributes. This makes it possible to provide a group-based recommendation for even a new user without personal viewpoint selection records. Each cluster is used as a target for the group estimation.

The user attribute characteristics are used in the form of a feature vector for group estimation. They are based on the Big Five personality traits and consist of one score per trait. The proposed method also includes users' interests in the target sport (soccer in this thesis), which is assessed on a seven-point scale via a questionnaire described in Section 6.4.1. Furthermore, there is an assumption that the experience level and the viewing frequency of the target sport also have an influence on the viewpoint selection. Thus, these are added into the feature vector. The values of these characteristics are normalized into a distribution of average 0 and variance 1 for each characteristic to generate the feature vector. Supervised learning by a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel is conducted for the group estimation.

6.3.3 Group-based Viewpoint Recommendation

The proposed method extends the viewpoint recommendation method based on object trajectory introduced in Chapter 5 by learning the common records of the group users for a group-based viewpoint recommendation.

The proposed method uses a machine learning method to learn the relationship between the trajectory distribution of the focused object and viewpoint selection tendencies of group members to achieve a group-based viewpoint recommendation. A Gaussian Mixture Model (GMM) was used to represent the ball-trajectory distributions of the different viewpoints by processing several frames segmented from the video sequence for each cut. In the following discussion, a sub-trajectory of the ball in cut C_i is represented by $T_{C_i} = \{\mathbf{x}_t | t = 1, 2, \ldots, F_i\}$, where i is the cut index and $\mathbf{x}_t \in \mathbb{R}^2$ is a point on the field coordinate system at frame t, and F_i is the length (number of frames) of cut C_i.

Here, the most selected viewpoint at each frame along the video duration are first extracted as the common records from users in the same group. For each viewpoint, the ball trajectories of cuts while the viewpoint v is selected are gathered according to the common records, and represented as T_v. The proposed method expresses T_v using GMM, which is a linear combination of several Gaussian components as follows:

$$p_v(\mathbf{x}_t) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}_t | \mu_k, \Sigma_k), \tag{6.3}$$

where $\mathbf{x}_t$ is a sample from T_v in the training dataset. K is the number of the Gaussian components, π_k is the weight of the k-th Gaussian component with $\sum_{k=1}^{K} \pi_k = 1$, $N(\mathbf{x}_t | \mu_k, \Sigma_k)$ is the Gaussian component density with parameters μ_k (mean) and

Σ_k (covariance). The parameters (π_k,μ_k and Σ_k) are estimated by applying the EM algorithm. The number of components is based on experimental results. Thus, the model of each viewpoint can be generated from the common records.

For each video sequence in the test data, it is first divided into cuts and ball trajectory T_{C_i} of cut C_i are extracted. The grand total of the log-likelihood for the points on the trajectory T_{C_i} is calculated under the generated GMM of each viewpoint. Viewpoint R is recommended for the group with the largest likelihood for each cut trajectory as follows,

$$R(T_{C_i}) = \underset{1 \le v \le V}{\arg\max} \sum_{\mathbf{x}_t \in T_{C_i}} \log p_v(\mathbf{x}_t). \tag{6.4}$$

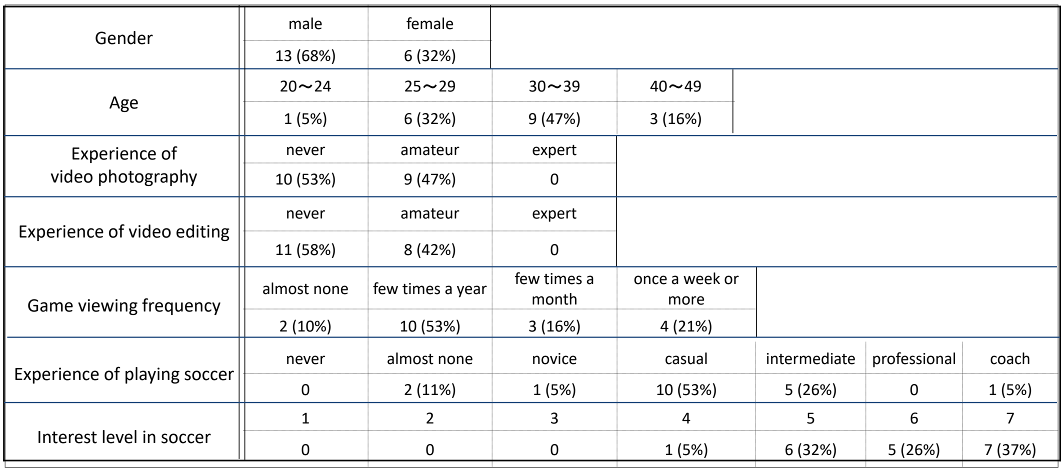
In this way, the proposed method can overcome the *cold-start* problem by providing viewpoint recommendation based on existing users' records with similar viewpoint selection tendencies to the user even without sufficient personal record.

6.4 Experiment

In this chapter, additional users' selection records are collected by conducting the same multi-view video editing experiment using multi-view video dataset of Game 1 and the same steps described in Chapter 3. The user attribute information was acquired through a questionnaire prior to the video editing experiment. Besides using a questionnaire, user attributes such as the sex, age, and personality could also be estimated through their behavior in both real world and social networks [Yi et al., 2015, Guntuku et al., 2015].

6.4.1 Participants

Experimental participants were first asked to state their profile information via a profile questionnaire. The information acquired from the 19 participants is summarized in Table 6.1. The participants included 13 males and 6 females of ages between 20 and 49. In the questionnaire, the interest level in soccer was assessed on a "Likert scale ranging" from "disagree strongly" to "agree strongly" in response to the statement "I am interested in soccer." All participants had a certain interest in soccer ($\geq$ 4: Neither agree nor disagree) and over 37% indicated that they strongly agreed to the statement. For the viewing frequency of the soccer game, 53% of the participants were occasional users, at least a few times a year, and 37% of the participants



Gender	male	female					
	13 (68%)	6 (32%)					
Age	20～24	25～29	30～39	40～49			
	1 (5%)	6 (32%)	9 (47%)	3 (16%)			
Experience of video photography	never	amateur	expert				
	10 (53%)	9 (47%)	0				
Experience of video editing	never	amateur	expert				
	11 (58%)	8 (42%)	0				
Game viewing frequency	almost none	few times a year	few times a month	once a week or more			
	2 (10%)	10 (53%)	3 (16%)	4 (21%)			
Experience of playing soccer	never	almost none	novice	casual	intermediate	professional	coach
	0	2 (11%)	1 (5%)	10 (53%)	5 (26%)	0	1 (5%)
Interest level in soccer	1	2	3	4	5	6	7
	0	0	0	1 (5%)	6 (32%)	5 (26%)	7 (37%)

Table 6.1: User attribute information acquired from the questionnaire.

viewed a few times a month or even more than once a week. Moreover, the largest participant group had a casual level of soccer playing experience (53%), followed by the participant group with intermediate experience (26%). Furthermore, nearly half of the participants had amateur video photography experience with a video camera, or video editing experience using an editing software.

6.4.2 Collection of Users' Personality

Following the profile questionnaire, participants were then asked to complete a ten-item personality measure (TIPI) questionnaire (also known as BFI-10), to establish their own personality, according to the Big Five personality traits. In the TIPI questionnaire, each item is assessed on a seven-point scale, and each trait is measured by a pair of opposite items [Gosling et al., 2003]. For example, the Openness to Experience trait was quantified by adding up the self-assessment of the user on the positive item "open to new experience, complex" and the inverse of the self-assessment on the negative item "conventional, uncreative." The average score for the five traits was 4.80 (StD = 0.99) out of 7.0, and had a range from 2.5 to 7.0.

6.4.3 Evaluation Framework

The evaluation was conducted in the following three kinds of recommendation frameworks, respectively: personal recommendation, group-based recommendation, and recommendation from another user.

For the personal recommendation to each user, a leave-one-sequence-out cross-validation was conducted by using one sequence of each user's viewing records as test data until all the sequences were used as test data. For the recommendation from another user or group-based recommendation to each user, the same leave-one-sequence-out cross-validation was also conducted using one sequence of each user's viewing records as test data while using the sequences of other user's records or each group's common records as training data.

The recommended viewpoints are then compared with the corresponding records to calculate the average concordance rate of all the test data at each frame.

6.5 Results and Analysis

6.5.1 User Grouping

The dendrogram of the user clustering result obtained by the proposed method based on the similarity in users' viewpoint selection records, is shown in Figure 6.4(b). The following discussion analyzes the result of the four group clustering indicated by a red dotted line in the dendrogram.

To easily visualize the relationship between the clustered groups and the similarity score, a confusion matrix of the similarity score between viewpoint selection records of each pair of users grouped with squares is shown in Figure 6.4(a). We can see that the users were clearly grouped according to the similarity.

To evaluate the user clustering result, a confusion matrix of the concordance rates of the personal recommendation is first compared with recommendation from another user, as shown in Figure 6.4(c), with the clustering dendrogram of the recommendation concordance rates in Figure 6.4(d) to investigate the meaning of the different groups. As Figures 6.4(a) and (c) show, the effectiveness of recommendation has a trend similar to the user similarity in regard to the viewpoint selection tendencies. Moreover, the two cluster types are almost the same, only having a difference with user no. 9. Namely, the users in Group 1, with a high similarity in viewing records (users no. 6, 12, 15, and 16), also achieved high recommendation concordance rates when the records of other group members were used as training data. The users in Group 4 with low similarity in viewing records (users no. 1, 7, and 13), could not acquire high recommendation concordance rates when using other users' records, whether within or without the group.

The effectiveness of the proposed group-based recommendation is then evaluated

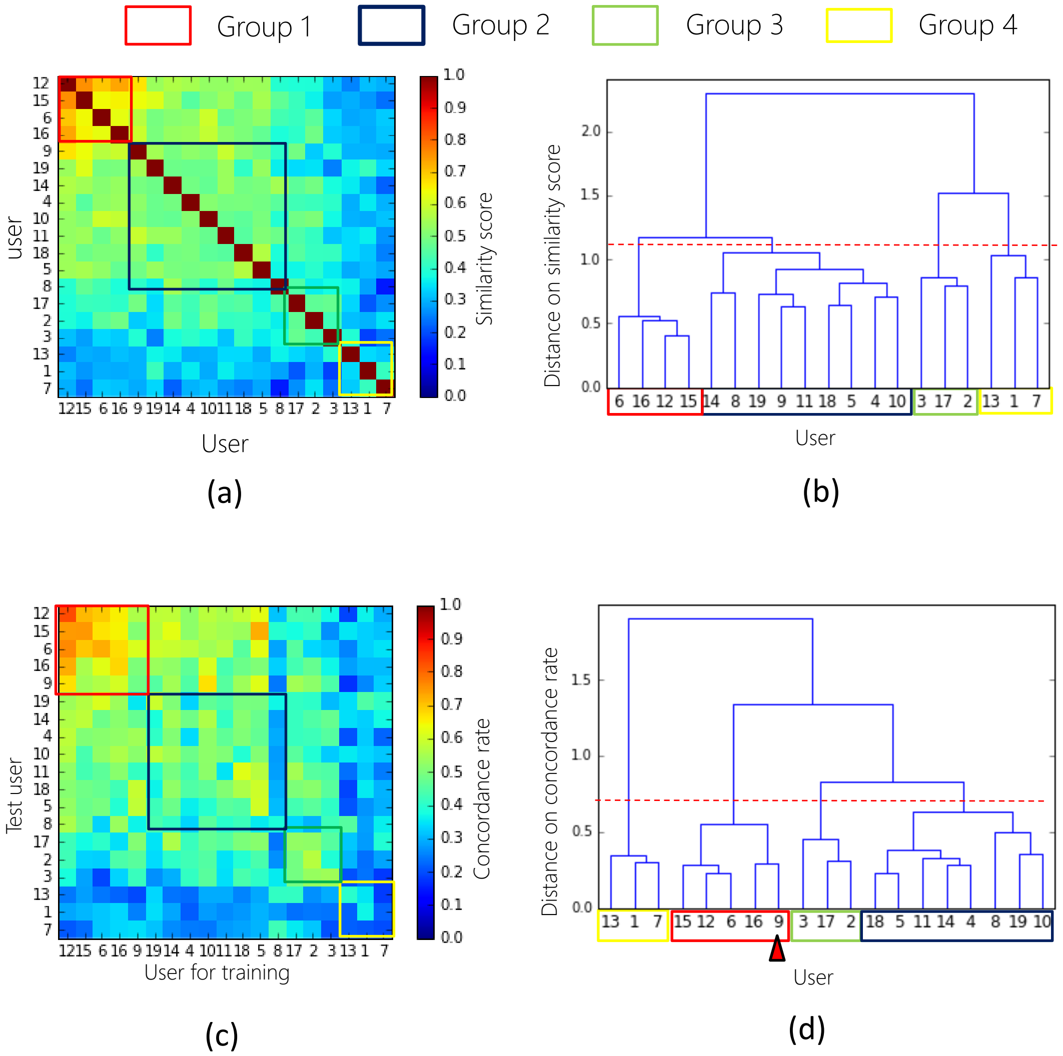


Figure 6.4: (a) Confusion matrix of the similarity score between viewpoint selecting records of each pair of users. (b) User clustering result based on similarity score of users. (c) Recommendation concordance rate of each test user when using each user's records as training data. (d) User clustering result based on recommendation concordance rate of each test user.

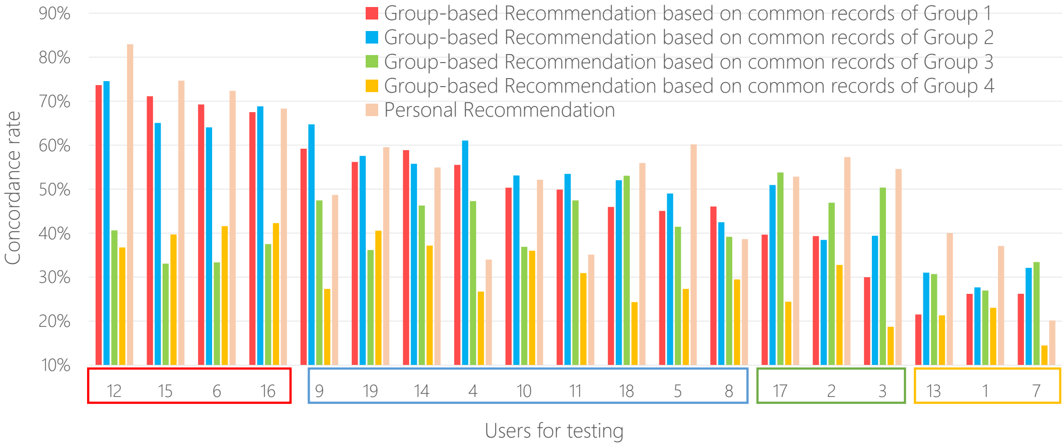


Figure 6.5: Concordance rates of each user as a test user by the group-based recommendation from each group and the personal recommendation.

and compared with the personal recommendation. The recommendation model of each group was constructed using the extracted common user records within the generated group. The concordance rates of the group-based recommendation, using the constructed models of the four groups for each user, while using their own data as test data, are shown in Figure 6.5. It also shows the concordance rates of the personal recommendation. The average concordance rates of each user group are summarized in Figure 6.6. From the graph, we can see that, for most users, the group-based recommendation from their own group yields better recommendation effectiveness than the one from other groups. This trend can be seen in the average rates of Groups 1, 2, and 3, compared with the concordance rates of the personal recommendation for each group in Figure 6.6, we can see that although the performance of the group-based recommendation for Groups 1, 3, and 4 is lower than the personal recommendation, the difference is not very distinct. However, that for the group-based recommendation for Group 2 is higher than the personal recommendation.

It can be inferred from these trends that different user groups have different viewing patterns and viewpoint recommendation needs. For some users, especially within the high similarity group, there exists a stable pattern in viewpoint selection tendencies. A better viewing experience can be acquired by not only personal recommendation, but also group-based recommendation, which is effective when the personal recommendation is difficult to realize. For those users in Group 2, the group-based recommendation from similar users was more effective than the personal recommendation. This shows that their viewing patterns were possibly not

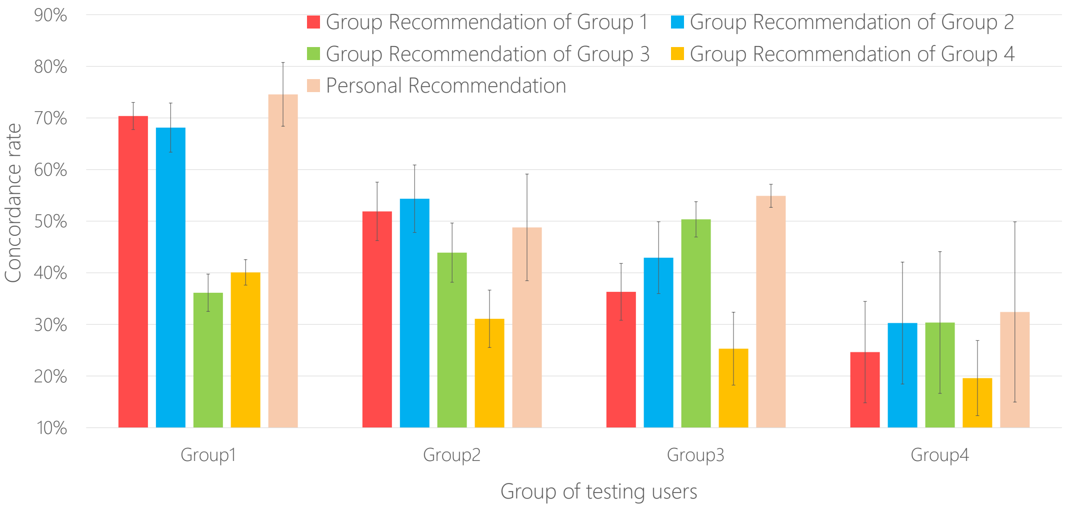


Figure 6.6: Average concordance rates of users in each group using different recommendation models.

stable enough and too difficult to be modeled by small-scale personal data. The group-based recommendation is more practical for these types of users. However, for the users with low similarity to most other users, such as those in Group 4, their viewpoint selection tendencies may not have a pattern consistent enough to construct any user-dependent recommendation model. For these types of users, a generic recommendation based on general rules may yield better performance.

6.5.2 User Attribute Analysis

The user attributes included in each group are analyzed to investigate the relationship between the users' attributes and viewpoint selection tendencies.

To easily analyze the user attributes, the TIPI scores on the Big Five personality traits and other attribute scores are z-normalized to have zero mean and unit variance. Then, the normalized scores of each user group are gathered and the average and the standard deviation for each distribution in each group are calculated as shown in Figure 6.7.

Regarding the Big Five personality traits, from the graph, we can see that the average TIPI scores of all traits in Group 1 had a trend higher than the average of all users (value > 0), while the average TIPI scores of most traits in Groups 3 and 4 had a trend lower than the average of all users (value < 0). In particular, there was significant difference between Groups 1 and 2 on Emotional Stability trait

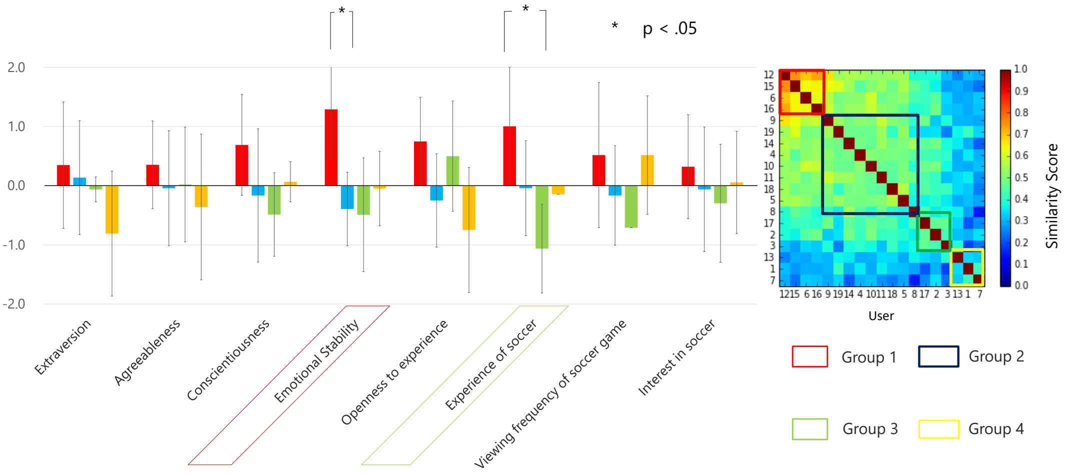


Figure 6.7: Z-normalized average and standard deviation on personality TIPI scores and other attribute scores of users in different user groups are shown in different color. The confusion matrix of the similarity score between viewpoint selecting records of each pair of users grouped with squares is shown on the right for reference. The colors of squares in the confusion matrix correspond to the colors of user attributes scores on the left.

by pairwise comparisons using T-test with Bonferroni adjustment. Thus, it can be inferred that users with stable emotion had a high degree of similarity in viewpoint selection records with a high recommendation effectiveness. However, the viewing patterns of users with low Emotional Stability were not similar to those of other users and were difficult to model because of the possible instability of their selection tendencies.

For the other user attributes, the distribution of experience and interest and viewing frequency of the target sport can also be seen in Figure 6.7. Certain characteristics can be found: regarding Group 1 with a high similarity in viewing records, all users in the group had a higher-level soccer experience and interest with higher viewing frequency (specifically 75% had an above-casual level of soccer experience, as shown in Table 6.1) than the other groups. Moreover, they were all male with experience in video photography and editing. For Group 2, 66% had a low soccer game viewing frequency (less than a few times a month). All users in Group 3 had a low soccer game viewing frequency and a low level of soccer experience, with no experience in video photography and editing. In Group 4, all users were female with a casual soccer experience level. In particular, the level of soccer experience of users in Group 1 was significantly higher than these in Group 3.

Thus, it can be summarized that the users with high emotional stability in personality, having high-level experience in the target sport, and with experiences in video photography and editing, have a high possibility of having consistent viewing patterns and receive better user-dependent viewpoint recommendations, whereas users with unstable emotion and less experience in the target sport, had the opposite trend in viewpoint selection.

6.5.3 User Group Estimation

This section evaluates the performance of group estimation, which includes user attributes in the feature vector and applies supervised classification by an SVM with an RBF kernel with a one-vs-rest method. A leave-one-user-out cross-validation was conducted for evaluation by using one user's attributes as test data and the other users' attributes as training data until all users were used as test data. Auto-balanced class weights are used to deal with the balance problem of different group sizes.

The average accuracy on group classification for all test data achieved 52.6%. For Group 2 with the most users, the estimation accuracy was 77.8%, while only 33.3% for Group 3 and can not estimate for Group 4 with less users. The insufficient user data scale is a possible reason for the unideal average accuracy.

6.6 Summary

In this chapter, a group-based recommendation method was proposed which considers users' viewpoint selection tendencies and attributes, including personality, interest, and experience level of the target sport. Users were clustered into groups based on similarities between their viewing records and a group estimation model was constructed using user attributes. The group-based recommendation from user's member group yielded better recommendation effectiveness than the one from other groups, and yielded similar effectiveness to personal recommendation. The users with high emotional stability in personality and high-level experience in the target sport, have a high possibility of having consistent viewing patterns and receive better user-dependent viewpoint recommendations. However, the proposed recommendation method has limited performance for some users with unique or unstable viewing patterns compared to other users.

Chapter 7

Conclusion

7.1 Summary

In this thesis, a context-aware user-dependent viewpoint recommendation framework was proposed to provide automatic viewpoint recommendation adapted to diverse video contexts and user contexts. In general, a user would select the best viewpoint to enjoy the event according to various video context and user context reflected by his/her viewpoint selection tendency related to personal experience and interests. To generate a viewpoint recommendation adaptive to diverse video and user contexts with high similarity to the user's viewpoint selection tendency, various factors were considered in the recommendation framework: context-dependent learning scheme, spatio-temporal feature representation, and user modeling for group-based recommendation.

First, an adaptive method to increase the correspondence between a user's viewpoint selection tendency and the viewing context was proposed by a *context-dependent learning scheme*. Viewpoint evaluation and transition processes were used to compute the contextual information of a scene and production using different weight parameters. This method selects context-dependent optimal sets of viewpoints by optimizing the parameters to represent different scene contexts and production contexts for both common and personal recommendation. The generated context-dependent recommendation outperformed the context-independent recommendations compared with actual selections made by users, which shows the effectiveness of this method. Moreover, the method can analyze or interactively alter user-specific viewpoint selection tendencies in different scene contexts. However, the performance could be limited if the video context is too complex to be represented mainly by individual spatial information, such as the strategy of different teams.

Next, a method to improve the effectiveness of context representation by *spatio-temporal feature representation* was proposed. This method learns the relationship between the user's personal viewpoint selection tendency and the spatio-temporal scene context represented by object trajectories. Three methods including GMM-based, SVM-Hist, and SVM-BoW were compared and the GMM-based method achieved the best performance. It assesses the degree of similarity between the recommendation and user selection records. However, since trajectory distribution could not handle, for example, direction information of an object's action, it will be difficult to be applied to applications which require detailed object action recognition and representation. Time-series models can represent some contextual information that trajectory distribution could not cover, such as direction information of object action. Applying such time-series model for scene context could improve the presented results. Besides, since this method is highly dependent to user's viewpoint selection record, the recommendation will be limited if the user's record is absent or not sufficient.

A method to overcome the *cold-start* problem of user-dependent recommendation is also proposed by generating and applying a *user model for group-based recommendation*. This method focused on the difficulty in acquiring sufficient personal viewing records in practice and the importance of the user's attribute information, such as the user's personality, interest, and experience level in the target sport. A group-based recommendation framework was proposed, which consists of a user grouping approach based on the similarity in viewpoint selection records of existing users, and a group estimation method based on the classification by user attributes. The group-based recommendation from user's member group yielded better recommendation effectiveness than the one from other groups, and yielded similar effectiveness to personal recommendation, so it can be used when personal recommendation is not avaliable. The proposed recommendation method worked better for users with high emotional stability and high-level experience in the target sport, who have a high possibility of having a stable viewing pattern. In the future, the number of groups should be improved by automatic, and more detailed analysis on the user's viewpoint selection tendency.

Other than the soccer game, the method can be easily applied to other field ball games such as baseball, basketball, hockey, etc. It is also probably to apply the proposed method to other events with region-dependent contexts containing primary objects of preference, such as large musical shows, group dance, etc.

7.2 Future Directions

Here, two future applications on the multi-view video field are introduced.

7.2.1 Multi-modal Feedback Analysis and On-line Updating

First, a method to use multi-modal feedback information to evaluate and improve the practical realization of a multi-view viewpoint recommendation system is considered.

For the *cold-start* problem, it was shown that the group-based method could provide a good initial recommendation. On-line updating should further improve the quality of the initial recommendation for a more personalized recommendation. For this, two problems are needed to be solved. One is how to detect the difference between the user's viewpoint selection tendency and the initial recommendation. Another is how to realize on-line updating according to the difference. For the detection, besides the direct viewpoint switching operation from a user (i.e., user click-through), multi-modal feedback information acquired from users may be more sensitive and convenient, such as the biological signal sensing, gaze movement, facial expression, etc. Thus, appropriate feedback information that can sensitively reflect the variation of a user's viewpoint selection tendency should be investigated. Furthermore, the switching operation should be more convenient by providing switching candidates in the appropriate timing by modeling the relationship between a user's multi-modal feedback information and his/her viewpoint selection tendency. Accordingly, on-line updating techniques could be developed using the switched records and multi-modal feedback based on the proposed recommendation model.

7.2.2 User-dependent Free-viewpoint Video Navigation

Next, let's consider the future application on free-viewpoint videos. Free-viewpoint videos allow users to interactively control the viewpoint and generate new views of a dynamic scene from any 3D position. It can enhance user viewing experience and interaction with others in the common space, such as make the TV viewing into a group activity. Currently, many works are focusing on generating a high quality free-viewpoint video or virtual reality application [Debevec et al., 1998, Ohm and Müller, 1999, Lawrence Zitnick et al., 2004, Kameda et al., 2004, Kim et al., 2006, Kilner et al., 2006, Horiuchi et al., 2012, Matsuyama et al., 2012]. There is a large potential to provide better experience at the viewpoint of viewers. User-dependent viewpoint navigation is also important and necessary for free-viewpoint videos due to numerous options for viewing. Compared to multi-view videos, free-viewpoint navigation should pay more attention to a user's experience while viewing, such as the representation performance for immersion experience, visually induced motion sickness during viewpoint changing, more interactive navigation interface through audio-visual approaches, and interaction support among different users in the common space. Thus, the viewpoint navigation not only needs to analyze and represent the viewing content with a user's interest the same as in multi-view navigation, but also detect his/her attention and status in real time by means of, for example, eye gaze change, emotion change, etc. Moreover, regarding the development of the social network and potential interaction application of free-viewpoint videos, user profile and behavior on social network may become an effective source of information for better user-dependent navigation and other applications.

Bibliography

[Ahmad, 2007] Ahmad, I. (2007). Multi-view video: Get ready for next-generation television. *IEEE Distributed Systems Online*, 8(3):1–6.

[Ariki et al., 2006] Ariki, Y., Kubota, S., and Kumano, M. (2006). Automatic production system of soccer sports video by digital camera work based on situation recognition. In *Proceedings of the 8th IEEE International Symposium on Multimedia*, pages 851–860.

[Ariki et al., 2008] Ariki, Y., Takiguchi, T., and Yano, K. (2008). Digital camera work for soccer video production with event recognition and accurate ball tracking by switching search method. In *Proceedings of the 2008 IEEE International Conference on Multimedia and Expo*, pages 889–892.

[Babaguchi, 2000] Babaguchi, N. (2000). Towards abstracting sports video by highlights. In *Proceedings of the 2000 IEEE International Conference on Multimedia and Expo*, volume 3, pages 1519–1522.

[Barrick and Mount, 1991] Barrick, M. R. and Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1):1–26.

[Basu et al., 1998] Basu, C., Hirsh, H., and Cohen, W. (1998). Recommendation as classification: Using social and content-based information in recommendation. In *Proceedings of the 15th AAAI National Conference on Artificial Intelligence*, pages 714–720.

[Bocconi et al., 2005] Bocconi, S., Nack, F., and Hardman, L. (2005). Vox populi: A tool for automatically generating video documentaries. In *Proceedings of the 16th ACM Conference on Hypertext and Hypermedia*, pages 292–294.

[Bramberger et al., 2005] Bramberger, M., Quaritsch, M., Winkler, T., Rinner, B., and Schwabach, H. (2005). Integrating multi-camera tracking into a dynamic task allocation system for smart cameras. In *Proceedings of the 2005 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 474–479.

[Cai and Aggarwal, 1999] Cai, Q. and Aggarwal, J. K. (1999). Tracking human motion in structured environments using a distributed-camera system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1241–1247.

[Chauhan et al., 2016] Chauhan, D., Patel, N. M., and Joshi, M. (2016). Automatic summarization of basketball sport video. In *Proceedings of the 2nd IEEE International Conference on Next Generation Computing Technologies*, pages 670–673.

[Chen et al., 2012] Chen, B., Wang, J., Huang, Q., and Mei, T. (2012). Personalized video recommendation through tripartite graph propagation. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1133–1136.

[Chen et al., 2013] Chen, C., Wang, O., Heinzle, S., Carr, P., Smolic, A., and Gross, M. (2013). Computational sports broadcasting: Automated director assistance for live sports. In *Proceedings of the 2013 IEEE International Conference on Multimedia and Expo*, pages 1–6.

[Chen et al., 2011] Chen, F., Delannay, D., and De Vleeschouwer, C. (2011). An autonomous framework to produce and distribute personalized team-sport video summaries: A basketball case study. *IEEE Transactions on Multimedia*, 13(6):1381–1394.

[Cheung et al., 2011] Cheung, G., Ortega, A., and Cheung, N.-M. (2011). Interactive streaming of stored multiview video using redundant frame structures. *IEEE Transactions on Image Processing*, 20(3):744–761.

[Collins et al., 2002] Collins, R. T., Amidi, O., and Kanade, T. (2002). An active camera system for acquiring multi-view video. In *Proceedings of the 2002 IEEE International Conference on Image Processing*, pages 517–520.

[Csurka et al., 2004] Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Proceedings of Workshop on Statistical Learning in Computer Vision of ECCV*, volume 1, pages 1–2.

[Cui et al., 2014] Cui, P., Wang, Z., and Su, Z. (2014). What videos are similar with you?: Learning a common attributed representation for video recommendation.

In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 597–606.

[Cutler et al., 2002] Cutler, R., Rui, Y., Gupta, A., Cadiz, J. J., Tashev, I., He, L.-w., Colburn, A., Zhang, Z., Liu, Z., and Silverberg, S. (2002). Distributed meetings: A meeting capture and broadcasting system. In *Proceedings of the 10th ACM International Conference on Multimedia*, pages 503–512.

[Daniyal and Cavallaro, 2011] Daniyal, F. and Cavallaro, A. (2011). Multi-camera scheduling for video production. In *Proceedings of the 2011 European Conference on Visual Media Production*, pages 11–20.

[Daniyal et al., 2010] Daniyal, F., Taj, M., and Cavallaro, A. (2010). Content and task-based view selection from multiple video streams. *Multimedia Tools and Applications*, 46(2–3):235–258.

[Davenport et al., 1991] Davenport, G., Smith, T. A., and Pincever, N. (1991). Cinematic primitives for multimedia. *IEEE Computer Graphics and Applications*, 11(4):67–74.

[Debevec et al., 1998] Debevec, P., Yu, Y., and Borshukov, G. (1998). Efficient view-dependent image-based rendering with projective texture-mapping. In *Proceedings of the 9th Eurographics Workshop on Rendering*, pages 105–116.

[DeMenthon et al., 1998] DeMenthon, D., Kobla, V., and Doermann, D. (1998). Video summarization by curve simplification. In *Proceedings of the 6th ACM International Conference on Multimedia*, pages 211–218.

[Du et al., 2006] Du, W., Hayet, J.-B., Piater, J., and Verly, J. (2006). Collaborative multi-camera tracking of athletes in team sports. In *Proceedings of the 2006 Workshop on Computer Vision Based Analysis in Sport Environments*, pages 2–13.

[D'Orazio and Leo, 2010] D'Orazio, T. and Leo, M. (2010). A review of vision-based systems for soccer video analysis. *Pattern Recognition*, 43(8):2911–2926.

[Ekin et al., 2003] Ekin, A., Tekalp, A. M., and Mehrotra, R. (2003). Automatic soccer video analysis and summarization. *IEEE Transactions on Image processing*, 12(7):796–807.

[Esterle et al., 2014] Esterle, L., Lewis, P. R., Yao, X., and Rinner, B. (2014). Socio-economic vision graph generation and handover in distributed smart camera networks. *ACM Transactions on Sensor Networks*, 10(2):20:1–20:24.

[Fu et al., 2010] Fu, Y., Guo, Y., Zhu, Y., Liu, F., Song, C., and Zhou, Z.-H. (2010). Multi-view video summarization. *IEEE Transactions on Multimedia*, 12(7):717–729.

[Gandhi et al., 2014] Gandhi, V., Ronfard, R., and Gleicher, M. (2014). Multi-clip video editing from a single viewpoint. In *Proceedings of the 11th ACM European Conference on Visual Media Production*, pages 9:1–9:10.

[Godi et al., 2017] Godi, M., Rota, P., and Setti, F. (2017). Indirect match highlights detection with deep convolutional neural networks. *arXiv preprint arXiv:1710.00568*.

[Gong and Liu, 2003] Gong, Y. and Liu, X. (2003). Video summarization and retrieval using singular value decomposition. *Multimedia Systems*, 9(2):157–168.

[Gosling et al., 2003] Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6):504–528.

[Guntuku et al., 2015] Guntuku, S. C., Qiu, L., Roy, S., Lin, W., and Jakhetiya, V. (2015). Do others perceive you as you want them to?: Modeling personality based on selfies. In *Proceedings of the 1st ACM International Workshop on Affect & Sentiment in Multimedia*, pages 21–26.

[Hayashi et al., 2013] Hayashi, Y., Doman, K., Ide, I., Deguchi, D., and Murase, H. (2013). Automatic authoring of a domestic cooking video based on the description of cooking instructions. In *Proceedings of the 5th International Workshop on Multimedia for Cooking & Eating Activities*, pages 21–26.

[Hilgard et al., 1975] Hilgard, E. R., Atkinson, R. C., and Atkinson, R. L. (1975). *Introduction to Psychology*. Oxford and IBH Publishing.

[Hill et al., 1995] Hill, W., Stead, L., Rosenstein, M., and Furnas, G. (1995). Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 194–201.

[Hirayama et al., 2013] Hirayama, T., Marutani, T., Tanoue, D., Tokai, S., Fels, S., and Mase, K. (2013). Agent-assisted multi-viewpoint video viewer and its gaze-based evaluation. In *Proceedings of the 6th Workshop on Eye Gaze in ACM Intelligent Human Machine Interaction: Gaze in Multimodal Interaction*, pages 29–34.

[Ho and Oh, 2007] Ho, Y.-S. and Oh, K.-J. (2007). Overview of multi-view video coding. In *Proceeding of the 14th International Workshop on Systems, Signals and Image Processing, the 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services*, pages 5–12.

[Hochberg, 1988] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

[Horiuchi et al., 2012] Horiuchi, T., Sankoh, H., Kato, T., and Naito, S. (2012). Interactive music video application for smartphones based on free-viewpoint video and audio rendering. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1293–1294.

[Hu and Pu, 2010] Hu, R. and Pu, P. (2010). Using personality information in collaborative filtering for new users. In *Proceedings of the 2nd ACM Workshop on Recommender Systems and the Social Web*, pages 17–24.

[Hua et al., 2003] Hua, X.-S., Lu, L., and Zhang, H.-J. (2003). AVE: Automated home Video Editing. In *Proceedings of the 11th ACM International Conference on Multimedia*, pages 490–497.

[Itoda et al., 2015] Itoda, K., Watanabe, N., and Takefuji, Y. (2015). Model-based behavioral causality analysis of handball with delayed transfer entropy. *Procedia Computer Science*, 71:85–91.

[Iwatsuki et al., 2013] Iwatsuki, A., Hirayama, T., and Mase, K. (2013). Analysis of soccer coach's eye gaze behavior. In *Proceedings of the 2nd IAPR Asian Conference on Pattern Recognition*, pages 793–797.

[Jiang et al., 2008] Jiang, H., Fels, S., and Little, J. J. (2008). Optimizing multiple object tracking and best view video synthesis. *IEEE Transactions on Multimedia*, 10(6):997–1012.

[Kabeya et al., 2016] Kabeya, Y., Tomiyasu, F., and Mase, K. (2016). Semi-automatic multiple player tracking of soccer games using laser range finders.

In *Proceedings of the 7th ACM International Conference of Augmented Human*, pages 40:1–40:2.

[Kameda et al., 2004] Kameda, Y., Koyama, T., Mukaigawa, Y., Yoshikawa, F., and Ohta, Y. (2004). Free viewpoint browsing of live soccer games. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo*, volume 1, pages 747–750.

[Kanade et al., 1997] Kanade, T., Rander, P., and Narayanan, P. (1997). Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, 4(1):34–47.

[Kazemi et al., 2013] Kazemi, V., Burenius, M., Azizpour, H., and Sullivan, J. (2013). Multi-view body part recognition with random forests. In *Proceedings of the 24th British Machine Vision Conference*, pages 1–11.

[Kilner et al., 2006] Kilner, J., Starck, J., and Hilton, A. (2006). A comparative study of free viewpoint video techniques for sports events. In *Proceedings of the 2006 European Conference on Visual Media Production*, pages 87–96.

[Kim et al., 2006] Kim, H., Kitahara, I., Kogure, K., and Sohn, K. (2006). A real-time 3D modeling system using multiple stereo cameras for free-viewpoint video generation. In *Proceedings of the 2006 International Conference on Image Analysis and Recognition*, pages 237–249.

[Kim and Wolf, 2010] Kim, H. and Wolf, M. (2010). Distributed tracking in a large-scale network of smart cameras. In *Proceedings of the 4th ACM/IEEE International Conference on Distributed Smart Cameras*, pages 8–16.

[Kumano et al., 2002] Kumano, M., Ariki, Y., Amano, M., and Uehara, K. (2002). Video editing support system based on video grammar and content analysis. In *Proceedings of the 16th IAPR International Conference on Pattern Recognition*, volume 2, pages 1031–1036.

[Kurutepe et al., 2007] Kurutepe, E., Aksay, A., Bilen, C., Gurler, C. G., Sikora, T., Akar, G. B., and Tekalp, A. M. (2007). A standards-based, flexible, end-to-end multi-view video streaming architecture. In *Proceedings of the International Packet Video Workshop*, pages 302–307.

[Lai et al., 2010] Lai, Y.-C., Liang, Y.-M., Shih, S.-W., Liao, H.-Y. M., and Lin, C.-C. (2010). Linear production game solution to a PTZ camera network. In

Proceedings of the 17th IEEE International Conference on Image Processing, pages 4317–4320.

[Lawrence Zitnick et al., 2004] Lawrence Zitnick, C., Kang, S. B., Uyttendaele, M., Winder, S., and Szeliski, R. (2004). High-quality video view interpolation using a layered representation. 23(3):600–608.

[Le Callet et al., 2012] Le Callet, P., Möller, S., and Perkis, A. (2012). Qualinet white paper on definitions of quality of experience. *European Network on Quality of Experience in Multimedia Systems and Services*, 3:1–20.

[Li and Bhanu, 2011] Li, Y. and Bhanu, B. (2011). A comparison of techniques for camera selection and hand-off in a video network. In *Distributed Video Sensor Networks*, pages 69–83. Springer.

[Li and Bhanu, 2012] Li, Y. and Bhanu, B. (2012). Camera pan/tilt control with multiple trackers. In *Proceedings of the 21st IAPR International Conference on Pattern Recognition*, pages 2698–2701.

[Li et al., 2005] Li, Z., Schuster, G. M., and Katsaggelos, A. K. (2005). Minmax optimal video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1245–1256.

[Liu et al., 2013] Liu, J., Carr, P., Collins, R. T., and Liu, Y. (2013). Tracking sports players with context-conditioned motion models. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1830–1837.

[Lou et al., 2005] Lou, J.-G., Cai, H., and Li, J. (2005). A real-time interactive multi-view video system. In *Proceedings of the 13th ACM International Conference on Multimedia*, pages 161–170.

[Marutani et al., 2012] Marutani, T., Mase, K., Fujii, T., and Kawamoto, T. (2012). Multi-view video contents viewing system by synchronized multi-view streaming architecture. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1277–1278.

[Mase et al., 2011] Mase, K., Niwa, K., and Marutani, T. (2011). Socially assisted multi-view video viewer. In *Proceedings of the 13th ACM International Conference on Multimodal Interfaces*, pages 319–322.

[Mase et al., 2006] Mase, K., Sumi, Y., Toriyama, T., Tsuchikawa, M., Ito, S., Iwasawa, S., Kogure, K., and Hagita, N. (2006). Ubiquitous experience media. *IEEE MultiMedia*, 13(4):20–29.

[Matsuyama et al., 2012] Matsuyama, T., Nobuhara, S., Takai, T., and Tung, T. (2012). *3D video and its applications*. Springer.

[Maugey and Frossard, 2013] Maugey, T. and Frossard, P. (2013). Interactive multiview video system with low complexity 2D look around at decoder. *IEEE Transactions on Multimedia*, 15(5):1070–1082.

[Memmert et al., 2017] Memmert, D., Lemmink, K. A., and Sampaio, J. (2017). Current approaches to tactical performance analyses in soccer using position data. *Sports Medicine*, 47(1):1–10.

[Mooney and Roy, 2000] Mooney, R. J. and Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the 5th ACM Conference on Digital Libraries*, pages 195–204.

[Morioka et al., 2008] Morioka, K., Kovacs, S., Lee, J.-H., Korondi, P., and Hashimoto, H. (2008). Fuzzy-based camera selection for object tracking in a multi-camera system. In *Proceeding of the 2008 IEEE Conference on Human System Interactions*, pages 767–772.

[Muramatsu et al., 2014] Muramatsu, Y., Hirayama, T., and Mase, K. (2014). Video generation method based on user's tendency of viewpoint selection for multiview video contents. In *Proceedings of the 5th ACM International Conference of Augmented Human*, pages 1:1–1:4.

[Nack and Ide, 2011] Nack, F. and Ide, I. (2011). Why did the prime minister resign?: Generation of event explanations from large news repositories. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 313–322.

[Ohm and Müller, 1999] Ohm, J.-R. and Müller, K. (1999). Incomplete 3-D representation of video objects for multiview applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(2):389–400.

[Onishi and Fukunaga, 2004] Onishi, M. and Fukunaga, K. (2004). Shooting the lecture scene using computer-controlled cameras based on situation understanding and evaluation of video images. In *Proceedings of the 17th IAPR International Conference on Pattern Recognition*, pages 781–784.

[Orriols and Binefa, 2001] Orriols, X. and Binefa, X. (2001). An EM algorithm for video summarization, generative model approach. In *Proceedings of the 8th IEEE International Conference on Computer Vision*, volume 2, pages 335–342.

[Ozeki et al., 2001] Ozeki, M., Nakamura, Y., and Ohta, Y. (2001). Camerawork for intelligent video production-capturing desktop manipulations. In *Proceedings of the 2001 IEEE International Conference on Multimedia and Expo*, pages 40–43.

[Pan et al., 2011] Pan, Z., Ikuta, Y., Bandai, M., and Watanabe, T. (2011). User dependent scheme for multi-view video transmission. In *Proceedings of the 2011 IEEE International Conference on Communications*, pages 1–5.

[Peng et al., 2016] Peng, X., Wang, L., Wang, X., and Qiao, Y. (2016). Bag of visual words and fusion methods for action recognition. *Computer Vision and Image Understanding*, 150(C):109–125.

[Perše et al., 2009] Perše, M., Kristan, M., Kovačič, S., Vučkovič, G., and Perš, J. (2009). A trajectory-based analysis of coordinated team activity in a basketball game. *Computer Vision and Image Understanding*, 113(5):612–621.

[Peterson, 2011] Peterson, B. (2011). *Learning to See Creatively*. Random House LLC.

[Quijano-Sanchez et al., 2010] Quijano-Sanchez, L., Recio-Garcia, J. A., and Diaz-Agudo, B. (2010). Personality and social trust in group recommendations. In *Proceedings of 22nd IEEE International Conference on Tools with Artificial Intelligence*, volume 2, pages 121–126.

[Ranjan et al., 2010] Ranjan, A., Henrikson, R., Birnholtz, J., Balakrishnan, R., and Lee, D. (2010). Automatic camera control using unobtrusive vision and audio tracking. In *Proceedings of the 36th Graphics Interface Conference*, pages 47–54.

[Ren et al., 2010] Ren, J., Xu, M., Orwell, J., and Jones, G. A. (2010). Multi-camera video surveillance for real-time analysis and reconstruction of soccer games. *Machine Vision and Applications*, 21(6):855–863.

[Resnick and Varian, 1997] Resnick, P. and Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3):56–58.

[Rui et al., 2000] Rui, Y., Gupta, A., and Acero, A. (2000). Automatically extracting highlights for TV baseball programs. In *Proceedings of the 8th ACM International Conference on Multimedia*, pages 105–115.

[Saini et al., 2012] Saini, M. K., Gadde, R., Yan, S., and Ooi, W. T. (2012). Movimash: Online mobile video mashup. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 139–148.

[Schein et al., 2002] Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–260.

[Schlipsing et al., 2017] Schlipsing, M., Salmen, J., Tschentscher, M., and Igel, C. (2017). Adaptive pattern recognition in real-time video-based soccer analysis. *Journal of Real-Time Image Processing*, 13(2):345–361.

[Shao et al., 2006] Shao, X., Xu, C., Maddage, N. C., Tian, Q., Kankanhalli, M. S., and Jin, J. S. (2006). Automatic summarization of music videos. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(2):127–148.

[Shen et al., 2007] Shen, C., Zhang, C., and Fels, S. (2007). A multi-camera surveillance system that estimates quality-of-view measurement. In *Proceedings of the 2007 IEEE International Conference on Image Processing*, pages 193–196.

[Sigari et al., 2015] Sigari, M.-H., Soltanianzadeh, H., and Pourreza, H. R. (2015). Fast highlight detection and scoring for broadcast soccer video summarization using on-demand feature extraction and fuzzy inference. *International Journal of Computer Graphics*, 6(1).

[Smith and Kanade, 1998] Smith, M. A. and Kanade, T. (1998). Video skimming and characterization through the combination of image and language understanding. In *Proceedings of the 1998 IEEE International Workshop on Content-Based Access of Image and Video Database*, pages 61–70.

[Snoek et al., 2005] Snoek, C. G., Worring, M., and Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th ACM International Conference on Multimedia*, pages 399–402.

[Sun et al., 2009] Sun, J., Wu, X., Yan, S., Cheong, L.-F., Chua, T.-S., and Li, J. (2009). Hierarchical spatio-temporal context modeling for action recognition. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2004–2011.

[Sundaram et al., 2002] Sundaram, H., Xie, L., and Chang, S.-F. (2002). A utility framework for the automatic generation of audio-visual skims. In *Proceedings of the 10th ACM International Conference on Multimedia*, pages 189–198.

[Tkalcic and Chen, 2015] Tkalcic, M. and Chen, L. (2015). Personality and recommender systems. In *Recommender Systems Handbook*, pages 715–739. Springer.

[Tokai et al., 2008] Tokai, S., Kawamoto, T., Fujii, T., and Mase, K. (2008). Pegged to point browsing: An approach to browse multi-view video with view-switching, and its applications. In *Proceedings of the 2008 ICPR Workshop on Sensing Web*, pages 41–46.

[Tomiyasu and Mase, 2015] Tomiyasu, F. and Mase, K. (2015). Human-machine cooperative viewing system for wide-angle multi-view videos. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, pages 85–88.

[Truong and Venkatesh, 2007] Truong, B. T. and Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(1):3.

[Wang, 2013] Wang, X. (2013). Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34(1):3–19.

[Wang et al., 2017a] Wang, X., Enokibori, Y., Hirayama, T., Hara, K., and Mase, K. (2017a). User group based viewpoint recommendation using user attributes for multiview videos. In *Proceedings of the 2017 ACM Multimedia Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes*, pages 3–9.

[Wang et al., 2016] Wang, X., Hara, K., Enokibori, Y., Hirayama, T., and Mase, K. (2016). Personal multi-view viewpoint recommendation based on trajectory distribution of the viewing target. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 471–475.

[Wang et al., 2015] Wang, X., Hirayama, T., and Mase, K. (2015). Viewpoint sequence recommendation based on contextual information for multiview video. *IEEE MultiMedia*, 22(4):40–50.

[Wang et al., 2014a] Wang, X., Muramatu, Y., Hirayama, T., and Mase, K. (2014a). Context-dependent viewpoint sequence recommendation system for multi-view video. In *Proceedings of the 2014 IEEE International Symposium on Multimedia*, pages 195–202.

[Wang et al., 2013] Wang, Y., Natarajan, P., and Kankanhalli, M. (2013). Multi-camera Skype: Enhancing the quality of experience of video conferencing. In *The Era of Interactive Media*, pages 243–253. Springer.

[Wang et al., 2012] Wang, Y.-S., Lin, H.-H., and Liao, Y.-W. (2012). Investigating the individual difference antecedents of perceived enjoyment in students' use of blogging. *British Journal of Educational Technology*, 43(1):139–152.

[Wang et al., 2017b] Wang, Z., Yu, J., and He, Y. (2017b). Soccer video event annotation by synchronization of attack–defense clips and match reports with coarse-grained time information. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(5):1104–1117.

[Wang et al., 2014b] Wang, Z., Yu, J., He, Y., and Guan, T. (2014b). Affection arousal based highlight extraction for soccer video. *Multimedia Tools and Applications*, 73(1):519–546.

[Xu et al., 2005] Xu, C., Shao, X., Maddage, N. C., and Kankanhalli, M. S. (2005). Automatic music video summarization based on audio-visual-text analysis and alignment. In *Proceedings of the 28th ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 361–368.

[Yang et al., 2007] Yang, B., Mei, T., Hua, X.-S., Yang, L., Yang, S.-Q., and Li, M. (2007). Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, pages 73–80.

[Yao et al., 2017] Yao, T., Wang, Z., Xie, Z., Gao, J., and Feng, D. D. (2017). Learning universal multiview dictionary for human action recognition. *Pattern Recognition*, 64:236–244.

[Yi et al., 2015] Yi, S., Li, H., and Wang, X. (2015). Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3488–3496.

[Zhang et al., 2008] Zhang, C., Rui, Y., Crawford, J., and He, L.-W. (2008). An automated end-to-end lecture capture and broadcasting system. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 4(1):6:1–6:23.

[Zhang et al., 2017] Zhang, W., Liu, Z., Zhou, L., Leung, H., and Chan, A. B. (2017). Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3D human pose estimation. *Image and Vision Computing*, 61:22–39.

[Zhu et al., 2016] Zhu, Y., Hanjalic, A., and Redi, J. A. (2016). QoE prediction for enriched assessment of individual video viewing experience. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 801–810.

List of Publications

Journal Papers

1. Xueting Wang, Kensho Hara, Yu Enokibori, Takatsugu Hirayama, and Kenji Mase, “Personal Viewpoint Navigation based on Object Trajectory Distribution for Multi-view Videos”, *IEICE Transactions on Information and Systems*, vol. E101-D, no. 1, pp. 193–204, Jan. 2018.

2. Xueting Wang, Takatsugu Hirayama, and Kenji Mase, “Viewpoint Sequence Recommendation Based on Contextual Information for Multiview Video”, *IEEE Multimedia*, vol. 22, no. 4, pp. 40–50, Oct. - Dec. 2015.

3. Fumiharu Tomiyasu, Xueting Wang, and Kenji Mase, “Video Cut Extraction Method for Wide-angle Multi-view Videos Using Spatial Relationship between Ball and Cameras”, *The Journal of the Institute of Image Electronics Engineers of Japan*, vol. 115, no. 494, pp. 175–180, Dec. 2016. (In Japanese)

Conference Papers (Peer Reviewed)

1. Xueting Wang, Yu Enokibori, Takatsugu Hirayama, Kensho Hara, and Kenji Mase, “User Group based Viewpoint Recommendation using User Attributes for Multi-view Videos”, *In Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes (ACMMM MUSA2 2017)*, pp. 3–9, Oct. 2017.

2. Xueting Wang, Kensho Hara, Yu Enokibori, Takatsugu Hirayama, and Kenji Mase, “Personal Multi-view Viewpoint Recommendation based on Trajectories Distribution of the Viewing Target”, *In Proceedings of the 24th ACM Multimedia Conference (ACMMM 2016)*, pp. 471–475, Oct. 2016.

3. Xueting Wang, Yuki Muramatu, Takatsugu Hirayama, and and Kenji Mase, "Context-dependent Viewpoint Sequence Recommendation System for Multi-view Video", *In Proceedings of the 10th IEEE International Symposium on Multimedia (ISM 2014)*, pp. 195–202, Dec. 2014.

4. Xueting Wang, "Viewing Support System for Multi-view Videos", *In Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*, Doctoral Consortium, pp. 527–531, Nov. 2016.

5. Fumiharu Tomiyasu, Yuki Muramatsu, Ryotaro Iida, Xueting Wang, Tomoko Yonezawa, Takatsugu Hirayama, and Kenji Mase, "Multi-View Video Viewing System by Viewpoint Recommendation for Target Tracking Viewing", *In Proceedings of 2014 Interaction of Information Processing Society of Japan (IPSJ)*, pp. 290–295, Mar. 2014. (in Japanese)

Conference Papers (Non-Peer Reviewed)

1. Xueting Wang, Yu Enokibori, Takatsugu Hirayama, Kensho Hara, and Kenji Mase, "Analysis of Relationship between Big Five Model based User Attributes and Camera Selecting Tendency for Multi-camera Videos", *In Technical Report of IEICE MVE* 2017 - 30, Oct. 2017. (in Japanese)

2. Xueting Wang, Kensho Hara, Yu Enokibori, Takatsugu Hirayama, and Kenji Mase, "Viewing Model based User Grouping for Viewing Support of Mutli-view Videos", *In Proceedings of 2016 IEICE HCG Symposium*, B-4-2, Dec. 2016. (in Japanese)

3. Xueting Wang, Kensho Hara, Yu Enokibori, Takatsugu Hirayama, and Kenji Mase, "User Grouping based on Viewpoint Selecting Pattern of Multi-view Videos", *In Proceedings of the 19th Meeting on Image Recognition and Understanding (MIRU* 2016), PS1-35, Aug. 2016. (in Japanese)

4. Xueting Wang, Yu Enokibori, Takatsugu Hirayama, Kensho Hara, and Kenji Mase, "Viewpoint Recommendation based on Trajectory Distribution of Viewing Target for Soccer Multi-view Video", *In Technical Report of IEICE MVE* 2015 - 96, Mar. 2016. (in Japanese)

5. Xueting Wang, Yuki Muramatsu, Takatsugu Hirayama, and Kenji Mase, "Context-aware Multi-view Viewpoint Selection", *In Proceedings of the 18th*

Meeting on Image Recognition and Understanding (MIRU 2015), SS2–22, Jul. 2015. (in Japanese)

6. Xueting Wang, Yuki Muramatsu, Takatsugu Hirayama, and Kenji Mase, "Viewpoint Recommendation based on Trajectory Distribution of Viewing Target for Soccer Multi-view Video", *In Technical Report of IEICE MVE* 2014 - 10, Jun. 2014. (in Japanese)

Award

1. Best Paper Award,
   Xueting Wang, Yuki Muramatsu, Takatsugu Hirayama and Kenji Mase, "Context-dependent Viewpoint Sequence Recommendation System for Multi-view Video", *In Proceedings of the 10th IEEE International Symposium on Multimedia (ISM 2014)*, pp. 195–202, Dec. 2014.