Context-dependent Viewpoint Sequence Recommendation System for Multi-view Video

Xueting Wang*, Yuki Muramatu†, Takatsugu Hirayama‡ and Kenji Mase§
Graduate School of Information Science
Nagoya University
Nagoya, Aichi, Japan
* *Email: wangxueting12@gmail.com*
† *Email: yuki.m830@gmail.com*
‡ *Email: hirayama@is.nagoya-u.ac.jp*
§ *Email: mase@nagoya-u.jp*

***Abstract*—Multi-view videos shot using multiple cameras are highly interested due to their considerable flexibility in enhancing the quality of our daily viewing experience, especially for large-scale events. However, the increase in the number of cameras burdens even experts on suitable viewpoint selection. Therefore, we propose in this paper an automatic viewpoint sequence recommendation system to support multi-view viewpoint selecting with a soccer game example. Unlike existing methods, our proposed system focuses on context-dependency using viewpoint evaluation and transition processes by two types of agents: a camera agent and a producer agent. The camera agent evaluates the view quality based on scene context such as positions of ball and players in given production context such as camera position and user's preference. The producer agent selects the optimal set of viewpoints by taking account of the view quality and the production objectives. The context-dependent optimization has been performed to generate variable viewing patterns which are adequate to various scene and production contexts. Sequences generated by the system and the human selection were experimentally compared to confirm the effectiveness of our proposed system. Our recommendation system has the potential to satisfy both common and personal viewing preferences for sports games.**

***Keywords*-multi-view video; context-dependent; recommendation system; agent;**

I. Introduction

As an extension of the widely used applications of the single-view video, such as regular TV broadcasting, multi-view video can provide better viewing experience through diverse event representation from different view angles. Thus, many systems for multi-view capturing, processing, and delivering have been developed to transmit information from various viewpoints for sports video broadcasts, education webcast systems, concerts, and other events [1][2]. Further, free-view video can be produced to provide more viewpoint options by interpolating scenes or modeling 3D scenes according with computer vision and image understanding techniques [3].

Several viewing interfaces have been developed to enable users to freely choose their preferred viewpoints [4][5]. These developments enhance users' viewing experience as they are more flexible and interactive than single-view video applications. However, the continual selection of the appropriate viewpoint among several viewpoints will be a complex task for an average viewer lacking professional awareness and experience. Furthermore, such applications would heavily burden even experts having the knowledge, awareness of preferences, and the intense concentration required to select among them. Therefore, an automatic recommendation of viewpoint sequence considering their preference that can guide and support users' viewpoint selection would be more convenient and enjoyable.

In this paper, we propose a context-dependent automatic viewpoint sequence recommendation system to support multi-view video viewing. To generate an optimal viewpoint sequence, we should first consider the contextual information that includes the scene information such as the ball and player positions, players actions and plays, and game events. Another contextual information includes the video production information such as the camera setups, their assigned roles, user's preference of viewing style, user's favorite players and events, the objective of watching the game, video production plan and rules and so on. They are named in this paper the scene context and the production context, respectively, and some of them are used for the viewpoint sequence recommendation system in its parameter optimization and recommendation. For instance, a camera view is defined by a scene context such as players positions and by a production context such as the camera position and the main target object. Moreover, a viewpoint sequence is generated by the evaluated quality of each camera view and by the production context such as a production rule on smooth viewpoint transition. In more detailed example of a scene context, the viewpoint selection tendencies of viewers depend on the location of the game event, e.g. the ball. If it is a scene that the ball locates in the penalty area, most viewers might choose viewpoints that show the ball from close-by and switch viewpoints to view it from various directions. In

978-0-7695-5437-2/14 $31.00 © 2014 IEEE
DOI 10.1109/ISM.2014.44

another case, they may prefer a wide and stable shot if the ball is in the midfield.

Therefore, it is important to achieve context-dependent viewpoint recommendation corresponding to various viewing patterns of users. We use viewpoint evaluation and transition processes to compute the contextual information of scene and production. First, in the evaluation process, a camera agent is designed to evaluate the view quality of each viewpoint by considering the scene contextual information represented by appearance features of objects in given production context such as user's preference in objects. A producer agent determines the most appropriate sequence of viewpoint transition with minimum transition cost by taking account of the view quality and production contextual information such as multi-video editing rules. In these two processes, our proposed system combines the appearance features and calculates the transition cost by using different weight parameters. The weight parameters are optimized to represent different viewing patterns corresponding to scene contexts and production contexts including users' preferences and camera position. Optimizations are performed based on the common viewpoint sequences selected by users according to their majority preference. Then it is performed based on personal preference to investigate the possibility of personalized recommendation.

This paper is organized as follows. In Section II, we introduce several related work. In Section III, we present the framework of our proposed approach for viewpoint sequence recommendation using scene and production contexts in Section IV. The detail experiments for viewing pattern modeling are in Section V and system evaluation using a real multi-camera dataset are in Section VI. We offer our conclusions in Section VII.

II. Related Work

In this paper, we discuss the limits of existing automatic viewpoint sequence editing methods.

First, in the representation of scene contextual information for each viewpoint quality evaluation, Chen et al. [6] proposed a method for automatic viewpoints suggestion by analyzing features of group of objects such as the visible number of objects. In [7] Daniyal et al. presented a novel evaluation algorithm based on individual information of the object, such as the position of each object. The results are sometimes limited because the importance of individual and group contextual information are lacking in objects analysis for scene features. Hence, for example, the camera view always focuses on the center field even when the ball go near the goal but far away from most of the players.

Shen et al. proposed a best-view selection method using a detailed content analysis based on Quality of View (QoV) [8]. In [9] Jiang et al. optimized multiple object tracking and formulated best-view video synthesis as a recursive decision problem. In these researches, viewpoints

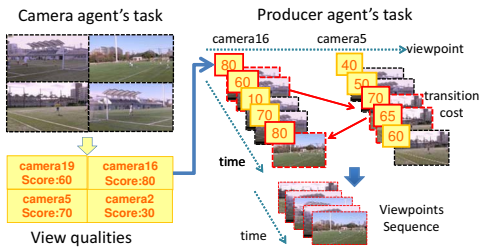

Figure 1. Outline of viewpoint sequence recommendation system for multi-camera videos through viewpoint evaluation and transition processes.

are frequently switched without considering the production contextual information on viewpoint transition, such as the relative positions of cameras. Such system would produce short timing and uncomfortable camera changes.

An extensive work of Daniyal [10] presented an algorithm to minimize the number of inter-camera switches but used the same criteria for content analysis and viewpoint transition arrangement of all durations regardless of users' different view preferences to scene contexts of different durations. Hence, such system may select a viewpoint with near detailed view while users prefer an overview for a midfield scene.

Therefore, the viewpoint evaluation capability and transition mechanism in these studies mentioned above are not dynamic enough against various scene contexts and user preferences, which is quite important for user-dependent viewpoint sequence recommendation.

III. Viewpoint Sequence Recommendation System

In this section, we introduce the framework for our proposed video sequence recommendation system.

The recommendation of viewpoints can be regarded as two kinds of tasks: the quality evaluation of the image at each viewpoint and the choice of viewpoint among many to generate a video sequence. Therefore, to simplify our system architecture, we divide it into two different processing parts: the evaluation process and the transition process.

Users are presented with the generated viewpoint sequence following the two processes at each frame. The overall flow of the system is shown in Figure 1.

A. Viewpoint Evaluation Process

We set a camera agent to each viewpoint. The task of the camera agent in the evaluation process is to assess the view by quality considering scene contextual information of each viewpoint in given production context such as user's preference in objects. According to the QoV measure [8], view quality is strongly related to, and can be evaluated based on, the appearance of view objects in the image. In this paper, we use numerical values for four appearance measures: in-frame measure f_e, proximity measure f_d, composition measure f_c, and context-related location

measure f_l. Furthermore, we use these measures to evaluate every object in the video frame. Viewpoint evaluation is conducted by a combination of all objects evaluations by considering the crowd and importance of objects. Following these evaluations, weight parameters are defined to represent the importance of each measure. Moreover, a combination pattern of these weight parameters is decided according to the scene contexts and optimized by production contexts including users' preferences and camera position. The evaluation steps are as follows:

1) Evaluation of single object in the frame:
   The camera agent first evaluates the characteristics of an object using the following appearance measures:
   - In-frame measure f_e: the visibility of the object in the field of view (FOV), as follows,

$$f_e = \begin{cases} 1 & \text{(object in view)}, \\ 0 & \text{(otherwise)}. \end{cases} \quad (1)$$

     The camera agent determines whether object o exists in the FOV of a camera. The position of the object, which is transformed from its position in the world coordinate system by a coordinate transformation, in an image of the view.
   - Proximity measure f_d: the distance between the object and the camera.
     To a certain extent, the clarity with which a user can view the object in an image is dependent upon its apparent size. The apparent size of the object is inversely proportional to the distance with the camera v_i. It can be normalized by a coefficient D_f, which represents the length of the diagonal of an event space such as a soccer filed. Therefore, we define the distance measure as

$$f_d = \begin{cases} 1 - \frac{D(o,v_i)}{D_f} & (D(o,v_i) < D_f), \\ 0 & \text{(otherwise)}. \end{cases} \quad (2)$$

   - Composition measure f_c: the composition of the object in the view.
     The visual attention to an object depends on its position in an image. Hence, we implement "the rule of thirds" [11] that is widely-used in visual arts, such as painting and photography, to evaluate the composition of the view images. It is well known that the center of an image attracts the viewer's attention. Besides, according to the rule, important compositional elements should be placed at the intersections of the lines dividing the image into nine (= 3×3) equal parts. Therefore, in this study, we evaluate the composition of view images by calculating the average distance between the position of the object, the center of the image p_c, and the intersections (p_l, p_r), which is normalized by the length of the diagonal line D_{img} of the image, as follows,

$$f_c = 1 - \frac{D(o,p_c) + D(o,p_l) + D(o,p_r)}{3D_{img}}. \quad (3)$$

   - Scene context-related location measure f_l: the relationship between the location of the object and the scene context.
     For instance, a player ($= object$) in the penalty area attracts more attention than one in other areas in a scene where the ball is also in the penalty area. We divide the event space into several non-overlapping areas (will be explained in section IV) and assign each area a score $f_l \in [0,1]$ that varies according to the scene context, e.g. ball positions. The score represents the relevant importance between the physical position of the object and the scene context. A detailed example of the measure and scene context is provided in Section IV.

   Therefore, a single object o can be evaluated by combining these measures with weights $\{\omega_d, \omega_c, \omega_l\}$ as follows,

$$V_s(o) = f_e * (f_d * \omega_d + f_c * \omega_c + f_l * \omega_l), \quad (4)$$

   where $\omega_d + \omega_c + \omega_l = 1$.
2) Evaluation of multiple objects: The camera agent evaluates multiple objects by considering both the average evaluation and the count number of them using weight ω_n as follows, where $|O|$ is the size of the set of objects o_k,

$$V_n(O) = |O| * \omega_n + \frac{1}{|O|} \sum_{o_k \in O} V_s(o_k) * (1 - \omega_n). \quad (5)$$

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

$$V(p) = V_n(\{O_m\}) * \omega_m + V_n(\{O_s\}) * (1 - \omega_m). \quad (6)$$

   The main objects are chosen either by the system or the user.

Through the above steps, the camera agents determine the overall evaluations of the viewpoints and submit them to the producer agent for further processing.

B. Viewpoint Transition Process

The producer agent determines the optimal sequence of viewpoints in this process. We solve this problem by calculating transition cost based on the view quality evaluated by camera agents and production contextual information: duration of viewpoint and load of camera transitions with

different weights. The weights vary with different viewing patterns according to scene contexts and production contexts including users' preferences and camera position. The viewpoint transition will be restricted if the cost is too high.

- Duration cost of previous viewpoint $D(p(t-1))$: When watching videos, people feel uncomfortable if the duration of a viewpoint is too short or if viewpoints are switched too frequently. In order to avoid this, the agent calculates the contiguous number of frames $T(p(t-1))$ until the viewpoint $p(t-1)$ of the previous frame $t-1$. Hence, the duration cost of the previous viewpoint $D(p(t-1))$ is defined as follows, where T_{ex} is the set for normalization.

$$D(p(t-1)) = 1 - \frac{T(p(t-1))}{T_{ex}}. \quad (7)$$

- Temporal view quality subtraction cost $Q(p_c(t))$: The change cost of view quality is defined as the difference in evaluation scores between the previous viewpoint $p(t-1)$ and the candidate viewpoint $p_c(t)$ of the current frame $V(p_c(t))$ as follows,

$$Q(p_c(t)) = -(V(p_c(t)) - V(p(t-1))). \quad (8)$$

  In contrast to the duration cost, the higher the evaluation score of a candidate viewpoint at the frame than the current score of the preceding viewpoint $V(p_c(t-1))$, the lower the cost of switching to candidate $p_c(t)$.
- Visual angles difference cost $A(p_c(t))$: To avoid the dizziness caused by too great a change in the visual angle and the position of the camera, we take two factors into account, the change in the view angles of the cameras, f_a, and the change in the position, f_n, between the previous and the candidate viewpoint as follows,

$$A(p_c(t)) = f_a(p_c(t), p(t-1)) + f_n(p_c(t), p(t-1)). \quad (9)$$

  The greater the difference in the view angles and the positions, the higher the cost of switching to $p_c(t)$.

As a consequence, the cost function of the current viewpoint is determined by combining these three factors as follows,

$$\begin{aligned} C(p_c(t)) &= T(p(t-1)) * \omega_t + Q(p_c(t)) * \omega_s \\ &\quad + A(p_c(t)) * \omega_a, \end{aligned} \quad (10)$$

where $\omega_t + \omega_s + \omega_a = 1$.

To find the best viewpoint for a transition, the producer agent calculates the cost functions of each candidate viewpoint and chooses the one when the cost is minimum and less than a pre-defined threshold as defined below.

$$p(t) = \begin{cases} p_c(t) & \min_i\{C(p_c(t))\} < \text{threshold}, \\ p(t-1) & \text{(otherwise)}. \end{cases}$$

The producer agent repeats this process at each frame and connects the viewpoints to form the recommended sequence.

IV. Viewing Patterns by View-dependent Contexts

We generate the viewpoint recommendation sequence with various viewing patterns by considering scene and production contexts.

A. Scene Context

In general, users' viewing patterns differ each other based on their preference to various scene contexts, which affect their viewpoint evaluation and transition. Therefore, we assume that there are differently weighted patterns for the measures and costs defined in the previous section. There is a rich description of scene contexts that is dependent on complex factors of the view objects and events. In this study, we focus on the coverage of soccer games.

According to a user survey from the experiment detailed in Section V, we found that the position of the ball is of the greatest preference to viewers of a soccer game. Hence, we regard the ball as an important scene context descriptor and quantify its influence based on the position of the ball to model the weight pattern for viewpoint evaluation and transition processes. We divided the soccer field into 15 regions, and assumed that the regions are grouped into six areas, namely the penalty(1), corner(2), transition(3), midfield(4), side-midfield(5), and then side (6), representing six characteristic contexts as shown in Figure 2(a). Thus, we can represent a video sequence as a series of scene contexts in terms of ball position and its influential regions. Moreover, we can achieve various patterns of viewpoint evaluation and transition by assigning corresponding weight parameters $\{\omega_d, \omega_c, \omega_l, \omega_n, \omega_m\}$ for viewpoint evaluation and weight parameters $\{\omega_t, \omega_s, \omega_a\}$ for viewpoint transition. For particular, a certain player of a viewer's favorite can be set as the main object of the scene by being assigned a high weight value ω_m. The preference is a production context. As a consequence, the system will generate a recommendation of viewpoint sequence that features the favorite player more.

The location measure f_l, described in Section III, is also given to each region according to the scene context descriptor, i.e., the current position of the ball as in Figure 2(b). The example in Figure 2(b) shows the score pattern of the location measure corresponding to the scene context of the penalty area where the ball is. The f_l of a player existing in the transition area of the opposite field is assigned low to represent his/her low relevance to the current penalty scene context.

The scene context is arbitrarily defined depending on the event space of the video such as the field of a soccer game and the position of objects of greater preference. In some case, the weights of context-dependent values are adjusted in parameter optimization. Therefore, besides soccer, our method can also be applied to many ball games, and other sports or events that require a fairly large space where the

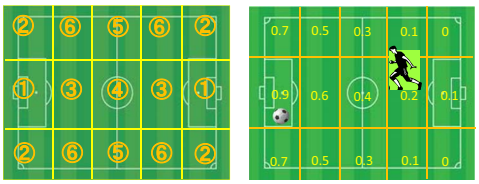

Figure 2. (a)Scene context area segmentation, (b)Sample of the score distribution of the location measure f_l for the penalty area scene context.

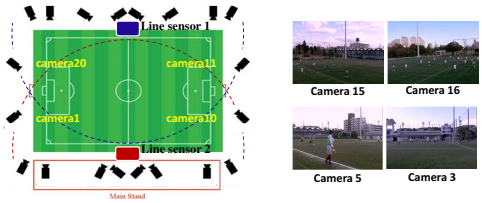

Figure 3. Cameras setups and samples of multi-camera videos.

scene context of the event is essentially region-dependent such as big shows with singers and dancers.

B. Production Context

In addition to scene contexts, users' viewing patterns differ according to production objectives, which are relevant to users' preferences and camera positions.

1) User Preference for weight parameter optimization: For viewing an event, some users have a common viewing pattern while some of them have personal preferences different from others. The viewing preferences can be represented by the tendency on viewpoint selection. Thus, we can learn the user preferences by analyzing viewpoint selection records of actual users and achieve various viewing patterns by optimizing the weight parameters. The optimization can be conducted in two ways. One is based on a common viewing pattern and the other is based on personal preferences gleaned from the viewing records.

2) Camera Position-dependent Evaluation Score: Since the similarity between the system generated viewpoint sequences and the users' selection is mostly related to the camera positions, we have to consider the camera position context for weight patterns optimization. We employ an evaluation score $S \in [0,1]$, to represent the similarity of the system's performance to humans' dependent on the camera positions. For each frame of a sequence, the similarity score S_t of the frame t is calculated by comparing the viewpoint selections of the system and the users using triangular functions, as follows,

$$S_t = \begin{cases} 1-|s-u|*0.45 & |s-u|<3\text{(same or nearby)} \\ 1-(|21-s-u|+1)*0.45 & |21-s-u|<1\text{(opposite)} \\ 0 & \text{(otherwise)} \end{cases}$$

$$S = \frac{\sum_t S_t}{L}$$

where s is the camera number of the selected viewpoint by the system at frame t, and u is the camera number of the viewpoint selected by the user. L is the length of the sequence. Thus, the evaluation score S of a sequence is the average of the similarity scores S_t along time. If the camera setup changes, the score function formula has to be changed to represent the appropriate similarity between selections of the system and the users.

C. Weight Parameter Optimization

We apply a brute-force attack algorithm in optimization, which searches for the optimal combination of weight parameters within the range $[0,1]$ over possible contexts combinations. User preference, a production context is given indirectly by the actual viewing sequences of each user. From this, we can extract the viewpoint sequences belonging to each scene context and calculate the camera position context-dependent evaluation score S_{cont}. For common preferences optimization, we calculate the evaluation score by comparing the generated viewpoint sequences by system with the common sequences generated from all users, detailed in Section V. The parameters values that lead to the highest evaluation score in a scene context will be the optimal parameters for the scene context. On the other hand, the selected viewpoint sequences of each viewer is compared to find the optimal parameters for personal preference as well as the common sequences.

V. Experiment

We conducted a video editing experiment on a real multi-viewpoint soccer dataset to optimize the weight parameters and verify the effectiveness of the proposed method.

A. Dataset

The multiple viewpoint videos used for the experiment were filmed using 20 digital cameras (CASIO EX-F1, at 30 fp 1920 $\times$ 1080 pixels) without pan, tilt, and zoom around a soccer field during a game. The cameras were synchronized after filming. Figure 3 shows the positions of the cameras, and four snapshots of the corresponding viewpoint videos. Moreover, the positions of the players on the field were obtained by two laser range sensors set on both sides of the field when filming the multi-camera videos, as shown in Figure 3. The position of the ball was obtained through a semi-manual procedure.

B. Procedure

Six male participants in their 20s who sometimes watch soccer games without special expertise in their daily lives took part in this experiment. The participants were instructed to watch the multi-view sequences and choose the most suitable viewpoint at any time. They were allowed to repeatedly replay the sequences while watching various viewpoints.

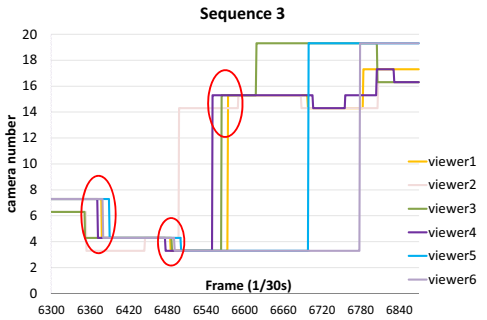

Figure 4. An example of viewpoint sequences by all participants. When most of the participants chose a common viewpoint, the difference on the timing of viewpoint switches was confined to a small range as shown in red circles. Besides, they performed the choice different from most of other participants from 6600 to 6800 frames by their personal preferences.

When they confirmed a viewpoint switch, the time and the viewpoint choice were recorded.

The participants took approximately 10 minutes to edit a sequence lasting 30 seconds. In view of how difficult it is for participants to perform video editing work for long periods of time, we extracted 17 video sequences from the game that containing typical soccer scenes. After the experiment, the participants filled out a questionnaire that asked them to list issues in choosing appropriate viewpoints that they considered important.

The editing record of each participant reflects his/her personal preference, which can be used for personal preference modeling. Further, by the comparison of the records of all participants, as shown in Figure 4, we found that they sometimes chose a common viewpoint as suitable. Although the timing of viewpoint switches differed for each participant, the difference was confined to a small range approximately 1 second. Thus, we concluded that there was a common selection preference among participants and generated the selected viewpoint sequence of common preference by connecting the viewpoint chosen most frequently by all participants for each experimental sequence. The generated common viewpoint sequences are used for common preference modeling.

Furthermore, we evaluate the selections using a two-fold cross-validation test: half the duration of each participant's selected sequence, which includes the common and personal sequences, was used as training data for optimization, while the other half was reserved for test data to evaluate the effect of the optimized parameters.

VI. Results and Analysis

A. Analysis of Optimized Parameters

We first analyze the effect of optimized parameters.

The optimization parameters of each scene context for common preference are listed in Table I. These results show that each view-related context combination yielded

Table I
Optimized weight parameters for common preference

Context scene	$\{\omega_d, \omega_c, \omega_l, \omega_n, \omega_m\}$	$\{\omega_t, \omega_s, \omega_a\}$
Penalty	$\{\underline{0.7}, 0.1, 0.2, 0.2, 0.2\}$	$\{0.3, 0.6, 0.1\}$
Corner	$\{0.2, 0.1, \underline{0.7}, 0.2, 0.1\}$	$\{0.2, 0.1, 0.7\}$
Transition	$\{0.5, 0.3, 0.2, 0.1, 0.2\}$	$\{0.2, 0.3, 0.5\}$
Midfield	$\{0.4, 0.2, 0.4, 0.2, 0.2\}$	$\{0.1, 0.5, 0.4\}$
Side-midfield	$\{0.7, 0.1, 0.2, 0.2, 0.2\}$	$\{0.1, 0.3, 0.6\}$
Side	$\{0.2, 0.2, 0.6, 0.2, 0.2\}$	$\{0.3, 0.5, 0.2\}$

Table II
Comparison for the evaluation score of the sequence using the optimized parameters for all scene contexts and the ones do not respectively.

Trained scene	S
All scene	65%
Penalty	61%
Corner	47%
Transition	62%
Midfield	59%
Side-midfield	60%
Side	62%

corresponding weight parameters. For instance, in the scenes where the ball existed in the penalty area, the value of the proximity measure w_d was larger than other weights whereas the value of the location measure w_l was the largest measure in the corner scene. Thus we can infer that the distance between the objects and the camera was the most important measure for viewpoint evaluation, indicating that many viewers might tend to watch views involving the penalty area more closely. On the other hand, in the corner area scene, viewers might be more interested in the location distribution of the ball and players, since the context-related location measure was the most important.

Furthermore, to quantitate the effect of using different weight parameters for different scene contexts, we compare the evaluation scores of sequences generated by combining the six optimized parameter sets of different contexts with the sequences generated by using just one parameter set, such as the optimized parameters of midfield scene context for all the sequence time long, for common preference. From the results in Table II, we see that sequences using the optimized parameters for all scene contexts achieved a better result than the ones do not. Therefore, it is important to optimize the weight parameters using training data set including various scene contexts.

Moreover, the personal viewing patterns of each participant can be analyzed through the optimized parameters of each participant. For instance, Table III show the optimized parameters for corner scene context for personal preferences of participant 1 and 2. In the weight parameters for transition process of participant 1, the value of weight w_t was larger

Table III
OPTIMIZED WEIGHT PARAMETERS OF CORNER AREA SCENE FOR PARTICIPANT 1'S AND PARTICIPANT 2'S PREFERENCES

Participant	Context scene	$\{\omega_d, \omega_c, \omega_l, \omega_n, \omega_m\}$	$\{\omega_t, \omega_s, \omega_a\}$
1	Corner	$\{0.2, 0.1, 0.7, 0.4, 0.2\}$	$\{\underline{0.7}, 0.1, 0.2\}$
2	Corner	$\{0.1, 0.8, 0.1, 0.8, 0.4\}$	$\{0.1, \underline{0.6}, 0.3\}$

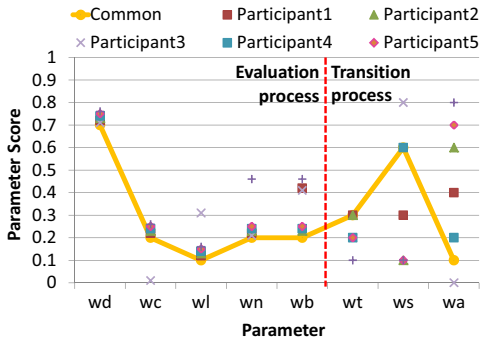

Figure 5. Common and personal parameters of penalty scene.

than other weights, which we consider that participant 1 paid closer attention to the duration cost of previous viewpoint. On the other hand, for participant 2 the value of weight w_s was the largest so that participant 2 might attach greater importance to the temporal view quality subtraction cost in the corner area scene. Furthermore, Figure 5 shows the distribution of optimized parameters of the penalty area scene for both common recommendation and personal ones. From Figure 5 we can find that in the viewpoint evaluation process of the penalty area scene, the participants had a common preference because the optimized parameters of the common recommendation and those for the viewers had a similar distribution. For the viewpoint transition process, the parameters of participants were different.

B. Effectiveness of the System

We then evaluate the effectiveness of the system by separately analyzing the performance of the camera and producer agents using the optimized parameters.

We verify the evaluation results of the camera agent by calculating a coverage rate. First, we rank the viewpoints according to their quality score after evaluation process by camera agent at each frame. The coverage rate is the rate that participants' selections were also the most top-ranked viewpoints of the camera agent for the sequence time long. This rate approached approximately 85% for the common preference, as shown in Figure 6, if we included the two top-ranked viewpoints. The high coverage of participants' selections shows that the evaluation of the camera agents was fairly reliable.

We also verify the producer agent's results using the evaluation score, as shown in Figure 7. The evaluation score for the common preference was approximately 65%,

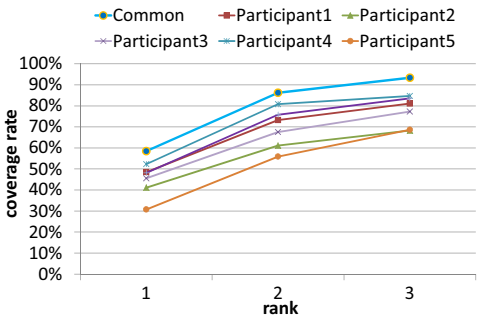

Figure 6. Average coverage rate of participants' selections for camera agents' top-ranked suggestions.

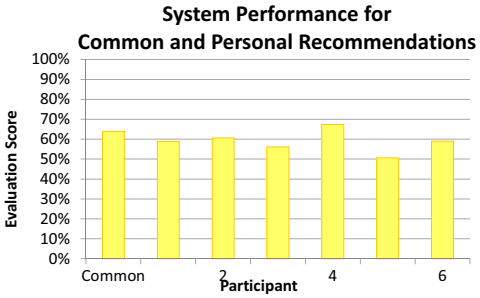

Figure 7. Evaluation score results of producer agent.

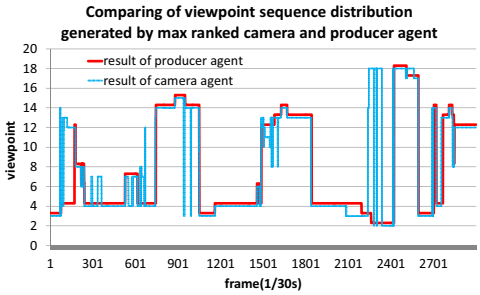

Figure 8. Comparison of viewpoint sequence distribution between the producer agent's result and the top-ranked one generated by camera agents.

whereas the average evaluation score of all participants was approximately 60%. We therefore believe that the choices of the producer agent can reflect both the common and personal demands of participants by comparing with random selections for 20 cameras whose evaluation score would be 5%. Moreover, Figure 8 shows the comparison of viewpoint sequence distribution between the recommendation of the producer agent and the top-ranked viewpoint sequence generated by camera agents. We can confirm that the producer agent is effective on decreasing frequent viewpoint transitions and too short viewpoint durations.

Further, the evaluation score of different context for both common and personal preference are shown in Figure 9. From it we can find that the proposed method performed better in the penalty area and the transition area scenes

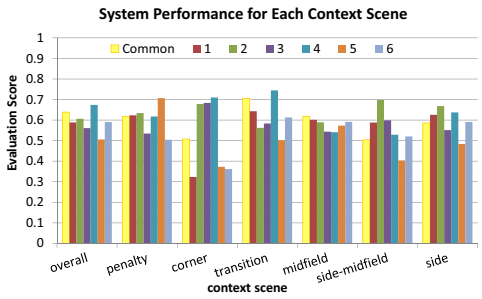

Figure 9. Common and personal recommendation evaluation scores for each context scene.

than in the corner area scene. The significantly varying scores of participants in the corner scene show that the proposed method was not able to adapt to some participants' preference for the corner area scene. We consider two reasons for this problem: one is that the factors for viewing considered in our method was still not enough to incorporate personal preferences of some users and the other is the data of corner area scene is not enough.

VII. Conclusion

In this study, we proposed an automatic video sequence recommendation system to provide convenient and high quality viewpoint selection to enhance viewing experience of multi-view videos. The proposed method carries out evaluation and transition processes using two types of agents to generate various viewing patterns which are adequate to various scene and production contexts. It elects optimal set of viewpoints for both common and personal preferences. The performance of both agents was tested by comparing the viewpoint selections of the proposed system and those made by human viewers in an experiment. Although we used common watchers of soccer game in the experiment, the system has the potential to satisfy professional needs if we learn the optimal parameters from the experts. The results may be a little different since the objects with the main focus of the experts are sometimes different from common watchers in various game contexts. In addition to automatic recommendations, our framework permits users to change their preferences during a recommendation by changing parameters or select the main object they are interested in. Once the parameters are specified, the computation of the viewpoint selection costs only few time. Although in our experiment the position data of the ball was acquired semi-manually, many researches have been carried on automatic object tracking in a soccer game using computer vision techniques [12]. If the position data can be obtained automatically, it will be possible to use the system for live captured events in real-time.

The proposed method can be applied to many ball games as well as events held in a region-dependent space and containing primary objects of preference, such as large shows with singers and dancers. Moreover, user-specific viewing patterns in different scene contexts can be analyzed, or even interactively altered, using our proposed framework.

In future research, we will carry out experiments to examine user satisfaction and develop a system with more interactive manner that allows users to change their preferences and receive greater recommendations.

Acknowledgment

This work was supported by NICT and JSPS KAKENHI Grant Number 26280074.

References

[1] I. Ahmad, "Multi-view video: get ready for next-generation television," *Distributed Systems Online, IEEE*, vol. 8, no. 3, pp. 6–6, 2007.

[2] R. T. Collins, O. Amidi, and T. Kanade, "An active camera system for acquiring multi-view video," in *in Proc. International Conference on Image Processing*, 2002.

[3] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz, "Multi-view stereo for community photo collections," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[4] J.-G. Lou, H. Cai, and J. Li, "A real-time interactive multi-view video system," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 161–170.

[5] K. Mase, K. Niwa, and T. Marutani, "Socially assisted multi-view video viewer," in *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 2011, pp. 319–322.

[6] C. Chen, O. Wang, S. Heinzle, P. Carr, A. Smolic, and M. Gross, "Computational sports broadcasting: Automated director assistance for live sports," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6.

[7] F. Daniyal, M. Taj, and A. Cavallaro, "Content and task-based view selection from multiple video streams," *Multimedia tools and applications,*, pp. 235–258, 2010.

[8] C. Shen, C. Zhang, and S. Fels, "A multi-camera surveillance system that estimates quality-of-view measurement," in *IEEE International Conference on Image Processing*, 2007, pp. 193–196.

[9] H. Jiang, S. Fels, and J. J. Little, "Optimizing multiple object tracking and best view video synthesis," *IEEE Transactions on Multimedia,*, pp. 997–1012, 2008.

[10] F. Daniyal and A. Cavallaro, "Multi-camera scheduling for video production," in *European Conference on Visual Media Production (CVMP),*, 2011, pp. 11–20.

[11] B. Peterson, *Learning to see creatively*. Random House LLC, 2011.

[12] T. D' Orazio and M. Leo, "A review of vision-based systems for soccer video analysis," *Pattern recognition*, vol. 43, no. 8, pp. 2911–2926, 2010.