PAPER

Personal Viewpoint Navigation Based on Object Trajectory Distribution for Multi-View Videos*

Xueting WANG[†a)], ***Student Member***, **Kensho HARA**[†b)], ***Member***, **Yu ENOKIBORI**[††c)], ***Nonmember***, **Takatsugu HIRAYAMA**[††d)], ***Member***, ***and*** **Kenji MASE**[††e)], ***Fellow***

SUMMARY Multi-camera videos with abundant information and high flexibility are useful in a wide range of applications, such as surveillance systems, web lectures, news broadcasting, concerts and sports viewing. Viewers can enjoy an enhanced viewing experience by choosing their own viewpoint through viewing interfaces. However, some viewers may feel annoyed by the need for continual manual viewpoint selection, especially when the number of selectable viewpoints is relatively large. In order to solve this issue, we propose an automatic viewpoint navigation method designed especially for sports. This method focuses on a viewer's personal preference for viewpoint selection, instead of common and professional editing rules. We assume that different trajectory distributions of viewing objects cause a difference in the viewpoint selection according to personal preference. We learn the relationship between the viewer's personal viewpoint-selection tendency and the spatio-temporal game context represented by the objects trajectories. We compare three methods based on Gaussian mixture model, SVM with a general histogram and SVM with a bag-of-words to seek the best learning scheme for this relationship. The performance of the proposed methods are evaluated by assessing the degree of similarity between the selected viewpoints and the viewers' edited records.

key words: *multi-view video navigation, user preference, Gaussian mixture model*

1. Introduction

Multi-view videos taken by multiple cameras from different angles play an important role in video services with the development of video capturing, processing and delivering technologies [1]–[4]. Furthermore, free-view videos can be generated to provide more viewpoint options by interpolating scenes or modeling 3D scenes [5]–[9]. With diverse information and viewing options, viewers can enjoy more interesting content by using multi-view video interfaces [10], [11] than broadcasting with a single forced viewpoint. Multi-view videos are suitable for events as diverse as news, web lectures, concerts and sports.

Manuscript received April 8, 2017.
Manuscript revised August 27, 2017.
Manuscript publicized October 12, 2017.

[†]The authors are with Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8603 Japan.
[††]The authors are with Graduate School of Informatics, Nagoya University, Nagoya-shi, 464–8603 Japan.

*This work was supported by JSPS KAKENHI Grant Number 26280074.

a) E-mail: wang@cmc.ss.is.nagoya-u.ac.jp
b) E-mail: kensho.hara@aist.go.jp
c) E-mail: enokibori@is.nagoya-u.ac.jp
d) E-mail: takatsugu.hirayama@nagoya-u.jp
e) E-mail: mase@nagoya-u.jp

DOI: 10.1587/transinf.2017EDP7122

However, the manual selection of appropriate viewpoints can be tiresome and difficult, especially for viewing a dynamic event in a wide-scale field with numerous viewing options without video editing experience. For such situations, an automatic viewpoint navigation based on a viewer's personal preference is ideal. In this study, we focus on the viewpoint navigation for watching sports involving wide-scale field spaces. Some related studies have been conducted on automatic viewpoint/camera selection. Game context related to the objects is effective to select the viewpoint [12]–[15]. The game context is often represented by frame-level object features, for instance, the size of a player visible in the view in a basketball game. These approaches are processed without enough representation of past and future object dynamics. Besides, most of the related studies focus mainly on common preferences [13], [16]–[18] and professional editing rules [19], [20]. Only several studies considered personal preference [21], [22]. In this study, we focus on the viewer's personal preference on spatio-temporal object dynamics for viewpoint selection.

We aim to achieve a personal viewpoint navigation as shown in Fig. 1 considering the spatio-temporal game context represented by the trajectories of the main viewing objects, i.e., the ball and players in a ball game. About the trajectory processing, some studies used time series model for player's dynamics representation or action recognition, such as Markov models [23], [24]. According to the observation of trajectory distributions of soccer games discussed in detail in Sect. 3, we find that different viewers show different viewpoint-selection tendencies for different trajectory distributions. We assume that the spatial distribution of object trajectory can include dynamical information or represent the object action. Thus, we focus on the spatial distribution of objects trajectories to represent the spatio-temporal game context in this study for the viewpoint selecting problem.

We apply a machine learning method to learn the re-

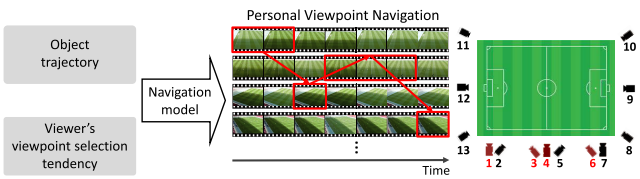

Fig. 1 Outline of personal viewpoint navigation by learning the relationship between the trajectory distribution of the viewing objects and each viewer's viewpoint-selection tendency.

lationship between the trajectory distribution of the viewing objects and each viewer's viewpoint-selection tendency to achieve personal viewpoint navigation. We first divide the video sequences into cuts. We then obtain the object trajectory distribution in cut unit as feature to learn each viewer's viewpoint-selection tendency. We achieve personal navigation by selecting appropriate viewpoints through the proposed learning methods. We evaluate three methods based on Gaussian mixture model (GMM), SVM with a two-dimensional histogram and SVM with a bag-of-words (BoW). We also compare the effectiveness of using combination of targets with just using each target to find viewer's interest in multiple objects.

We now summarize our contribution in this paper as follows. We realize personal viewpoint navigation using the trajectory distribution of the viewing objects and verified the navigation effectiveness. We also show that most viewers focus on the main object in a sport game while some of them have unique interest in specific objects.

This paper is organized as follows. In Sect. 2, we introduce some related studies conducted on multi-view video editing. In Sect. 3, we analyze the relationship between game context and viewer's viewpoint-selection tendency. In Sect. 4, we present our framework of object trajectory based viewpoint navigation. The detailed experiments conducted to acquire viewers' viewpoint selection records are described in Sect. 5 and the result of the evaluation conducted using real multi-camera data sets are described in Sect. 6. We offer our conclusions in Sect. 7.

2. Related Work

In this section, we introduce some existing automatic viewpoint selecting methods.

2.1 Multiple Information Based Viewpoint Navigation

First, we went through some previous researches that selected viewpoints using multiple audio-visual information. In studies [16]–[18], the authors selected viewpoints based primarily on audio features, face trajectory and speaker position for web lectures and meeting broadcasts. The audio information is difficult for viewpoint selection due to smaller differences among cameras and noise of the crowd for some sports with wide scale field. Saini et al. [20] 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 and shakiness.

For sports, we assume that the game context consisting of individual object information, such as positions of the ball and players in a soccer game, is more reliable and effective.

2.2 Game Context Based Viewpoint Navigation

Several other researchers have focused on the game context. Chen et al. [14] focused on features of a group of objects such as the number of players who are visible from a viewpoint. 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 [13], [25]. 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 viewpoint evaluation by considering the view angle and distance from objects for each frame [12]. The extended approach [15] optimized the viewpoint transition by viewpoint-quality evaluation with dynamic features corresponding to the game context represented by the object position. Muramatsu et al. [21] selected viewpoint by using 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.

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

2.3 User Preference Based Viewpoint Navigation

Most of the related studies focused mainly on common preferences, like assigning a viewpoint evaluation score in proportion to the size and the number of players visible in the view [13], [16]–[18], [25]. Saini et al. [20] applied professional editing rules such as the shot length between the directors' viewpoint transitions.

However, general viewers might prefer a view that corresponds to their own preferences. There are only a few existing studies on user-dependent viewpoint selection [15], [21]. They optimized the weight parameters for features extracted from each user's viewpoint-selection record. The performance of these approaches is limited due to the lack of representation of the objects' temporal information.

Therefore, we aim to realize viewpoint navigation by using spatio-temporal game context and learning from personal viewpoint-selection record to realize personal navigation. Our previous study [22] proposed a personal viewpoint recommendation method by modeling the relationship between a viewer's personal viewpoint-selection tendency and the ball trajectory distribution of a soccer game. In this paper, we include trajectory distributions of more objects as the features to model personal viewpoint-selection tendency to adapt to possible interests of different viewers.

3. Qualitative Analysis of Trajectory Distribution and Viewpoint-Selection Tendency

In this study, we collected the viewers' viewpoint-selection records to analyze their viewpoint-selection tendencies through a video editing experiment for sports game, in particular, the soccer games. We provided multi-view videos of

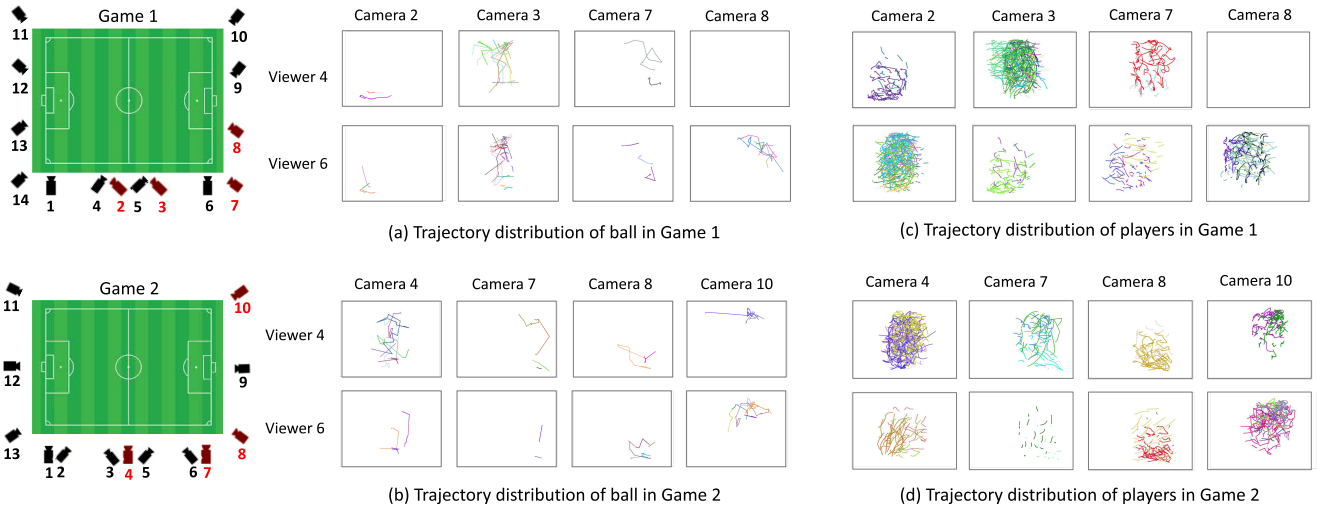

Fig. 2 Ball and players trajectory distribution for each selected viewpoint (extracted) of two viewers in the two games. Each of different colored lines shows the trajectory during a cut unit in overall soccer field. Camera settings of the two games are shown on the left. The cameras in red are the extracted ones for showing the trajectory distribution on the right.

two soccer games to viewers and let them edit a summary video of a soccer game by continually selecting their preferable viewpoint from various viewpoints around the game field shown in Fig. 2. We will introduce the detail of the experiment setting in Sect. 5.

We assume that the viewers select the appropriate viewpoints based on the game context, which can be represented by the spatio-temporal movement of focusing objects, i.e., the ball and players in the case of a soccer game. Thus, we analyze the relationship between the trajectory distribution of viewing objects and each viewer's viewpoint-selection records.

3.1 Ball Trajectory Distribution

Most viewers of soccer games have a tendency to follow the ball [26]. Thus, we first analyzed the relationship between trajectory of the ball and personal viewpoint-selection tendency. Figures 2 (a) and (b) shows the ball trajectories when each camera was selected by two viewers for two games (Game 1 and 2). The rectangles represent the ground plane of the soccer field. Each of different colored lines in the rectangle shows the trajectory during a cut unit in the overall soccer field. For each viewer, Figs. 2 (a) and (b) show that the ball trajectories centered around a location in the soccer field when the viewer selected different viewpoint in both games. Besides, they have different trends between the viewers. For example, in Game 1, Viewer 6 preferred Camera 8 when ball moved in the corner area on the right side of the field while Viewer 4 did not select Camera 8. The difference of viewers also existed when Cameras 4 and 10 were selected in Game 2.

Therefore, we consider that using the ball trajectory distribution can be effective to learn different viewpoint-selection tendency.

3.2 Players Trajectory Distribution

Players are also important objects in soccer games. Thus, we also analyzed their trajectories to represent the game context that has an impact on personal viewpoint navigation.

First, we collected trajectories of all players except the two keepers and carried out the same analysis as the ball. Figures 2 (c) and (d) show the players trajectory distributions when each camera was selected. From these distributions, we found that players trajectories also have different trends among not only the viewpoints but also the viewers. However, the difference is not as much absolute as ball trajectory.

We further analyzed the player who will get the ball at the next moment since the viewers seem to pay their attention to the player. We also found that the players trajectories have similar trends with the ball trajectories for each selected viewpoint.

As a result, we consider that using players trajectory distribution also can be effective to learn different viewpoint-selection tendency for some viewers who pay more attention to the players.

4. Viewpoint Navigation Approach

Based on the analysis result obtained in Sect. 3, we decide to use the following three kinds of objects as viewing targets:

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

Our navigation framework learns the relationship between personal viewpoint-selection tendency and trajectory distribution of the viewing target.

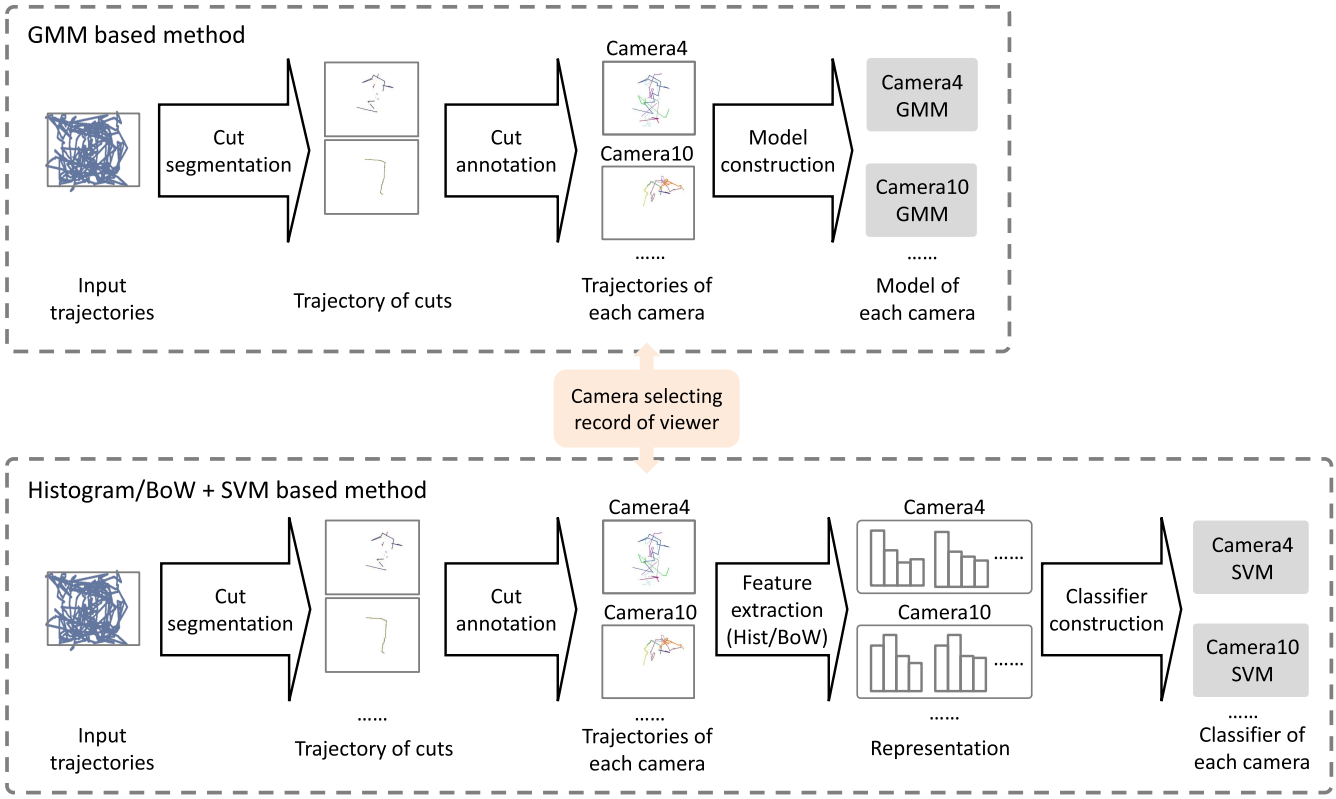

Fig. 3 The outline of learning schemes for relationship between personal viewpoint-selection tendency and trajectory distribution of viewing targets in cuts. *Cut annotation* means to gather and label the cuts when the same viewpoint is selected.

In this section, we first discuss three methods for learning the relationship by using each viewing target and discuss a combination method of using different targets trajectories data. Then, we introduce two definitions of cut unit including several frames segmented from the video sequence. We compare all the combinations of 3 learning schemes × 5 kinds of target combinations × 2 kinds of cut unit to find the best one for viewpoint navigation.

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

4.1 Machine Learning Scheme

We compare three methods based on a maximum likelihood decision rule with GMM, SVM with a two-dimensional histogram, and SVM with BoW to seek the best learning scheme for the relationship. The outline of these methods is shown in Fig. 3. We consider GMM is appropriate descriptor of trajectory distribution for each selected viewpoint since it is widely used to express object-position distribution. We also try two baseline representations using the two-dimensional histogram and the BoW with soft assignment to describe trajectory in a cut considering the softer boundary than simple two-dimensional histogram.

4.1.1 Gaussian Mixture Model Based Method

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

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

where K is the number of the Gaussian components, π_k is the weight of the k-th Gaussian component with $\sum_{k=1}^{K} \pi_k = 1$, $N(\mathbf{x}_f | \mu_k, \Sigma_k)$ is the Gaussian component density with parameters μ_k and Σ_k. The number of components is based on the experimentally varied results, as discussed later. In this study, we use it to represent the target trajectory distributions of each viewpoint for each viewer. We gather the target trajectories of cuts while viewpoint v is selected, and represent them by T_v. Thus, to generate GMM (p_v) of the viewpoint v, $\mathbf{x}_f$ is a sample from T_v in a training dataset. We apply EM algorithm to estimate the parameters (π_k, μ_k and Σ_k).

For each video sequence in a test dataset, we first divide the video sequence into cuts and extract target trajectory T_{C_i}

from cut i. We calculate the total of the log-likelihood for the points on the trajectory T_{C_i} under the generated GMM of each viewpoint for the viewer. Thus, given a trajectory T_{C_i}, we output a viewpoint R with the largest log likelihood as follows,

$$R(T_{C_i}) = \arg\max_{1 \le v \le V} \sum_{\mathbf{x}_f \in T_{C_i}} \log p_v(\mathbf{x}_f). \quad (2)$$

4.1.2 Histogram and SVM Based Method

We also employ a method based on SVM with a two-dimensional histogram called *Hist-SVM* for short in this paper. The soccer field is spatially divided equally into $M \times N$ bins. We calculate the two-dimensional histogram of points of target trajectory T_{C_i} in cut i. Histogram normalization is conducted considering the differences in the lengths of the cuts.

In the training step, we use the normalized histogram as a feature vector and perform the learning step by SVM with a RBF kernel. The supervised signals are the viewpoint-selection records of each viewer. Thus, we build a one-vs-all classifier corresponding to each viewpoint.

For the test data, the learned classifiers output a viewpoint with maximum score for each input trajectory T_{C_i} of cut i. The output viewpoints of cuts compose the navigation sequence. In this study, M and N are set to their optimum values based on the experimentally varied results.

4.1.3 BoW and SVM Based Method

We furthermore employ a method based on SVM with a bag-of-words technique [27] called *BoW-SVM* for short. We first cluster all the points of the target trajectories in the training data to codewords using the unsupervised GMM shown as Eq. (1). The number of Gaussian distribution is set to the optimum value based on the experimentally varied results. A codeword is defined as each Gaussian distribution in the GMM. Thus, the number of codewords is same as K, i.e., the number of Gaussian distribution. We then apply a soft assignment to generate the histogram of codewords in cut i. Under the k-th Gaussian distribution, the value of the bin of k-th codeword histogram is calculated as responsibility $p(k|\mathbf{x}_f)$ as follows,

$$p(k|\mathbf{x}_f) = \frac{\pi_k N(\mathbf{x}_f|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(\mathbf{x}_f|\mu_k, \Sigma_k)}. \quad (3)$$

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

$$\mathbf{A}_i = \frac{\sum_{f=1}^{F_i} \mathbf{a}(\mathbf{x}_f)}{F_i}, \quad (4)$$

where $\mathbf{x}_f \in T_{C_i}$, 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 viewer is learned with RBF kernel based SVM. Thus, we build one-vs-all classifiers as with *Hist-SVM*.

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

4.2 Viewing Targets Fusion

Different viewers may have interest in different viewing targets. Thus, we use each viewing target or different combination of targets as a feature vector to find which target or combination is most focused by each viewer. We apply the proposed learning schemes mentioned in Sect. 4.1 when only use each target trajectory. In this section, we add fusion methods corresponding to the proposed learning schemes for two kinds of combined targets: $B + P_{\mathrm{n}}$ and $B + P_{\mathrm{all}}$.

For the combination, there are several methods usually used to fuse multiple features, typically including early fusion conducted at representation level and late fusion conducted at score level [28], [29]. For the representation level fusion, multiple features are integrated into a single feature representation, which is fed into one supervised learning phase. The integrated feature representation may reflect better multiple information and correlation though the higher dimension will add the difficulty on 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. The late fusion focuses on the individual strength but it may result in the loss of correlation of features.

As shown in Fig. 4, we use different fusion methods depending on the learning scheme. For GMM based method, if we apply early fusion, we have to integrate 2-dimensions trajectory data of ball and players included in the targets (21×2 dimensions for B combined with P_{all}) into a single feature vector before GMM generation. Considering the difficulty on learning high dimensional feature, we use the late fusion. We first generate GMMs of viewpoints for ball (B) trajectory data and GMMs for player (P_{n} or P_{all}) trajectory separately. Then we integrate the likelihood scores of each GMM of ball (B) and player (P_{n} or P_{all}) into one feature vector and learn by SVM afterwards instead of the maximum likelihood decision rule used for each target. For *Hist/BoW-*

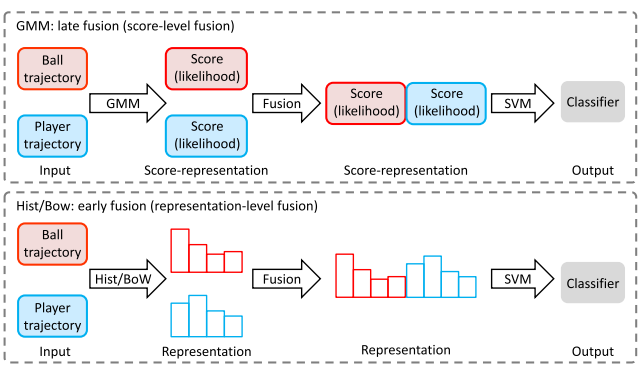

Fig. 4 Fusion methods of combining viewing target trajectories.

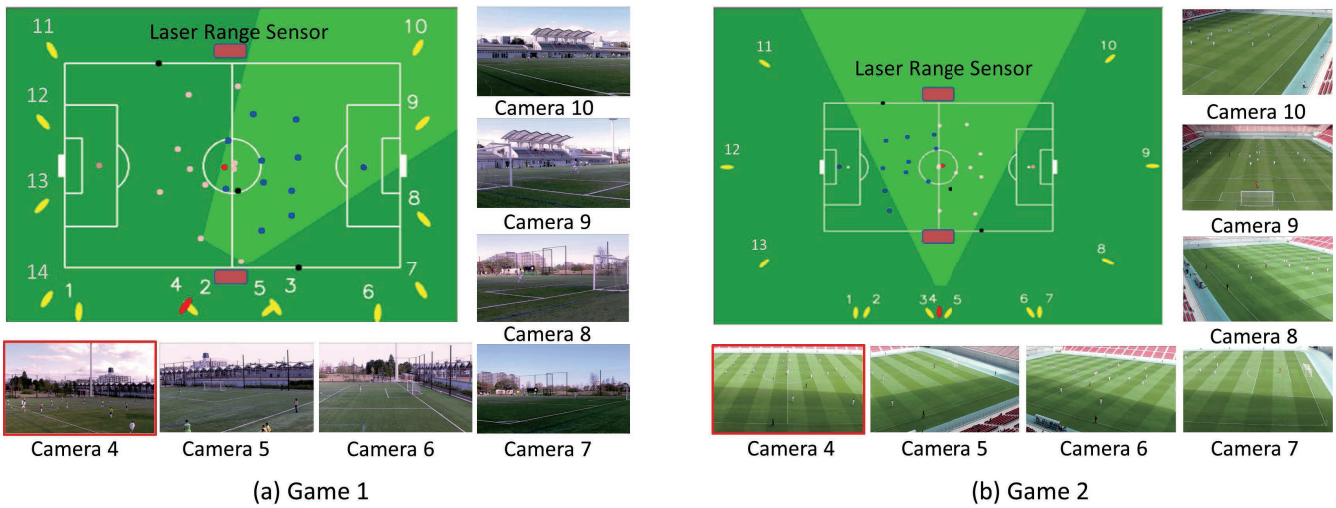

Fig. 5 Camera positions and sample viewpoint images of two soccer games.

SVM based method, we use the early fusion considering the correlation of ball and players trajectories. We first input the data of different targets into histogram/BoW generator separately and the resulting representations are then fused into a single integrated feature vector for the supervised learning using SVM.

4.3 Video Cut Segmentation

In this study, we focus on the fact that viewers select viewpoints according to the game context in past and future periods, and not only solely on the current frame. We use the cut consisting of multiple frames to represent the spatio-temporal game context. Therefore, it is necessary to determine how the video sequence should be segmented. In this study, we consider the following two kinds of segmentation.

4.3.1 Ideal Segmentation (*SegU*)

Here, ideal segmentation is a result that the game context is appropriately classified according to personal preference, which causes viewer's viewpoint selection. We call this segmentation *SegU* for short. It is unavailable for viewpoint navigation in practice. In this study, we record the viewer's viewpoint switching timing to ideally segment the video sequences to verify the upper bound of navigation accuracy. Thus, for *SegU* based experiments, we apply the switching timing for cut segmentation both in the training and test steps, and select the best viewpoint for each segment.

4.3.2 Equal Segmentation (*SegS*)

Equal segmentation is a result that the video sequence is segmented into cuts of a fixed length. We call this segmentation *SegS* for short. We adopt the sliding-window method to compensate for the overlap in each cut. Concretely, we generate the cut with a window size around each frame. The selected viewpoint based on the trajectory distribution in the cut is assigned to the center frame of the cut. Thus, for *SegS* based experiments, we conduct the sliding window processing along the video sequence both in the training and test steps, and select the best viewpoint for each frame. Unlike *SegU*, we can perform *SegS* 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 set to the optimum value based on the experimentally varied results.

5. Experiment

In this study, we collected the viewers' viewpoint-selection records through a video-editing experiment using soccer game multi-view video dataset.

5.1 Multi-View Video Dataset

We used the multi-view video dataset of two soccer games held in different venues with different camera settings. Both games were filmed using digital cameras (CASIO EX-F1, at 30 fps and 1920 × 1080 pixels) with no pan, tilt, and zoom around the soccer field. We used only the cameras near the main stand as shown in Fig. 5 because radical changes occurring when the viewpoint transfers from one side of the field to another can cause cognitive discomfort. Figure 5 includes sample images of cameras placed on the right half of the field. The camera IDs are given counterclockwise.

The videos of all cameras were synchronized manually after filming. In consideration of difficulty of video editing study for long periods of time, we provided the extracted short video sequences (of about 30 seconds each) to participants. These video sequences contained typical play scenes to attract viewers' interest. For example, these are soccer play scenes of dribble (players sparse/dense), passing (short/long passing), sliding, shooting, cross, throw-in, heading, body check, goal-kick, and free-kick, while with

absent of corner-kick and penalty-kick. Thus, the performance of the proposed method might be limited in such absent situations.

We obtained the positions of the players through semi-automatic processing by two laser range sensors placed on both sides of the field as shown in Fig. 5 used in study [30]. We obtained the position of the ball through manual labeling and a basic interpolation procedure since our main focus is the viewpoint navigation but not automatic ball tracking. Some vision-based and sensor-based tracking techniques are being researched separately for this purpose [31].

5.2 Collection of Viewers' Selection Records

We conducted the multi-view video editing experiment to collect viewers' selection records.

We asked experimental participants to state their profile information via a profile questionnaire. The 10 participants comprised six males and four females, all in the age group 20–39. In the questionnaire, the interest level in soccer was assessed on a four-level selection from "almost not" to "very." 10% of the participants was strongly interested in soccer, 40% had a general interest in soccer and 50% agreed to the statement "a little" (level 2). For the viewing frequency of the soccer game, 50% of the participants were occasional viewers 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, with no particular expertise. Further, two of them had amateur video photography experience with a video camera, or video editing experience using an editing software.

We randomly presented 11 and 10 short video sequences to the participants for each game. The participants could repeatedly replay each scene, select viewpoints, and confirm the selected viewpoints with a simple action on a graphical user interface. The editing record of each participant would reflect their personal preference.

5.3 Comparative Methods

AveragePos uses the centroid of the ball positions during a cut as the feature and trained RBF kernel based SVM, which is used in [21] as mentioned in Sect. 2.
WeightOptm uses context-dependent weights optimized by 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, which is used in [15] as mentioned in Sect. 2.

5.4 Evaluation Framework

To quantitatively evaluate the effectiveness of navigation, we calculated the concordance rate between each participant's viewing record $R^u(f)$ and the output viewpoints $R^s(f;\mathbf{x}_f)$ of proposed methods at each frame of a video sequence as follows,

$$Rate = \frac{\sum_f E(f;\mathbf{x}_f)}{L}, \tag{5}$$

$$E(f;\mathbf{x}_f) = \begin{cases} 1 & (\text{if } R^s(f;\mathbf{x}_f) = R^u(f)), \\ 0 & (\text{otherwise}). \end{cases} \tag{6}$$

where L is the length of a sequence.

We conducted a leave-one-sequence-out cross-validation by using one sequence of each participant's viewing records as test data until all the sequences are used as test data. We calculated the concordance rate of each test sequence and the average rate over all the test data as result.

6. Results

We use the evaluation framework mentioned above to evaluate the performance of proposed methods using three learning schemes, five kinds of target combinations with two kinds of cut segmentation methods, and find the best combination for personal viewpoint navigation.

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 target for Games 1 and 2 were 33 and 39, and using the target combinations were 12 and 13, respectively. With regard to *Hist-SVM*, 21 (7×3) bins were the best for both games.

6.2 Comparing of Different Factors

We conducted a analysis of variance (three-way ANOVA) including all the results of 10 participants using the proposed methods (3 learning schemes, 5 target combinations and 2 cut segmentations) to investigate the effects of different factors and their interaction. Table 1 summarizes the ANOVA results.

Each main effect for target combinations and the cut segmentations was statistically significant ($p < .01$) in both games. Nonetheless, there was no significant difference in the learning schemes in Game 1 and marginally 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 targets combination was statistically significant ($p < .01$) in Game 2. For Game 1, all factors were independent since there was no significant interaction of different factors. We discuss the detail of different factors below.

6.2.1 Comparing on Learning Schemes

We first discuss the navigation performance of proposed method under ideal situation using ideal segmentation, i.e,

Table 2 Concordance rates of the three learning schemes using the five target combinations with segmentation *SegU* of the two games. *SegU*: segmentation according to viewer's viewpoint switching.

Fusion		B	P_n	P_{all}	B + P_n	B + P_{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%	35.31%±9.81%	41.84%±5.60%	49.65%±8.14%	48.06%±9.50%

Table 1 Result of analysis of variance (three-way ANOVA) for the concordance rates of 10 participants using all the proposed methods (3 learning schemes, 5 target combinations and 2 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	
Target Combination (TC)	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	+
Target Combination (TC)	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						

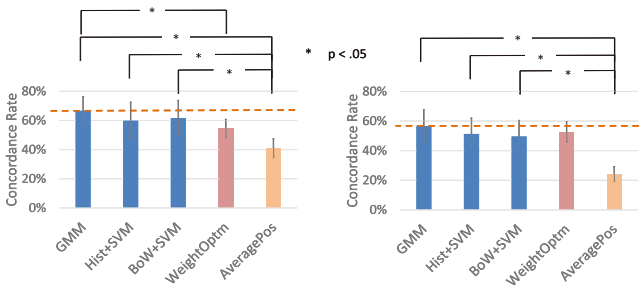

Fig. 6 The average and standard deviation of the concordance rates of 10 participants of the proposed methods using only ball trajectory with *SegU* target combinations methods for the two games.

SegU. The average and standard deviation of the concordance rates of 10 participants using the three learning schemes and the five target combinations with ideal segmentation for the two games are shown in Table 2. From this table, we find that the GMM based method achieved the best concordance rates 66.64% and 56.65% for the two games, respectively. Besides, considering the average concordance rate of all kinds of targets combination of the two games, the GMM based method also achieved the best average concordance rate 52.23%. Thus, with regard to the average concordance rates, the results of the GMM based method was higher than those based on *Hist-SVM* and *BoW-SVM*, although there was no significant difference in the learning schemes by ANOVA analysis.

Moreover, we qualitatively compared the recommendation sequences generated by different learning schemes using only ball trajectory *B*. We found that the recommendation of *Hist-SVM* based method was unstable when the trajectory distribution existed around the border lines of histogram division, while *BoW-SVM* and the GMM based methods selected the same viewpoint stably as the user record. *BoW-SVM* based method made mistakes when the trajectory distribution existed at the center area especially on the far side of field where multi cameras (i.e., camera 3, 4, 14, 7 in Game 1) can cover the game. By contrast, the GMM based method worked better in this situation. We consider this resulted from the difference between the construction methods of GMM. *BoW-SVM* 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 trajectories at local areas). The localized GMM acquired more discriminative ability and achieved the better performance for the overlapped trajectory distributions from multiple selection.

Furthermore, we compared the concordance rates of the comparative methods (i.e., AveragePos and WeightOptm mentioned in Sect. 5.3) with the proposed methods using only ball trajectory *B* with ideal segmentation (*SegU*). The average and standard deviation of the concordance rates of these methods are shown in Fig. 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 game context enough. 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 better average concordance rate. The WeightOptm method employed the distance information between cameras and viewing targets besides game context representation with targets information. Thus, we assume that using the trajectory distribution learned by the GMM based method was more effective for game context

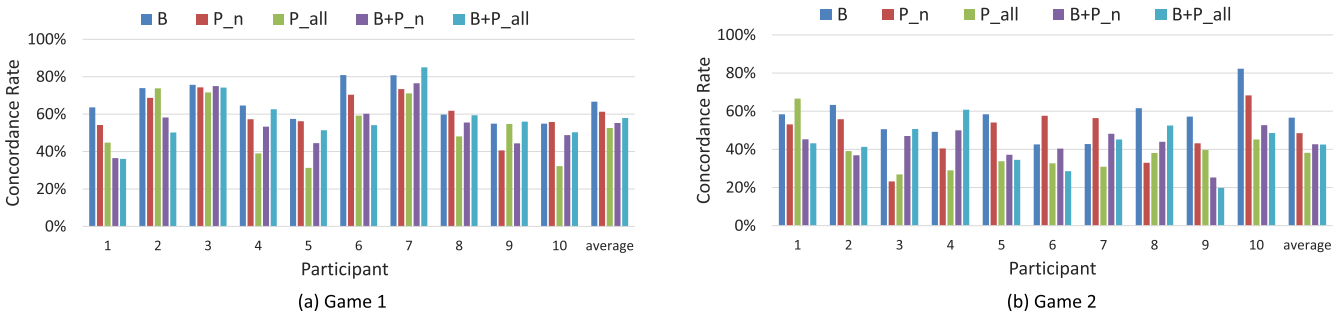

Fig. 7 Concordance rate of each participant using the five target combinations with the GMM based method and *SegU* for the two games. B: the ball of a soccer game; P_{n}: the player who will get the ball at next time; P_{all}: all players of a soccer game except goal keepers. $B + P_{\mathrm{n}}$: the combination trajectories of ball and the player who will get the ball at next time as feature. $B + P_{\mathrm{all}}$: the combination of ball and players trajectories as feature.

representation without camera information.

We discuss more detailed result using the GMM based method below.

6.2.2 Comparing on Viewing Target Combinations

To compare the difference among target combinations, we conducted the pairwise comparisons using T-test with Bonferroni adjustment for the results of the five target combinations (B, P_{n}, P_{all}, $B + P_{\mathrm{n}}$, $B + P_{\mathrm{all}}$) using the three learning schemes with the ideal segmentation of 10 participants for the two games. For Game 1, the comparisons showed that using only ball trajectory B (mean = 62.72%, SD = 12.15%) was significantly better than targets combination P_{all} (mean = 51.38%, SD = 13.57%) and $B + P_{\mathrm{n}}$ (mean = 57.47%, SD = 11.87%), with $p < .05$. P_{all} was also significantly lower than P_{n} (mean = 59.01%, SD = 11.58%), $B + P_{\mathrm{n}}$ and $B + P_{\mathrm{all}}$ (mean = 58.50%, SD = 13.15%), with $p < .05$. For Game 2, only ball trajectory B (mean = 52.50%, SD = 11.49%) was significantly better than the following target combinations P_{n} (mean = 39.82%, SD = 12.92%) and P_{all} (mean = 40.35%, SD = 9.47%), with $p < .05$. However, there were no significant difference between using only ball trajectory B and using both ball and all players, i.e. $B + P_{\mathrm{all}}$ for both games.

Regarding the interaction of the learning schemes and the target combinations from Table 2, we find that the GMM based method achieved better average concordance rate than *Hist-SVM* and *BoW-SVM* based methods when using only ball trajectory B, while there existed a contrary tendency among the learning schemes when using the target combinations $B + P_{\mathrm{n}}$ and $B + P_{\mathrm{all}}$.

Further, we show concordance rates of each participant using the five target combinations in Fig. 7. We can find that most of these participants acquired the best concordance on the navigation using only ball trajectory. For Participants 7 and 9 in Game 1, the navigation using both ball and all players achieved the best concordance rate. In Game 2, Participants 6 and 7 achieved the best concordance rate using trajectory of the player who will get the ball next. Therefore, we consider that the ball was the main viewing target of interest in the soccer games, which attracted more attention than players for viewpoint selection. Some participants such as Participant 7 would pay more attention to players.

Therefore, for personal recommendation, the navigation can achieve better effectiveness if we use the appropriate objects to reflect viewer's preference.

6.2.3 Comparing on Cut Segmentations

We calculated the average concordance rates of proposed methods using the two kinds of cut segmentation with only ball trajectory for the two games. 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 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 on *Hist-SVM* and *BoW-SVM* based methods.

We compared the generated recommendation sequences to investigate the difference between *SegS* and *SegU*. Figure 8 showed a sample of recommendations using *SegS* and *SegU*. Through the figure we can find 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 achieve possible improvement on the performance.

6.3 Need of Personal Navigation

In addition, we compared the effectiveness of learning from each viewer's own record with learning from other viewers to verify the need of personal navigation. The results of learning the two kinds of records data using only ball trajectory for each participant in the two games are shown in Figs. 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.

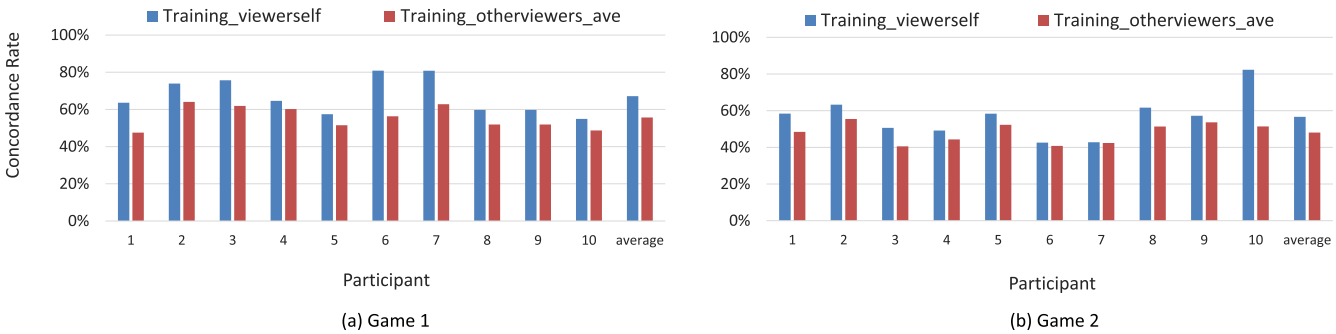

Fig. 9 Comparing with results of learning using the other viewers' records for each viewer. *Training_viewerself* represents the result of learning from each participant's own record. *Training_otherviewers_ave* represents the average performance of learning from other participants' records.

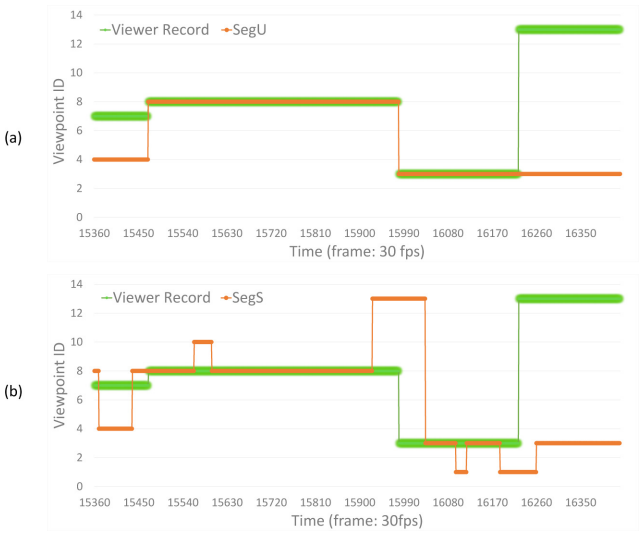

Fig. 8 A sample of comparison on the generated recommendation sequences using the equal segmentation with sliding window (*SegS*) and the ideal segmentation (*SegU*) for a viewer. *Viewer Record* represents the viewpoint-selection record of the viewer. (a) Viewpoint sequence generated using *SegU* and viewer's record. (b) Viewpoint sequence generated using *SegS* and viewer's record.

The performance of learning from each participant was better than learning by other participants. This result shows that each participant had a different viewing tendency against other participants and our navigation reflected their personal preference.

7. Conclusions

In this study, we proposed an automatic viewpoint navigation method based on personal preference. We generated the personal navigation by learning the relationship between personal viewpoint-selection tendency and the spatio-temporal game context in the form of the trajectory distribution of viewing targets. The experimental results showed the GMM based method outperformed other methods. Other than the soccer game, we can apply our method to events occurring in a large space for which the event context is essentially object-dependent such as baseball, basketball, or large theatrical shows. In the future, we intend to discuss on better methods for video cut segmentation and applying time series model for game context that trajectory distribution could not cover such as direction information of object action.

References

[1] I. Ahmad, "Multi-view video: get ready for next-generation television," IEEE Distributed Systems Online, vol.8, no.3, p.6, 2007.
[2] R.T. Collins, O. Amidi, and T. Kanade, "An active camera system for acquiring multi-view video," Proc. International Conference on Image Processing, 2002.
[3] K. Mase, Y. Sumi, T. Toriyama, M. Tsuchikawa, S. Ito, S. Iwasawa, K. Kogure, and N. Hagita, "Ubiquitous experience media," IEEE Multimedia Mag., vol.13, no.4, pp.20–29, 2006.
[4] E. Kurutepe, A. Aksay, C. Bilen, C.G. Gurler, T. Sikora, G.B. Akar, and A.M. Tekalp, "A standards-based, flexible, end-to-end multi-view video streaming architecture," Packet Video 2007, pp.302–307, 2007.
[5] Y. Kameda, T. Koyama, Y. Mukaigawa, F. Yoshikawa, and Y. Ohta, "Free viewpoint browsing of live soccer games," Proc. IEEE International Conference on Multimedia and Expo, pp.747–750, 2004.
[6] H. Kim, I. Kitahara, K. Kogure, and K. Sohn, "A real-time 3d modeling system using multiple stereo cameras for free-viewpoint video generation," Proc. International Conference Image Analysis and Recognition, vol.4142, pp.237–249, 2006.
[7] J.J.M. Kilner, J.R. Starck, and A. Hilton, "A comparative study of free viewpoint video techniques for sports events," Proc. European Conference on Visual Media Production, pp.87–96, 2006.
[8] T. Horiuchi, H. Sankoh, T. Kato, and S. Naito, "Interactive music video application for smartphones based on free-viewpoint video and audio rendering," Proc. 20th ACM international conference on Multimedia, pp.1293–1294, 2012.
[9] T. Matsuyama, S. Nobuhara, T. Takai, and T. Tung, 3D video and its applications, Springer, 2012.
[10] J.G. Lou, H. Cai, and J. Li, "A real-time interactive multi-view video system," Proc. 13th annual ACM international conference on Multimedia, pp.161–170, 2005.
[11] K. Mase, K. Niwa, and T. Marutani, "Socially assisted multi-view video viewer," Proc. 13th ACM international conference on multimodal interfaces, pp.319–322, 2011.
[12] C. Shen, C. Zhang, and S. Fels, "A multi-camera surveillance system that estimates quality-of-view measurement," Proc. IEEE International Conference on Image Processing, pp.193–196, 2007.
[13] F. Daniyal, M. Taj, and A. Cavallaro, "Content and task-based view selection from multiple video streams," Multimedia tools and applications, vol.46, no.2-3, pp.235–258, 2010.

[14] C. Chen, O. Wang, S. Heinzle, P. Carr, A. Smolic, and M. Gross, "Computational sports broadcasting: Automated director assistance for live sports," Proc. IEEE International Conference on Multimedia and Expo, pp.1–6, 2013.
[15] X. Wang, T. Hirayama, and K. Mase, "Viewpoint sequence recommendation based on contextual information for multiview video," IEEE Multimedia Mag., vol.22, no.4, pp.40–50, 2015.
[16] R. Cutler, Y. Rui, A. Gupta, J.J. Cadiz, I. Tashev, L.w. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: A meeting capture and broadcasting system," Proc. tenth ACM international conference on Multimedia, pp.503–512, 2002.
[17] C. Zhang, Y. Rui, J. Crawford, and L.-W. He, "An automated end-to-end lecture capture and broadcasting system," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol.4, no.1, pp.1–23, 2008.
[18] A. Ranjan, R. Henrikson, J. Birnholtz, R. Balakrishnan, and D. Lee, "Automatic camera control using unobtrusive vision and audio tracking," Proceedings of Graphics Interface, pp.47–54, 2010.
[19] M. Kumano, Y. Ariki, M. Amano, and K. Uehara, "Video editing support system based on video grammar and content analysis," Proc. 16th IEEE International Conference on Pattern Recognition., pp.1031–1036, 2002.
[20] M.K. Saini, R. Gadde, S. Yan, and W.T. Ooi, "Movimash: online mobile video mashup," Proc. 20th ACM international conference on Multimedia, pp.139–148, 2012.
[21] Y. Muramatsu, T. Hirayama, and K. Mase, "Video generation method based on user's tendency of viewpoint selection for multi-view video contents," 5th Augmented Human International Conference, AH '14, pp.1:1–1:4, 2014.
[22] X. Wang, K. Hara, Y. Enokibori, T. Hirayama, and K. Mase, "Personal multi-view viewpoint recommendation based on trajectory distribution of the viewing target," Proc. ACM Multimedia Conference, pp.471–475, 2016.
[23] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," Proc. IEEE International Conference on Computer Vision and Pattern Recognition., pp.2004–2011, 2009.
[24] K. Itoda, N. Watanabe, and Y. Takefuji, "Model-based behavioral causality analysis of handball with delayed transfer entropy," Procedia Computer Science, vol.71, pp.85–91, 2015.
[25] F. Daniyal and A. Cavallaro, "Multi-camera scheduling for video production," Proc. European Conference on Visual Media Production, pp.11–20, 2011.
[26] A. Iwatsuki, T. Hirayama, and K. Mase, "Analysis of soccer coach's eye gaze behavior," Proc. 2nd IAPR Asian Conference on Pattern Recognition, pp.793–797, 2013.
[27] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," Proc. workshop on statistical learning in computer vision of ECCV, pp.1–2, 2004.
[28] C.G.M. Snoek, M. Worring, and A.W.M. Smeulders, "Early versus late fusion in semantic video analysis," Proc. 13th annual ACM international conference on Multimedia, pp.399–402, 2005.
[29] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," Computer Vision and Image Understanding, vol.150, pp.109–125, 2016.
[30] Y. Kabeya, F. Tomiyasu, and K. Mase, "Semi-automatic multiple player tracking of soccer games using laser range finders," Proc. 7th ACM International Conference of Augmented Human, p.40, 2016.
[31] T. Orazio and M. Leo, "A review of vision-based systems for soccer video analysis," Pattern Recognition, vol.43, no.8, pp.2911–2926, 2010.

Xueting Wang received the B.E. degree from Tianjin University of Science and Technology, and the M.S. degree in Information Science from Nagoya Univeristy, in 2012 and 2015, respectively. She is currently a Ph.D. student at the Graduate School of Information Science, Nagoya University. Her research interests include multi-view video editing and viewing support. She is a member of IEICE and IEEE.

Kensho Hara received the B.E. degree in Information Engineering and the M.S. and Ph.D. degrees in Information Science from Nagoya Univeristy, in 2012, 2014 and 2017, respectively. He is currently a postdoctoral researcher at the National Institute of Advanced Industrial Science and Technology. His research interests include human action recognition and detection. He is a member of IEICE and IEEE.

Yu Enokibori received the B.E., M.E., and Ph.D. degree in Engineering Science from Ritsumeikan University in 2005, 2007, and 2010, respectively. He is now an assistant professor at Graduate School of Informatics, Nagoya University. His research interests include ubiquitous computing, wearable computing, invisible computing, healthcare and medical computing, and human computer interaction. He is a member of the Information Processing Society of Japan (IPSJ), Japan Society of Artificial Intelligence (JSAI), Society of Biomechanisms Japan (SOBIM), Japan Academy of Nursing Science (JANS), and ACM.

Takatsugu Hirayama received the M.E. and D.E. degrees in Engineering Science from Osaka University in 2002 and 2005, respectively. From 2005 to 2011, he was a research assistant professor at the Graduate School of Informatics, Kyoto University. He is currently a designated associate professor at the Graduate School of Informatics, Nagoya University. His research interests include computer vision, human vision, human communication, and human-computer interaction. He has received the best paper award from IEICE ISS in 2014. He is a member of IEICE, the Information Processing Society of Japan (IPSJ), the Human Interface Society of Japan, ACM, and IEEE.

Kenji Mase received the B.E. degree in Electrical Engineering and the M.E. and Ph.D. degrees in Information Engineering from Nagoya University in 1979, 1981 and 1992, respectively. He became a professor of Nagoya University in August 2002. He is now with the Graduate School of Informatics, Nagoya University. He joined the Nippon Telegraph and Telephone Corporation (NTT) in 1981 and had been with the NTT Human Interface Laboratories. He was a visiting researcher at the Media Laboratory, MIT in 1988-1989. He has been with ATR (Advanced Telecommunications Research Institute) in 1995-2002. His research interests include gesture recognition, computer graphics, artificial intelligence and their applications for computer-aided communications. He is a member of the Information Processing Society of Japan (IPSJ), Japan Society of Artificial Intelligence (JSAI), Virtual Reality Society of Japan, Human Interface Society of Japan and ACM, a senior member of IEEE Computer Society, and a fellow of IEICE.