Personal Multi-view Viewpoint Recommendation based on Trajectory Distribution of the Viewing Target

Xueting Wang, Kensho Hara, Yu Enokibori, Takatsugu Hirayama, Kenji Mase
Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan
{wang, kensho}@cmc.ss.is.nagoya-u.ac.jp, {enokibori, hirayama, mase}@is.nagoya-u.ac.jp

ABSTRACT

Multi-camera videos with abundant information and high flexibility are expected to be useful in a wide range of applications, such as surveillance systems, web lecture broadcasting, concerts and sports viewing, etc. Viewers can enjoy a high-presence viewing experience of their own choosing by means of virtual camera switching and controlling viewing interfaces. However, some viewers may feel annoyed by 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-recommending method designed especially for soccer games. This method focuses on a viewer's personal preference for viewpoint-selection, instead of common and professional editing rules. We assume that the different trajectory distributions cause a difference in the viewpoint selection according to personal preference. We therefore analyze the relationship between the viewer's personal viewpoint selecting tendency and the spatio-temporal game context. We compare methods based on a Gaussian mixture model, a general histogram+SVM and bag-of-words+SVM to seek the best representation for this relationship. The performance of the proposed methods are verified by assessing the degree of similarity between the recommended viewpoints and the viewers' edited records.

Keywords

multi-view video recommendation; user preference; Gaussian mixture model

1. INTRODUCTION

Multi-view videos taken by multiple cameras from different angles are playing an important role in video services with the development of video capturing, processing and delivering technologies [1, 2]. Furthermore, free-view videos can be generated to provide more viewpoint options by interpolating scenes or modeling 3D scenes [8, 10, 13]. Because of the diverse information and viewing options, viewers can enjoy more interesting contents in their own way than that provided by the professional TV broadcasters with a single forced viewpoint. Thus, multi-view videos are suitable to represent events as diverse as web lectures, concerts and sports viewing.

However, the continual selection of appropriate viewpoints by a completely manual process on the existing multi-view video interface [11, 12] could be annoying, especially when the selection options are relatively numerous. A high level of viewer stress could occur when viewing a dynamic event in a wide-scale field. Thus, an automatic viewpoint recommendation based on a viewer's personal preference is important for multi-view video watching in such situations. In this work, we focus on the viewpoint recommendation for the sports game, in particular, the soccer game.

There are several related studies on automatic viewpoint selecting. In works [4, 15, 18], camera selecting are based mainly on audio feature, face trajectory and speaker position for web lectures and meeting broadcasts. The audio context are not suitable for sports games due to the noise of the crowd and far distance of wide-angle cameras. Saini et.al. [16] proposed a framework for the automatic mashup of dance performance videos taken by mobile phones. They chose the best angle based on view quality context depending on such as illumination and shakiness, and professional editing rules. We assume that game context consisting of individual object information, such as positions of the ball and players, is more reliable than the audio-visual information used in the above methods for sports games.

Other researchers have also focused on the game context. In work [5, 6] Daniyal et al. presented an algorithm for viewpoint-quality ranking based on object scoring for *frame-level* features including size and location of the players in a basketball game. The extended approach in work [17] optimized the viewpoint transition by viewpoint-quality evaluation with dynamic measurements corresponding to game context represented by the object position calculated in each frame. Muramatsu et al. [14] selected viewpoint by learning the average of object features, such as position, distance to the camera and size in the view among a short time from user's viewpoint-selection records. However, these approaches are processed either without or with poorly represented past and future object dynamics such as trajectory and temporal behavior. We assume that the game context is described better by information about the object's trajectory than by frame-level information, and that the recent machine-learning representation is more effective than simple average processing.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '16, October 15-19, 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-3603-1/16/10...$15.00
DOI: http://dx.doi.org/10.1145/2964284.2967265

Most of the related studies discussed above focus mainly on common preferences [4, 5, 6, 15, 18] such as assigning the viewpoint evaluation score in proportion to the size and number of players visible in the view, and professional editing rules [16] such as learning the shot length from the directors' viewpoint transition. However, viewers might prefer viewing that is based on their own preferences and styles. There are only a few existing studies on user-dependent viewpoint recommendation [14, 17]. They learn the object's features and optimize the weight parameters for feature combination from each user's viewpoint-selection record. The performance of these approaches is limited due to the lack of representation of the object's temporal information.

We therefore aim to achieve a personal viewpoint recommendation considering the spatio-temporal game context represented by the trajectory of the main viewing target, i.e., the ball. We assume that different viewers show different viewpoint-selection tendency for different trajectory distributions. Thus, we apply a machine leaning to learn the relationship between the trajectory distribution of the viewing target and each viewer's viewpoint-selection tendency to achieve an effective personal viewpoint recommendation. We first divide short video sequences into cuts. We then use the ball-trajectory distribution in each cut as a feature vector to learn each viewer's viewpoint selection preference for the spatio-temporal game context. We compare and seek the best one from three methods each based on a Gaussian mixture model(GMM), a two-dimensional histogram with SVM, and a bag-of-words(BoW) with SVM to achieve the best performance for personal recommendation based on the viewer's viewpoint-selection records.

Contributions. We now summarize our contribution in this paper as follows.

- We realize personal viewpoint recommendation using the trajectory distribution of the main viewing target in order to featuring spatio-temporal context for personal preference learning.
- We achieve the better performance over other representations by using GMM-based method.

2. PERSONAL VIEWPOINT RECOMMENDATION

The proposed method describes trajectory distribution of the viewing target as feature vectors in order to learn each viewer's preference from viewpoint-selection records of the viewer's in cuts dividing the video sequences. Most viewers of a soccer game have a trend to view the game by following the ball [9]. Thus, we use the ball as the main viewing target, and we use its trajectory to represent the game context that is required for a personal viewpoint recommendation.

In this section, we first detail the analysis of the relationship between the ball trajectory distribution and each viewer's viewpoint-selection tendency. Then, we discuss three methods for describing the relationship. We also discuss two definitions of cut unit includes several frames segmented from the video sequence. In the following discussion, we represent the ball trajectory by $B = \{T_i\}$, where i is the trajectory index of each cut. The trajectory of each cut is $T = \{\mathbf{x}_f\}_{f=1}^{F}$ is the trajectory of each cut, where $\mathbf{x}_f \in \mathbb{R}^2$ is the point on the field coordinate system at frame $f(1 \leq f \leq F)$.

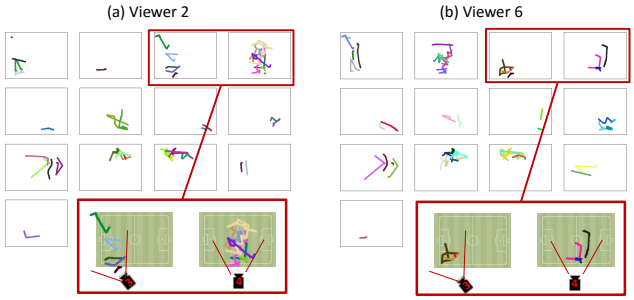

Figure 1: Ball-trajectory distribution for each selected viewpoint of two viewers in Game 2. Lines of different colors show the trajectories during cuts unit in overall soccer field.

2.1 Qualitative Analysis of Trajectory Distribution and Viewpoint-selection Tendency

We assume that viewers select the appropriate viewpoints based on the game context, which can be represented by the spatio-temporal movement of the main object, i.e., the ball in the case of a soccer game. We analyze the relationship between the ball trajectory and each viewpoint-selection records acquired in our video-editing experiment. Figure 1 shows that the ball trajectories centered around a location in the soccer field when each viewer selected each viewpoint. In other words, the ball trajectories have different trends for the different viewpoints that are selected. Thus, we consider that using the ball-trajectory distribution is effective to learn different viewpoint-selection tendency.

2.2 Learning Personal Preference

Personal preference can be represented by the relationship between personal viewpoint-selection tendency and distribution of ball trajectories. We compare three methods each based on GMM, a two-dimensional histogram with SVM and BoW with SVM to seek the best representation for the relationship.

2.2.1 *Gaussian Mixture Model based Method*

GMM is the linear combination of several Gaussian components, and it is widely used to express an object-position distribution. In this work, we use it to represent the different ball-trajectory distributions of the different viewpoints for each viewer. For each viewpoint c, we gather the ball trajectories B_c while it is selected. We express the points in the gathered trajectories using GMM with EM algorithm to estimate the parameters (mean and covariance).

We generate the model of each viewpoint for each viewer from the training data. For each video sequence in the test data, we first divide the video sequence into cuts and extract the ball trajectory T from each cut. We calculate the grand total of the log-likelihood of each trajectory T under the generated GMM of each viewpoint for the viewer. We recommend the viewpoint with the largest likelihood for each trajectory. The number of components is based on the experimentally varied results, as discussed later.

2.2.2 *Histogram with SVM based Method*

We also employ a method based on two-dimensional histogram with SVM called Hist-SVM for short. We spatially divide the soccer field equally into $M * N$ bins. We calculate

the histogram of the ball trajectory T in each cut. A histogram normalization is conducted in consideration of the differences in the length of a cut.

In the training steps, we use the normalized histogram as s feature vector and perform the learning by SVM with a RBF kernel. The supervised signals are the viewpoint-selection records of each viewer. Thus, we build the multi-class classifiers corresponding to each viewpoint in order to recommend viewpoints. For the test data, the SVM assigns a viewpoint to the trajectory T in each cut in the same way as in the training step. In this study, M and N are set to their optimum values based on the experimentally varied results.

2.2.3 *BoW with SVM based Method*

Finally, we employ a bag-of-words technique with SVM [3] called BoW-SVM for short. We first cluster all the points in the ball trajectory B of the training data to cordwords using the Gaussian Mixture Model. A codeword in this work is defined as each Gaussian distribution in the Gaussian Mixture Model. We then apply a soft assignment to generate the histogram of codewords in the cut. The scores of bins of the codewords histogram are calculated by the sum of the normalized likelihood under each Gaussian distribution for points of the trajectory T in the cut. We generate the codewords histogram as a feature vector. We then learn the relationship between the feature vectors and the selected viewpoints by each viewer with RBF kernel based SVM using the viewpoint-selection records as the supervised signals. Thus, we build the multi-class classifiers as with Hist-SVM. In the test step, the SVM assigns a viewpoint to the feature vector. The number of codewords is set to the optimum value based on the experimentally varied results.

2.3 Video Cut Segmentation

In this study, we take care to reflect 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.

2.3.1 *Ideal Segmentation*

Here, ideal segmentation is a result that the game context is appropriately classified according to personal preference, which causes viewer's viewpoint selecting. We call this segmentation SegU for short. It is unavailable for viewpoint recommendation in practice. In this study, we record the viewer's viewpoint switching timing to segment the video sequences to verify the upper bound of recommendation accuracy.

2.3.2 *Equal Segmentation*

Equal segmentation is a result that the video sequence is segmented into cuts of a fixed length, which can be easily realized in practice. 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 recommended viewpoint based on the trajectory distribution in the cut unit is assigned to the center frame of the cut. The window size, which is the length of the cut, is set to the optimum value based on the experimentally varied results.

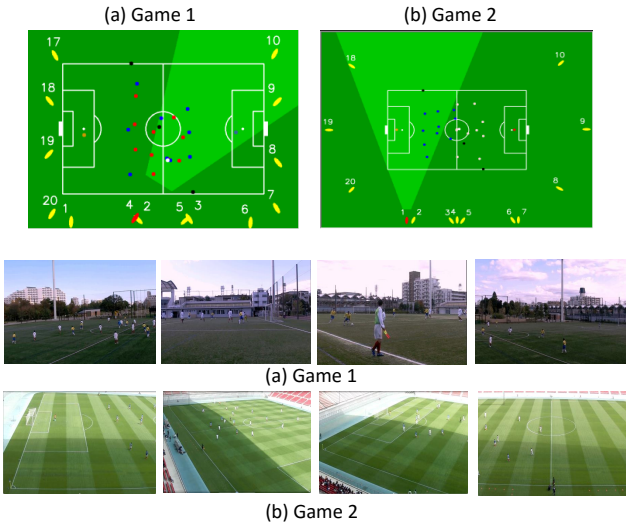

Figure 2: Camera position and sample viewpoint images.

3. EXPERIMENT

In this study, we collect the viewers' viewpoint-selection records through video-editing experiment using soccer game multi-view videos dataset. Construction of the proposed methods and effectiveness evaluation are carried out using the same dataset.

3.1 Multi-view Video Dataset

We use multi-view video datasets of two soccer games in different venues with different camera settings to evaluate the proposed method. The games were filmed using 13 digital cameras (CASIO EX-F1, at 30 fps 1920 × 1080 pixels) with no pan, tilt, or zoom) around the soccer field. The camera settings and sample viewpoint images are shown in Figure 2. The video sequences were synchronized after filming. We obtained the position of the ball through manual labeling and an interpolation procedure since our main focus is the viewpoint selection but not automatic ball tracking. Some vision-based and sensor-based tracking techniques are being researched separately for this purpose [7].

3.2 Collection of Viewers' Selection Records

We conducted a multi-view video editing experiment to collect viewers' records. The participants in this experiment comprised six males and four females of ages between 20-39. They were all occasional viewers of soccer games, with no particular expertise. In view of the difficulty of participants performing video editing work for long periods of time, we randomly presented 11 and 10 short video sequences (of about 30 seconds each) containing typical play scenes to the participants for each game in the experiment. 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 so that could be used for personal preference learning.

3.3 Training Using Mirror Samples

In a practical application of the method, we assume that training data acquired from viewers are limited because we can only use a few parts of the game for this purpose without annoying the viewer. Thus, we introduce a useful method to increase the training data as follows.

Since soccer field is symmetrical in the left-right direction, the mirror-reversed positions of the ball with respect to the center line of the field can also be expected to represent virtual game context. Thus, we add the mirror trajectories to the training data. Besides, since the cameras surrounding the field are also symmetrical in most games, we transfer the viewer's selected viewpoint corresponding to the mirror trajectory to the camera in a correspondingly symmetrical position as the new records.

3.4 Comparative Methods

AveragePos uses centroid of the ball positions during a cut as the feature, which is used in [14] as mentioned in the first section.

WeightOptm uses context-dependent optimized weights to combine the features in each frame, which is used in [17] as mentioned in the first section.

3.5 Evaluation Framework

We conducted a leave-one-sequence-out cross-validation as the evaluation by using one sequence of each participant's viewing records as testing data until all the sequences are used as test data. We then compared the viewpoints recommended by each method and each participant's viewing records and calculated the average concordance rate of all the test data on each frame.

4. RESULTS

We use the evaluation framework mentioned above to verify the performance of each method.

4.1 Parameters

The proposed methods achieved the highest average concordance rate using the following parameters. The number of component in GMM were 4 and 1, the number of notecodes in BoW-SVM were 31 and 36 for the two games separately. With regard to Hist-SVM, $7 * 3 = 21$ bins was the best for both games.

4.2 Concordance Rate

The average concordance rates of 10 participants with different methods for the two games are shown in Figure 3(a) and (b). From these figures, we find that the method based on GMM with ideal segmentation (GMM+SegU) achieved the best concordance rates 66.46% and 56.65% for the two games, separately. With regard to the game-context representation, GMM was more effective than those based on Hist-SVM and BoW-SVM. The lower result of the comparative methods (i.e., AveragePos and WeightOptm) show that using the spatio-temporal game context was effective. All the proposed methods achieved higher concordance rates than those of AveragePos for the two games. This result shows that the centroid of the ball position during a cut could not represent the game context.

In addition, Figure 4(c) and (d) show the concordance rates of GMM for each participant obtained in the experiment of the two games. In the figure, we show the result

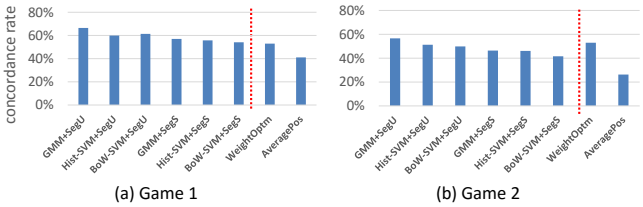

Figure 3: Concordance rates of different methods with different segmentation for the two games

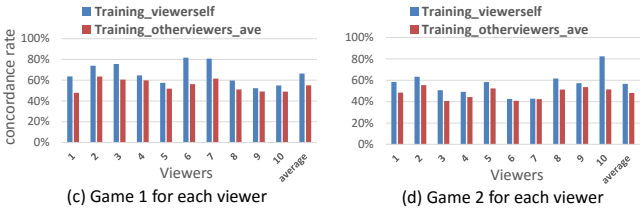

Figure 4: Comparision with results of training using other viewers' records for each viewer.

of leave-one-participant-out cross validation as the average performance. The performance of learning from each participant through the leave-one-sequence-out cross-validation was better than that of the average one. This result shows that each participant had a different viewing tendency and our recommendation reflected their personal preference.

The performance when using the mirror trajectories in the training step was evaluated in the same way. The concordance rates for GMM based method increased from 66.46% and 56.65% to 67.70% and 61.96% of the two games respectively, showing the effectiveness of increasing the training data.

5. CONCLUSIONS

In this study, we proposed an automatic viewpoint recommendation method based on personal preference. We predicted the personal recommendation by learning the relationship between personal viewpoint-selection tendency and the spatio-temporal game context in the form of the trajectory distribution of the main viewing target. The experimental results showed the GMM based method outperformed other methods. In the future, we intend to include the spatio-temporal features of other objects related to the main target and to discuss on better methods for segmentation.

6. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 26280074.

7. REFERENCES

[1] I. Ahmad. Multi-view video: get ready for next-generation television. *IEEE Distributed Systems Online*, 8(3):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] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[4] 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. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 503–512. ACM, 2002.

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

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

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

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

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

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

[11] 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*, pages 161–170. ACM, 2005.

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

[13] T. Matsuyama, S. Nobuhara, T. Takai, and T. Tung. *3D video and its applications*. Springer, 2012.

[14] Y. Muramatsu, T. Hirayama, and K. Mase. Video generation method based on user's tendency of viewpoint selection for multi-view video contents. In *5th Augmented Human International Conference, AH '14*, pages 1:1–1:4, 2014.

[15] A. Ranjan, R. Henrikson, J. Birnholtz, R. Balakrishnan, and D. Lee. Automatic camera control using unobtrusive vision and audio tracking. In *Proceedings of Graphics Interface*, pages 47–54. Canadian Information Processing Society, 2010.

[16] M. K. Saini, R. Gadde, S. Yan, and W. T. Ooi. Movimash: online mobile video mashup. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 139–148. ACM, 2012.

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

[18] 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)*, 4(1):6, 2008.