Court-based Volleyball Video Summarization Focusing on Rally Scene

Takahiro Itazuri[1], Tsukasa Fukusato[1], Shugo Yamaguchi[1], and Shigeo Morishima[2]
[1]Waseda University [2]Waseda Research Institute for Science and Engineering, Japan
{s132800732, tsukasa, wasedayshugo}@{fuji, aoni, suou}.waseda.jp, shigeo@waseda.jp

Abstract

In this paper, we propose a video summarization system for volleyball videos. Our system automatically detects rally scenes as self-consumable video segments and evaluates rally-rank for each rally scene to decide priority. In the priority decision, features representing the contents of the game are necessary; however such features have not been considered in most previous methods. Although several visual features such as the position of a ball and players should be used, acquisition of such features is still non-robust and unreliable in low resolution or low frame rate volleyball videos. Instead, we utilize the court transition information caused by camera operation. Experimental results demonstrate the robustness of our rally scene detection and the effectiveness of our rally-rank to reflect viewers' preferences over previous methods.

1. Introduction

Many people enjoy watching sports videos in their spare time. The development of broadcasting system made it possible for people to watch videos whenever they want. However, watching sports videos is a time-consuming task because of the long duration of a sports game. Consequently it is difficult for viewers who do not have enough time to enjoy watching a lot of games. In addition, the viewers have no choice but to watch whole videos since the number of delivered highlight videos is small. Therefore, a video summarization technique is required to enable viewers to efficiently watch sports videos. Video summarization is a task of creating a summary video to overview its content. The summary video must be self-consumable and contain high priority entities.

We propose a video summarization system for volleyball videos (Figure 1). Our video summarization system is composed of (i) rally scene detection for self-consumability and (ii) rally-rank evaluation for deciding priority of content. The generated summary video by our system contains only more important rally scenes with higher rally-rank.

Sports videos have a characteristic structure. On our observation, the broadcasting sports video can be divided into three parts: "play scene," "replay scene," and "break scene." The play scene is a scene where the ball is in play ("rally scene" in volleyball) and it is a long global shot so that viewers can comprehend the game overall (Figure 2a). The replay scene is a part of the play scene with effects such as slow and close-up (Figure 2b). The break scene includes other scenes such as advertisements, audiences, and

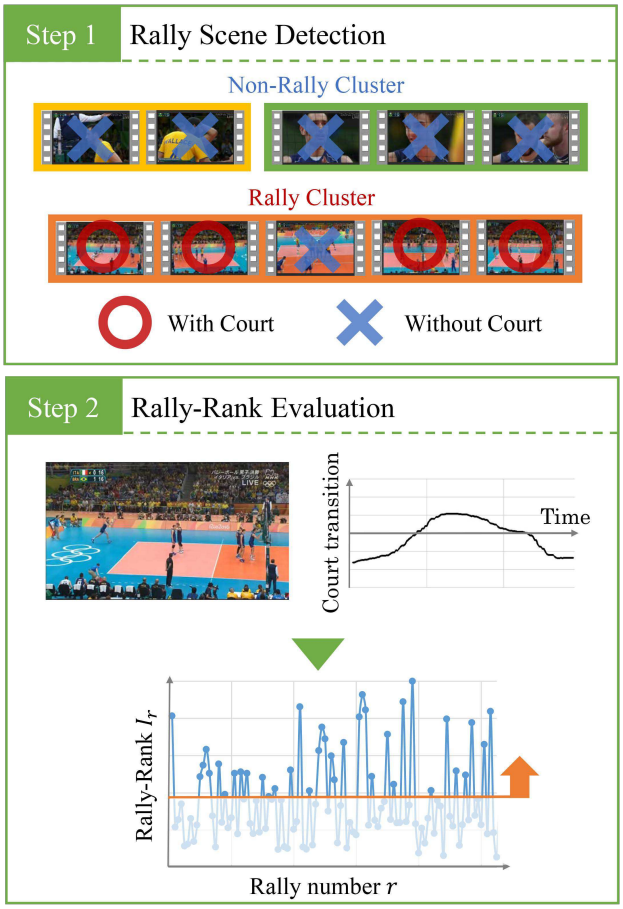


Figure 1: System overview of our volleyball video summarization.

(a) Play scene (b) Replay scene (c) Break scene

Figure 2: An example of play, replay, and break scene. The play scene is wide field shot where the ball is in play. The replay scene is close-up shot and a part of the play scene. The break scene is a scene not directly related to plays.

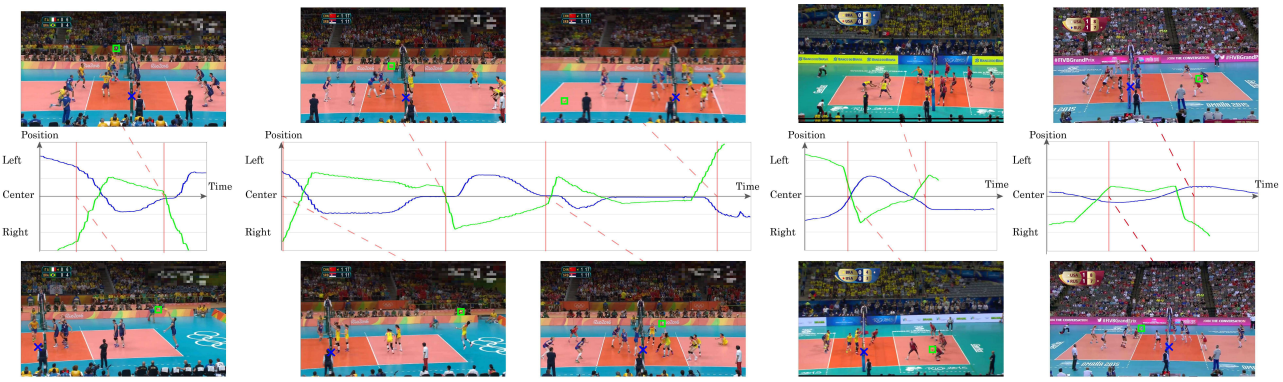


Figure 3: Examples of ball transition and court transition in the rally scenes. In the rally scenes, the global ball position (green) is estimated by the center position of the court (blue) because the camera tends to follow the ball.

before and after the game (Figure 2c). Since the break scene has nothing to do with the game content directly, it is not suitable for the summary video. The replay scene is not self-consumable because its content is limited to a part of specific play scenes. Furthermore, since the replay scene is close-up shot, it is powerful and exciting, but it is not suitable for understanding the situation of the game. On the other hand, the play scene includes all plays and each play scene is self-consumable because it has the play from the beginning to the end. Therefore, our system automatically detects the rally scenes (the play scenes in volleyball) from the input video.

To decide priority of content for video summarization, features representing the game content must be extracted from the input video. Since balls and players are objects of interest, the transition information of balls and players is an representative example of such features. However, acquisition of such features is still unreliable and not robust in low resolution or low frame rate volleyball videos because of the challenging situation such as quick movement of the balls, occlusion between the players wearing the same uniform, and image blur by the camera operation [1, 2, 7, 11, 12, 13]. We focus on a tendency that the camera follows the ball, and assume that the global ball transition can be approximated by the court transition caused by camera operation (Figure 3). Our system utilizes the court transition information caused by the camera operation that can be acquired robustly and evaluates rally-rank to decide priority. The court transition information can be acquired more easily and robustly than the information about the players and the balls. In section 3, we perform several tests in order to verify that the court transition information has a strong correlation with the ball transition and it represents the game content.

2. Related Work

Recently, researches into summarizing sports videos based on its characteristic and periodic structure have been proposed [4, 6, 8, 10, 14, 15, 16, 17]. These methods can be classified into three approaches: (i) event detection based approach, (ii) excitement modeling based approach (iii) combination approach. Event detection based approach detects characteristic events (e.g., goal in soccer and hit in baseball), replay scenes [4, 17], and audio keywords [14, 16], and then generates a summary video which consists of only these events. This approach can perfectly preserve the characteristic events in the input video, whereas its content is restricted to the detected events and it requires heuristic event detection algorithm for each sport.

For general sports video, excitement modeling based ap-

Table 1: Volleyball videos used for our experiments.

Video	Competition name	Match	Gender	Teams
1	Olympic Games 2016	Gold Medal Match	Men	Italy vs. Brazil
2	Olympic Games 2016	Gold Medal Match	Women	China vs. Serbia
3	Olympic Games 2016	Pool A	Women	Japan vs. Cameroon
4	FIVB World League 2015	Finals	Men	Brazil vs. USA
5	FIVB World Grand Prix 2015	Finals	Women	Russia vs. USA

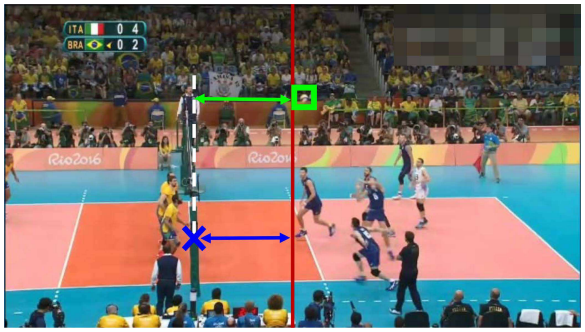


Figure 4: How to measure the ball position and the court position. The ball position (green) is a position with the court center (white dotted line) as the origin and the court position (blue) is a position with the image center (red line) as the origin.

proach [6, 15] quantifies the degree of excitement based on affective features such as sound information (e.g., sound volume and pitch) and editorial information (e.g., cut density), and then generates a summary video including more exciting scenes. These features are based on cheers of the audience and photographing technique occurring simultaneously with unusual and exciting events. This approach can be applicable to general sports videos and change the duration of the summary video by generating it that only the scenes with the higher degree of excitement than a threshold. However, the extracted scene is often not self-consumable because it contains only exciting events and scenes after exciting events.

On the other hand, an approach combining excitement modeling and event detection has been proposed [8, 10]. It detects the play scenes as events in general sports and then quantifies the degree of excitement for each play scene using affective features. This approach is useful for sports with many scoring points (e.g., tennis, badminton, and table tennis) because it can determine the priority of many scoring points. However, since the priority decision is based on only affective features, it does not consider the game content sufficiently.

Table 2: Correlation analysis between the court position and the ball position. Each Pearson correlation coefficient value is an average value in 10 rally scenes of each video.

Video	Pearson correlation coefficient
1	-0.825
2	-0.802
3	-0.836
4	-0.921
5	-0.718
Average	$\mathbf{-0.820}$

3. Verification

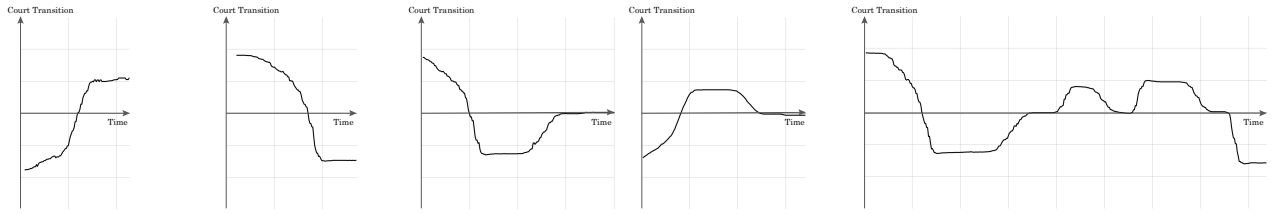
3.1. Ball-Court Correlation Analysis

We performed correlation analysis for verifying our assumption that the ball position can be approximated by the court position (Figure 4). Figure 3 shows examples of the court transition and the ball transition. Correlation analysis is tested on total 5341 frames (10 rally scenes from each of 5 volleyball videos (Table 1)). In correlation analysis, we calculate Pearson correlation coefficient between the court position and the ball position obtained manually. The Pearson correlation coefficient is a measure of the linear correlation between two variables. As a result, the absolute value of Pearson correlation coefficients is more than 0.7 in all videos, which proved that there is strong correlation (Table 2). Therefore, it can be concluded that the court position information includes the global ball position information.

3.2. Rally Scene Retrieval and Cluster Analysis

We perform rally scene retrieval and cluster analysis by using the court transition information (time series data of x-component of the court center position) in order to verify that it represents the game content. We calculate the dissimilarity D between the two rally scenes by applying dynamic time warping to the court transition data.

In rally scene retrieval, we output a rally scene with both the same scoring team and the smallest dissimilarity for the input rally scene. For determining the scoring team, we use the court position in the first frame of the next rally scene since the scoring team gains the next service right in volleyball rules. To evaluate rally scene retrieval, we use two crite-



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3 (d) Cluster 4 (e) Cluster 5

Figure 5: The representative court transitions of generated clusters.

ria that measure (1) the difference between the rally counts (how many times the ball came back and forth between the net) of the input query and those of the retrieval results and (2) the accuracy of the scoring team estimation for 120 rally scenes in video 1. Note that the court transition data is correctly acquired in these rally scenes. As a result, the average value of the rally counts difference is 0.14 times and the accuracy of the scoring team estimation is 92.6%. Therefore, the court transition information represents the game content such as the rally counts and the scoring team.

In cluster analysis, we perform a hierarchical agglomerative clustering with group average method. The hierarchical agglomerative clustering is a bottom-up clustering method. it starts with every single sample in a single cluster. Thin, in each successive iteration, it merges the closest pair until all of the data in one cluster. We perform rally cluster analysis on 120 rally scenes and generate 5 clusters. Figure 5 shows the representative court transitions of the generated clusters. The generated five clusters can be classified into three types by considering left and right symmetry: (i) only service (Figure 5a, 5b), (ii) service and one return (Figure 5c, 5d), (iii) long rally (Figure 5e). Therefore, the court transition information includes the semantics such as the rally counts, which team hits a service, and which team attacks in the end of the rally. Therefore, it can be concluded that the court transition information represents the semantics of the game.

4. Method

4.1. Court Detection and Tracking

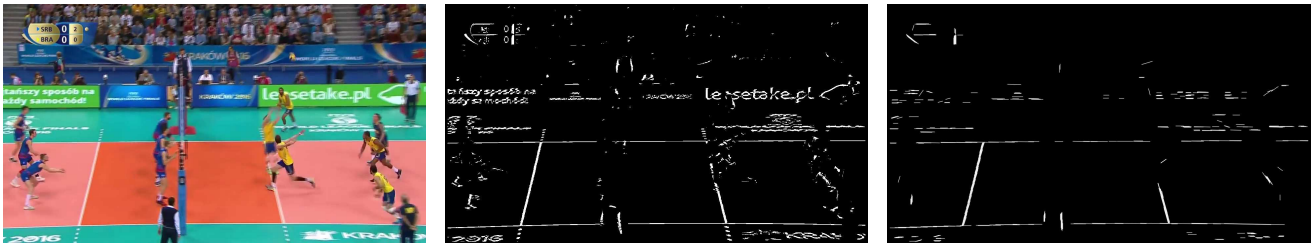
Several court detection technique has been proposed for camera calibration in sports videos [3, 5]. The task of court detection is to provide a geometric transformation that maps the points in the world coordinate to those in the image coordinate. Since the sports courts can be assumed to be planar, the mapping can be described by a 3×3 matrix H (homography). To calculate H, eight parameters have to be determined because of scaling invariant. In order to determine the eight parameters, at least four point-correspondences between positions in the world coordinate and those in the image coordinate. Here, we use the intersection points of the court-lines to compute these parameters. Our algorithm has two steps; court-line extraction and model fitting. In general, these processing is difficult because of occlusion and lens distortion. We perform a robust court-line extraction using iterative line segment detection (LSD) and a court fitting technique without correcting lens distortion by extending previous methods [3, 5].

4.1.1 Court-Lines Extraction

In general, the court consists of several white and thin lines, so we generate a binary image by two constraints; (i) a threshold σ_l of a target pixel's luminance and (ii) a threshold σ_d of the relative luminace differences between it and surround pixels. The surround pixels are either two pixels at a horizontal distance of $\pm\tau$ or at a vertical distance of $\pm\tau$. Parameter τ equals approximately the double court-line width. However, still, several noise regions such as letters in logos, the stadium, spectators and players might appear in the resulting binary image (Figure 6b). Then we apply an iterative line segment detection (LSD) into the binary image for detecting longer and continuous line-region (Figure 6c, 6d, 6e, 6f). Here, LSD is performed by probabilistic Hough transform which has three parameters; an accumulator threshold T, a minimum line length L and a maximum allowed gap G between linked points on the same line. Our system automatically increases these parameter values per LSD step. In this paper, based on our preliminary experiment, the number of iterations is set to 2 times for getting a balance between the iterative LSD results (convergence) and a computational cost. After iterative LSD, we detect lines based on a standard Hough transform with RANSAC.

4.1.2 Model Fitting

The model fitting step determines correspondences between the four detected lines and the template lines in the model. Since the correspondences are not known beforehand, we iterate assignments of two horizontal lines and two vertical lines between the image and the model. For each assignment, we compute a matching score S. The homography



(a) Input image (b) Binary image (c) First LSD ($T = 15, L = 15, G = 2$)

(d) Second LSD ($T = 45, L = 45, G = 6$) (e) Third LSD ($T = 75, L = 75, G = 10$) (f) Fourth LSD ($T = 105, L = 105, G = 14$)

Figure 6: Iterative Line Segment Detection.

matrix H that gives the maximum score S_{max} higher than a threshold is selected as the best solution. The score S is the degree of overlap between the detected court-lines and the transformed model court-lines. However, even if the intersection points are detected completely, the transformed model court-lines do not overlap court-lines in the image completely because of lens distortion. Then, we generate a score map obtained by the court-line image with Gaussian filter and the score is calculated by accumulating the score map values (luminance) on the transformed model court-line pixels as follows:

$$S = \sum_i \mathrm{ScoreMap}(\boldsymbol{p}_i) \tag{1}$$

where $\mathrm{ScoreMap}(\cdot)$ is luminance of the score map and $\boldsymbol{p}_i$ are the transformed model court-line pixels. This measurement are robust against lens distortion because this score map depends on the distance from the detected court-line pixels.

4.1.3 Court Tracking

In the subsequent frames, we predict a homography matrix for the next frame $t + 1$ as follows:

$$\hat{H}_{t+1} = H_t H_{t-1}^{-1} H_t \tag{2}$$

Next, we perform court-line extraction and model fitting considering only neighborhood pixels around model court-lines transformed by the predicted homography matrix. However, strong motion blur can occur in these frames by the camera motion. If the model fitting fails, we perform

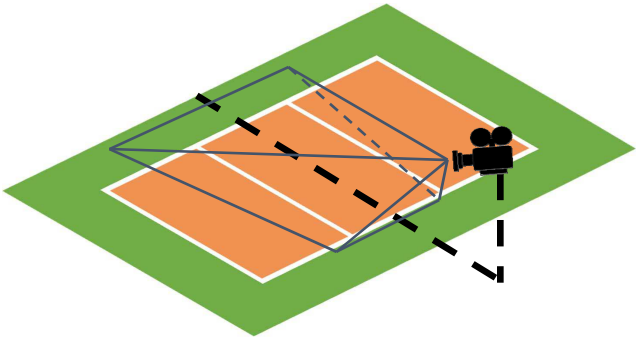


Figure 7: Assumption of camera position and orientation in volleyball videos.

non-linear minimization of a back projection error E using Levenberg Marquardt algorithm. The back projection is defined as follows:

$$E = \sum_i d(\hat{H}_{t+1}^{-1} \boldsymbol{p}_i) \tag{3}$$

where $d(\cdot)$ is distance from the nearest model court-line, $\boldsymbol{p}_i$ are back projected points near the predicted court-lines in the binary image (Figure 6b).

4.2. Rally Scene Detection

Some rally scene detection techniques in racquet sports videos have been proposed recently. Liu *et al.* [10] perform an unsupervised shot clustering based on HSV histogram and detect rally scenes based on a scene classifier using trained audio data by support vector machine. This method requires large training of audio annotation data, which are

constructed manually for the first 30 minutes of the input video. For automatic rally scene detection, Kawamura *et al*. [8] assume that rally scenes are usually obtained with a fixed camera in racquet sports. That is, the rally scenes are easily determined by whether the white-line region of the court can be detected from an average image of each shot or not. They determine rally clusters containing white-line regions as rally scenes from clusters generated by Liu's shot clustering. However, rally scenes in volleyball videos contain the camera operation such as panning, so it remains problematic to apply Kawamura's method to them directly. In contrast, we assume that "the court is taken by a specific direction in rally scenes" extended from the Kawamura's assumption. Based on our assumption, we detect shots with the court facing a specific direction as the rally scenes. In the case of volleyball videos, since the rally scene is taken from the extension of the center-line, we detect the court with its long side oriented horizontally with respect to the image (Figure 7).

Our algorithm applies court detection based hierarchical rally scene selection approach; cluster-level and shot-level. First, we segment the input video by shot [9] and generate clusters using unsupervised shot clustering based on HSV histogram [10]. From the generated clusters, we select the rally clusters where the court detection rate for the center frames of the shots is higher than a threshold R_c (in this paper, 50%). However, the rally clusters contain the non-rally scenes with similar HSV histogram to the correct rally scenes. Therefore, after rally cluster selection, we extract several frames (in this paper, 15 frames) from each shot in the rally clusters and perform court detection again. If the court detection rate of the shot is higher than a threshold R_s (in this paper, 50%), we detect the shot as a rally scene.

4.3. Video Summarization

We evaluate a rally-rank for each rally scene, and generate a summary video containing only the rally scenes with higher rally-rank. The rally scenes included in the summary video are selected in order from the rally scene with the highest rally-rank so that the summary video becomes within the viewer-specified duration. In addition to affective features (e.g., sound information and editorial information) in previous method [8, 10], the court-based features is used for rally-rank evaluation. Given rally length L_r, pitch P_r, volume V_r, total court movement distance D_r, court average speed S_r, and court maximum speed M_r, each rally-rank I_r for the r-th rally is calculated as follows:

$$I_r = \alpha L_r + \beta P_r + \gamma V_r + \delta D_r + \epsilon S_r + \zeta M_r + \eta \quad (4)$$

where $\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \eta$ are weight coefficients determined by multiple regression analysis of the subjective experimental values. Since the variance of feature values differs between videos, all feature values are normalized for each video. The subjective experiments was performed for 10 rally scenes from each of 5 volleyball videos and we asked 45 subjects to rank the interest of the rally scene at 7-point Likert scale.

5. Experimental Results

5.1. Rally Scene Detection

To verify the effectiveness of our rally detection, we used criterion, "precision," "recall," and "F-measure," as follow:

$$\text{Precision} = \frac{C}{D} \quad (5)$$

$$\text{Recall} = \frac{C}{T} \quad (6)$$

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (7)$$

where T is the actual number, D is the detected number, and C is the correctly detected number of rally scenes. In the experiment, we used 5 volleyball videos (Table 1). The ground truth of the rally scenes are manually labeled. For comparison, we used Kawamura's method and the results of the detection accuracy are shown in Table 3. These results show that our detection method provides higher accuracy than the previous method. This is because our method performs court detection instead of white-line detection and selects the rally shots from the rally clusters (hierarchical selection). Recall values are smaller than precision values since it is difficult for our method to detect rally scenes taken by a different direction from the camera orientation constraint (Figure 7).

5.2. Video Summarization

To verify that our rally-rank evaluation can reflect viewers' interest, we calculated adjusted R-squared between average subjective values and evaluated values (Table 4). The adjusted R-squared is a modified version of R-squared that is a statistical measure of how close the data to the fitted regression line. The results show that all adjusted R-squared values in our method is higher than those in Kawamura's method. Therefore, our rally-rank evaluate function demonstrates the effectiveness of satisfying viewers' interest.

Table 5 shows variance inflation factor (VIF) and regression coefficients for each independent variable. VIF quantifies how much the variance is inflated for detecting multicollinearity (how much independent variables are correlated with each other). Since all VIF values are smaller than 10.0, the correlation between the independent variables is low. Regression coefficients represent the weight coefficients in rally-rank evaluation. The absolute values of the regression coefficients mean the degree of the impact on our rally-rank. Since there is weak correlation between the independent variables and the absolute regression coefficient

Table 3: Accuracy of rally scene detection.

Video	Our method			Kawamura *et al.* [8]		
	Precision	Recall	F-measure	Precision	Recall	F-measure
1	1.000	0.863	0.926	0.948	0.877	0.811
2	0.876	0.966	0.919	0.942	0.830	0.882
3	1.000	0.903	0.949	0.957	0.903	0.929
4	0.994	0.855	0.919	0.927	0.825	0.873
5	1.000	0.698	0.822	0.515	0.269	0.353
Total	**0.974**	**0.857**	**0.907**	0.858	0.740	0.770

Table 4: Adjusted R-squared in rally-rank evaluation. (AF: Affective features, CF: Court-based features)

Video	Our method	AF [8]	CF
All	**0.544**	0.474	0.483
1	**0.935**	0.759	0.707
2	**0.735**	0.430	0.685
3	**0.702**	0.569	0.612
4	**0.896**	0.708	0.697
5	**0.976**	0.868	0.903

Table 5: Variance inflation factor (VIF) and regression coefficients (RC)

Independent Variable	VIF	RC
Rally Length L	3.79	0.531
Pitch P	1.09	0.026
Volume V	1.11	0.109
Total Movement Distance D	2.62	0.380
Average Speed S	2.77	−0.055
Maximum Speed M	1.18	−0.210

Table 6: Time Compression rate (TCR).

Video	Input [s]	All Rallies [s]	TCR [%]
1	6275	876	14.0
2	6743	1516	22.5
3	4823	1096	22.7
4	9452	1587	16.8
5	9446	1364	14.4
Average	-	-	**18.1**

values of the court-based features are large, the court-based features have the large impact in the rally-rank.

Figure 8 shows an example of our rally-rank evaluation in Video 1. Low-rank rally scenes included a lot of "service miss" and "service ace," while high-rank rally scenes have many rally counts. In this way, the rally scene with more rally counts tends to be evaluated with a higher rank. On the other hand, our rally-rank cannot distinguish "service ace" and "service miss."

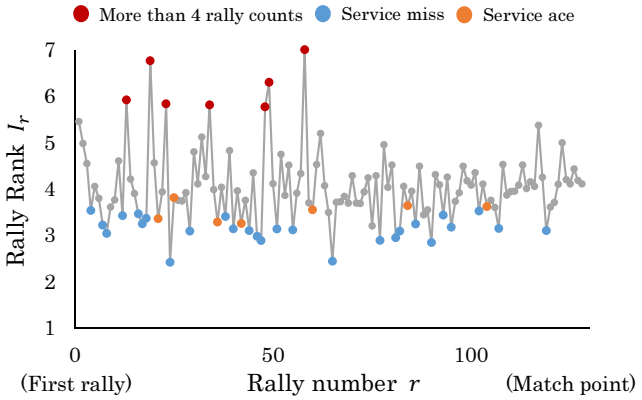


Figure 8: An example of our rally-rank evaluation.

Table 6 shows the time compression rate of summary videos containing all rally scenes against original input videos. The results indicate that viewers can watch all rally scenes within a short period of time.

6. Conclusion

We have proposed a volleyball video summarization system. Our system detects rally scenes automatically and robustly, and then introduce a rally-rank evaluation based on the court transition information. It is verified that the court transition has the strong correlation with the ball transition and represents the game content by performing correlation analysis, rally scene retrieval and cluster analysis. By using the court-based features, our video summarization made it possible to reflect viewers' preferences more.

Since our rally-rank evaluation is performed independently for each rally scene, it is difficult to consider the connection between the rallies and there is the difference between the game results and the impressions felt from the summarized video. In the future, we aim to create a summary video that consider the interaction between the rallies and is not contradictory to the game result. We will apply our court-based technique to other court sports with similar camera work by only changing the court template model.

Acknowledgement

This work was supported in part by JST ACCEL (grant JPMJAC1602).

References

[1] S. Baysal and P. Duygulu. Sentioscope: A soccer player tracking system using model field particles. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(7):1350–1362, 2016. 2

[2] H.-T. Chen, W.-J. Tsai, S.-Y. Lee, and J.-Y. Yu. Ball tracking and 3d trajectory approximation with applications to tactics analysis from single-camera volleyball sequences. *Multimedia Tools and Applications*, 60(3):641–667, 2012. 2

[3] D. Farin, S. Krabbe, W. Effelsberg, and P. H. N. de With. Robust camera calibration for sport videos using court models. In *SPIE Storage and Retrieval Methods and Applications for Multimedia*, volume 5307, pages 80–91, 2004. 4

[4] Z. Feng, D. Yuan, W. Zhe, and W. Haila. Matching logos for slow motion replay detection in broadcast sports video. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pages 1409–1412, 2012. 2

[5] J. Han, D. Farin, and P. H. N. de With. Broadcast court-net sports video analysis using fast 3-d camera modeling. *IEEE Transactions on Circuits and System for Video Technology*, 18(11):1628–1638, 2008. 4

[6] A. Hanjalic. Adaptive extraction of highlights from a sport video based on excitement modeling. *IEEE Transactions on Multimedia*, 7(6):1114–1122, 2005. 2, 3

[7] C.-C. Hsu, H.-T. Chen, C.-L. Chou, and S.-Y. Lee. 2d histogram-based player localization in boradcast volleyball videos. *Multimedia Systems*, 22(3):325–341, 2016. 2

[8] S. Kawamura, T. Fukusato, T. Hirai, and S. Morishima. Rsviewer: An efficient video viewer for racquet sports focusing on rally scenes. In *International Conference on Information Visualization Theory and Applications*, volume 2, pages 247–254, 2016. 2, 3, 6, 7

[9] S. Lian, Y. Dong, and H. Wang. Efficient temporal segmentation for sports programs with special cases. In *Advances in Multimedia Information Processing - PCM 2010*, volume 6297, pages 381–391, 2010. 6

[10] C. Liu, Q. Huang, S. Jiang, L. Xing, Q. Ye, and W. Gao. A framework for flexible summarization of racquet sports video using multiple modalities. *Computer Vision and Image Understanding*, 113(3):415–424, 2009. 2, 3, 5, 6

[11] A. Maksai, X. Wang, and P. Fua. What players do with the ball: a physically constrained interaction modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[12] H. Morimitsu, R. M. Cesar, and I. Bloch. Attributed graphs for tracking multiple objects in structured sports videos. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 34–42, 2015. 2

[13] K. Soomro, S. Khokhar, and M. Shah. Tracking when the camera looks away. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 742–750, 2015. 2

[14] J. Wang, C. Xu, E. Chng, and Q. Tian. Sports highlight detection from keyword sequences using HMM. In *IEEE International Conference on Multimedia and Expo*, pages 599–602, 2004. 2

[15] Z. Wang, J. Yu, Y. He, and T. Guan. Affection arousal based highlight extraction for soccer video. *Multimedia Tools and Applications*, 73(1):519–546, 2014. 2, 3

[16] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang. Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, volume 5, pages 632–635, 2003. 2

[17] Z. Zhao, S. Jiang, Q. Huang, and G. Zhu. Highlight summarization in sports video based on replay detection. In *IEEE International Conference on Multimedia and Expo*, pages 1613–1616, 2006. 2