User-selectable Event Summarization in Unedited Raw Soccer Video via Multimodal Bidirectional LSTM

Tomoki Haruyama (student member)†, Sho Takahashi (member)††, Takahiro Ogawa (member)†††, Miki Haseyama (fellow)†††

Abstract A new method that generates user-selectable event summaries from unedited raw soccer videos is presented in this paper. Since there are more unedited raw soccer videos than broadcasted/distributed soccer videos and unedited videos have various viewers, it is necessary to analyze these videos for meeting the demands of various viewers. The proposed method introduces a multimodal CNN-BiLSTM architecture for analyzing unedited raw soccer videos. This architecture extracts candidate scenes for event summarization from unedited soccer videos and classifies these scenes into typical events. Finally, our method generates user-selectable event summaries by simultaneously considering the importance of candidate scenes and the event classification results. Experimental results using real unedited raw soccer videos show the effectiveness of our method.

Key words: Video semantic analysis, sports videos, convolutional neural network, long short-term memory, multimodal analysis, event summarization.

1. Introduction

In recent years, various video distribution services such as Youtube* and DAZN** have become popular due to the development of network technologies and various devices1). Services specializing in sports such as DAZN have become particularly popular since sports are played worldwide and attract many viewers2). In TV broadcasting, it is possible to provide not only sports videos but also various contents related to those videos by services such as data broadcasting and Hybridcast3). Sports videos can therefore be watched in various ways, and information related to sports can now be easily obtained.

Soccer matches have many viewers all over the world. Most viewers prefer to watch matches in a condensed form of a highlight since a soccer match lasts for at least 90 minutes. Broadcast highlights are usually edited by analysts working for television stations or video distribution companies. The process of editing requires going through the whole match, selecting scenes that are important and then making a highlight of the selected scenes, which is a very time-consuming process4). Many researchers have developed technologies for soccer video summarization5)-15) without time-consuming human input. The targets of these methods are broadcasted/distributed soccer videos, and analysis is performed by using the unique characteristics of edited videos. Specifically, these methods perform various processes including shot boundary detection, event detection and summarization of soccer videos by effectively using editing features such as a scoreboard, logo marks, dissolves and replay scenes.

On the other hand, an entire view of a match has generally been recorded as shown in Fig. 1 for tactical analysis and scouting of players in various teams. Videos that have recorded in this way are defined as unedited raw soccer videos in this paper. Such soccer videos are not limited to official matches that are broadcasted/distributed but also include practice matches and many other matches. In addition, un-
like broadcasted/distributed soccer videos owned only by the top teams, many teams own unedited raw soccer videos. Therefore, a technique for generation of summaries from unedited raw soccer videos is necessary for many teams. However, since these videos are unedited and have no editing features, accurate generation of a summary based on existing methods for broadcasted/distributed soccer videos is difficult. Specifically, since unedited raw soccer videos have no shot boundaries, even the pre-processing used in existing methods cannot be performed. Therefore, it is necessary to establish a method for generation of summaries from unedited raw soccer videos.

Since unedited raw soccer videos especially have various types of viewers and since summaries are utilized for various purposes, it is necessary to generate summaries that meet the demands of various viewers. For example, some viewers want to watch a summary composed of one type of event briefly, and whereas other viewers want to watch a summary composed of various events sufficiently. Therefore, a user-selectable framework in which the user can select the desired events and the time scale of a summary they wish for is needed.

In this paper, we propose a novel method that generates user-selectable event summaries from unedited raw soccer videos by utilizing bidirectional long short-term memory (BiLSTM). We call extracting events from an original video and summarizing the video by aggregating extracted events as “event summarization”. The proposed method generates a summary that contains the events the user wants and the length the user wants from a match video by a multimodal convolutional neural network (CNN)-BiLSTM architecture. In this architecture, we first calculate multiple CNN-based features from audio-visual components representing the entire view of soccer video clips, movements of players and cheering by the spectators of unedited raw soccer videos. Furthermore, we apply BiLSTM to these CNN-based features to consider bidirectional time series of unedited raw soccer videos. BiLSTM is a model that bidirectionally trains long short-term memory (LSTM) networks, and it is often used for natural language tasks and video recognition tasks since it can capture not only forward but also backward contexts. In soccer, it is also important to capture not only forward but also backward contexts. For example, when a counter-attack starts, it is difficult to tell if it will end with a shot or if the counter-attack will be stopped by a foul. In soccer matches, there are many situations that cannot be judged in a unidirectional context as in the above situations, and BiLSTM is therefore essential for accurate analysis of soccer videos. Finally, by integrating the outputs of BiLSTMs applied to multiple CNN-based features, high-performance analysis of unedited raw soccer videos is realized. By multimodally using heterogeneous features, each modality can compensate for the shortcomings of the other modalities. In the proposed method, by utilizing this multimodal CNN-BiLSTM architecture, the generation of user-selectable event summary is performed. Specifically, we first extract candidate scenes for event summarization from unedited raw soccer videos by utilizing a multimodal CNN-BiLSTM architecture tuned for importance calculation. Our framework uniquely determines the level of importance independent of the user’s intentions, taking into account that the cost of retraining the model for each user is very high and the importance is unique to some degree from a tactical point of view. Next, we classify the candidate scenes into typical events (i.e., shot, corner kick, free kick and foul) by utilizing another multimodal CNN-BiLSTM architecture tuned for event classification. Finally, a user selects the desired events included in the summary and the time scale of the summary, and our method generates an event summary that can adapt to such user selections by simultaneously considering the importance of candidate scenes and the event classification results. Consequently, the proposed method can generate user-selectable event summaries from unedited raw soccer videos.

2. User-selectable Event Summarization via Multimodal BiLSTM Architecture

In this section, we present the details of the proposed method. As shown in Fig. 2, our method consists of three steps, extraction of candidate scenes for event summarization, event classification of candidate scenes and generation of event summary. In the first step, we calculate the importance of video clips divided evenly with overlaps from an unedited raw soccer video by using a multimodal CNN-BiLSTM architecture tuned for importance calculation. Then we extract candidate scenes for event summarization based on the calculated importance. In the second step, event classification of the candidate scenes is performed by using a multimodal CNN-BiLSTM architecture tuned for event classification. Finally, in the third step, we generate a user-selectable event summary by assembling classified
candidate scenes and adapting them to user selections. The details of these three steps are shown in the following subsections: extraction of candidate scenes for event summarization (Section 2.1), event classification of candidate scenes (Section 2.2) and event summarization (Section 2.3).

2.1 First Step: Extraction of Candidate Scenes for Event Summarization

We divide the target unedited raw soccer video that has no shot boundary into \( T \) second video clips at one-second intervals by a sliding window scheme. Then we calculate the importance of these divided video clips by a multimodal CNN-BiLSTM architecture as shown in Fig. 3. Specifically, in this architecture, we first extract multiple features from audio-visual components of input video clips by a single CNN model. Furthermore, we calculate the importance by applying multiple BiLSTMs tuned for importance calculation to the calculated heterogeneous features to take the time series of soccer video clips into account. Finally, candidate scenes for event summarization can be extracted on the basis of the calculated importance.

(1) Feature Extraction

The proposed method realizes feature extraction from multimodal data via a single CNN-based architecture as shown in Fig. 4. Generally, it is difficult to optimize multiple CNN architectures from a limited number of training samples. This is particularly difficult for unedited raw soccer videos since the environments of soccer video clips such as stadium conditions are quite different, and unedited raw soccer videos of other teams are not available for the opposing team. Therefore, in our method, we utilize a pre-trained CNN architecture for solving the above problem. In order to obtain CNN-based features via the single CNN architecture, we have to obtain images from all modalities. In the following, we show the details of the images of each modality.

Images for representing the entire view of soccer video clips

The proposed method simply uses image data at the \( t(=1,2,...,T) \)-th frame. Note that \( T \) is the time length of each video clip. Simple image data include various types of information such as the location of the playing field and players. Therefore, we directly use the image data.

Images for representing movements of players

Semantic information cannot be sufficiently perceived by only using simple image data. However, features
embedding positional information of players are desired since such information is very important for understanding the video. Hence, to extract player regions from image data, we apply the YOLOv3 model\(^{22}\), which is an object detection model pre-trained by the Microsoft Common Objects in Context (MSCOCO) dataset\(^{23}\), to image data. Then we detect regions of “person” as player regions and generate their mask images.

Images for representing cheering by spectators
Since the spectators cheer loudly in important situations, analysis of audio components of soccer videos is effective. The proposed method divides each video clip into \(T\) equal-length segments. Then, for each segment, the image representing cheering by the spectators is obtained by calculating its spectrogram.

In this way, the proposed method obtains three kinds of images for each segment. Note that the lengths of time between video frames and the duration of audio sequences are determined in such a way that they become the same. The proposed method inputs these images into the VGG16 model\(^{24}\) that is pre-trained with the ImageNet dataset\(^{25}\). We obtain the outputs of its final pooling layer as \(D\)-dimensional CNN-based features. Then, by aligning these features for all \(T\) segments, the three kinds of CNN-based features representing the entire view of soccer video clips, movements of players and cheering by the spectators can be obtained as visual features \(X^{\text{visual}} \in \mathbb{R}^{DT}\), player features \(X^{\text{player}} \in \mathbb{R}^{DT}\) and audio features \(X^{\text{audio}} \in \mathbb{R}^{DT}\), respectively. Note that calculation of spectrogram-based features from a pre-trained CNN model is known to be an effective representation of audio data\(^{26,27}\), and its effectiveness in the soccer video analysis has been reported\(^{28}\). By using the above approach, our method realizes feature extraction from multimodal data via the single CNN-based architecture.

(2) Multimodal Output Calculation
We calculate the importance of the divided video clips by using multiple BiLSTMs tuned for importance calculation as shown in Fig. 3. Note that the “Output” is calculated as the “Importance” of the target clips in the first step (Section 2.1). In the same way, the “Output” is calculated as the “Event” of the target clip in the second step (Section 2.2).

BiLSTM is an extended version of the LSTM network\(^{17}\) that can capture the time series. Compared to traditional recurrent neural networks (RNNs), an
The LSTM network consists of a single LSTM block from multiple hidden units as shown in Fig. 5. Every memory block consists of one self-connected linear memory cells and three multiplicative gates. An input gate $i$, a forget gate $f$ and an output gate $o$ plays the role of writing, reading and resetting the memory cell values $c$, respectively. Given an input $x_t$ at time $t$, activations of the input gate $i_t$, forget gate $f_t$, memory cell state $c_t$ and output gate $o_t$ are respectively updated by the following equations:

$$
i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i),$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f),$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c),$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o),$$

$$h_t = o_t \cdot \tanh(c_t),$$

where $\sigma$ denotes the logistic sigmoid, $W$ is the weight matrix of the mutual connections, $h_t$ presents the output of the hidden block, and $b$ indicates the block bias. From the above equations, it is observed that the values of all memory cells and block outputs in the previous time $t-1$ will certainly affect the activations of all three gates, even the input units in the current time $t$ in the same layer, except in the case between the memory cell and output gate. The main advantage of using such memory-enhanced blocks in RNNs over traditional units is that the LSTM block is to total activation over time. Since the derivative function is distributed over the sum, there is no vanishing gradient problem where the backpropagation error blows up or decays over time.

BiLSTM has a forward LSTM network and a backward LSTM network constructed from such LSTM blocks as shown in Fig. 6, and it is often used for natural language tasks and video recognition tasks because it can capture not only forward but also backward contexts. In soccer, it is also important to capture not only forward but also backward contexts. For example, when a counter-attack starts, it is difficult to tell if it will end with a shot or if the counter-attack will be stopped by a foul. In soccer matches, there are many situations that cannot be judged in a unidirectional context as in the above situations, and BiLSTM is therefore essential for accurate analysis of soccer videos. Specifically, in the proposed method, as shown in Fig. 6, when each feature $X^m (m \in \{\text{visual}, \text{player}, \text{audio}\})$ is input to a forward LSTM network and a backward LSTM
network, the output $h^m$ is calculated by combining the output of a forward LSTM network $\hat{h}_T^m$ and a backward LSTM network $\hat{h}_1^m$ as follows:

$$h^m = [\hat{h}_T^m, \hat{h}_1^m]^	op. \quad (6)$$

The multiple BiLSTMs are tuned to classify whether an input video clip is an important scene or not. Then, by inputting the output $h^m$ to a softmax function, $p^m_k$, the probability that an input video clip is an important scene or not is calculated as follows:

$$p^m_k = \frac{\exp(z_k^m \cdot h^m + b_k^m)}{\sum_{k'} \exp(z_k^m \cdot h^m + b_k^m)}. \quad (7)$$

Note that $z_k^m$ and $b_k^m$ are the weight vector and the bias corresponding to the importance or no importance for each feature, respectively, and $k \in \{\text{important}, \text{non-important}\}$. $k'$ is the index used to compute $\Sigma$ and is distinguished to avoid mixing it with $k$. Also, the proposed method calculates $p_k$ by integrating $p_k^m$ as the following formula so that multiple aspects can be considered.

$$p_k = \sum_m \alpha^m p_k^m \quad \text{s.t.} \quad \sum_m \alpha^m = 1, \quad (8)$$

where $\alpha^m$ are the parameters that balance the percentages of each modality. In the proposed method, $p_{\text{important}}$, the probability that an input video clip is an important scene, is defined as the importance of the video clip, that is, “Output” in Fig. 3 is regarded as “Importance”. Finally, parts in which $p_{\text{important}}$ continuously exceeds a threshold $\tau$ for time length $T'$ or more are regarded as candidate scenes for event summarization. Note that $T'$ is a parameter that plays a role as the time threshold for accurately extracting important scenes.

### 2.2 Second Step: Event Classification of Candidate Scenes for event summarization

We perform event classification of the candidate scenes obtained in the previous step (Section 2.1). Specifically, the proposed method classifies the candidate scenes into typical events (i.e., shot, corner kick, free kick, and foul) by a multimodal CNN-BiLSTM architecture shown in Fig. 3. In the second step, the multiple BiLSTMs are tuned to classify what events an input video clip includes. Multiple CNN-based features are calculated from the candidate scenes, and we input these features to tuned multiple BiLSTMs for event classification to calculate the probability that the candidate scene includes event $l \in \{\text{shot, corner kick, free kick, foul}\}$ as $p_l^m$. As you can see, “Output” in Fig. 3 is regarded as “Event” in the second step. Finally, we estimate the event of the candidate scene $\hat{l}$ as follows:

$$\hat{l} = \arg \max \sum_m p_l^m. \quad (9)$$

### 2.3 Third Step: Generation of Event Summarization

The proposed method generates a event summary that adapts to the user selections. Specifically, the user first selects the event(s) and the time scale of the summary they wish for. Then our method generates a summary by assembling only user-selected types of event(s) from the candidate scenes in descending order of the calculated importance until the total time reaches the length of the time scale determined by the user. Consequently, our method can generate a user-selectable event summary from an unedited raw soccer video.

### 3. Experimental Results

We describe experiments using actual unedited raw soccer videos in this section. Since it is difficult to quantitatively evaluate user-selectable event summarizations, we first quantitatively evaluate the importance calculation (Section 3.1) and event classification (Section 3.2), which are the main shafts of our method. After these evaluations, we qualitatively evaluate the user-selectable event summarizations in a subject experiment (Section 3.3).

#### 3.1 Performance Evaluation of Importance Calculation

We evaluate the performance of the multimodal CNN-BiLSTM architecture for importance calculation in the first step (Section 2.1) by monitoring the important scene classification. Specifically, the multimodal CNN-BiLSTM architecture classifies video clips in a dataset (Dataset1) into important scenes and non-important scenes. Dataset1 consists of the 373 video
Table 1 shows the experimental results. In this table, "PM" and "CM" are abbreviations for "proposed method" and "comparative method", respectively. CM1-CM4 are methods based on multimodal soccer analysis for similar scene retrieval by simply distance scale (SDS), and these methods do not take the time series into account. CM5-CM8 are methods based on the previous paper that uses support vector machine (SVM), and these methods do not take the time series into account. CM9-CM12 are methods based on forward LSTM that take the forward time series into account. CM13-CM16 are methods based on backward LSTM that take the backward time series into account. CM17-CM19 and PM are methods based on BiLSTM that take the bidirectional time series into account. Furthermore, CM4, CM8, CM12, CM16 and PM are methods that integrate multiple features to take into account multiple aspects of unedited raw soccer videos. In this experiment, the number of hidden layers of LSTM and BiLSTM was two. We used 80% of the dataset as a training set and the remaining 20% as a test set and conducted 5-fold cross-validation. Also, Adam with a learning rate of 0.001 was used for cost minimization. The batch size was 512, and the number of epochs was 50. We determined $\alpha_m$ by performing a grid search using the training set. Note that each $\alpha_m$ was tried in a total of 11 stages from 0.0 to 1.0 at an interval of 0.1. We evaluated the performance by F-measure, which is the harmonic mean of Recall and Precision.
we used Micro-F1 and Macro-F1, which are generally used for multi-label classification problems, as evaluation indexes of the average F-measure. Specifically, Micro-F1 is the F-measure for all scenes, and Macro-F1 is the average of the F-measures for each class. First, by comparing the results obtained by a single feature (CM1–CM3, CM5–CM7, CM8–CM11, CM13–CM15 and CM17–CM19) and the results obtained by multiple features (CM4, CM8, CM12, CM16 and PM), we confirm the effectiveness of multimodally using heterogeneous features obtained from unedited raw soccer videos. Furthermore, by comparing the results obtained by SDS, SVM, f-LSTM, b-LSTM and BiLSTM, we confirm the effectiveness of considering the time series, especially the bidirectional time series. Also, it can be seen that the proposed method outperforms CM5–CM8, methods based on the previous paper \(^{30}\). On the other hand, the progress of the proposed method from the previous paper itself \(^{30}\) is the introduction of player features that can take tactics into account for summarization, rather than just simple audio-visual features, and the introduction of a framework that allows for time series analysis of those features. As a result, the proposed method not only improved the accuracy from the previous paper itself \(^{30}\), but also allowed us to train scenes of different time lengths. Consequently, since the proposed method outperforms other comparative methods, we confirm the effectiveness of the multimodal CNN-BiLSTM architecture for importance calculation. Furthermore, the results suggested that this level of accuracy is sufficient for extracting candidate scenes. Also, in the proposed method, \(\alpha^{visual} = 0.4\), \(\alpha^{player} = 0.2\), \(\alpha^{audio} = 0.4\) were set, respectively. These parameters show that the cheering by spectators also contributes just as significantly to important scene detection as the visual components of soccer videos. This is due to the fact that the spectators get excited about important scenes.

### 3.2 Performance Evaluation of Event Classification

We evaluate the performance of the multimodal CNN-BiLSTM architecture for event classification in the second step (Section 2.2). Specifically, the multimodal CNN-BiLSTM architecture classifies video clips in another dataset (Dataset2) into four typical events (i.e., shot, corner kick, free kick and foul). Dataset2 consists of 958 video clips obtained from unedited raw soccer videos of 33 matches in MEIJI YASUDA J1 LEAGUE. The numbers of video clips including shot, corner kick, free kick and foul are 427, 173, 111 and 247, respectively. Each of these clips has a unique length of time, ranging from 5 to 68 seconds, and the total duration is about 2 hours and 23 minutes. The average time length of these clips are about 8.926 seconds. These scenes were labeled on the basis of match reports published on J.LEAGUE.jp\(^*\). In order to confirm the effectiveness of the multimodal CNN-BiLSTM architecture for event classification, we used the same comparative methods and experimental conditions as those described in Section 3.1, with the exception of CM5–CM8, a method for important scene detection.

Table 2 shows the experimental results. By comparing the experimental results in the same way as in Section 3.1, we confirm the effectiveness of the multimodal CNN-BiLSTM architecture for event classification. Also, in the proposed method, \(\alpha^{visual} = 0.4\), \(\alpha^{player} = 0.4\), \(\alpha^{audio} = 0.2\) were set, respectively. These parameters make it clear that it is important to focus on the movements of players for event classification. This is because the movements of players are somewhat specific to each event.

### 3.3 Performance Evaluation of User-selectable Event Summarization

We performed a subject experiment to evaluate the proposed method. In this experiment, we generated user-selectable event summaries from actual unedited raw soccer videos of 2 matches in MEIJI YASUDA J1 LEAGUE. First, 10 subjects (2 females and 8 males, aged 22–25 years) selected the event(s) and time scale of the summary they wished for. Then our method generated 22–25 years) selected the event(s) and time scale of the summary they wished for. Then our method generated user-selectable event summaries according to their selections. In this experiment, the parameters were experimentally set as \(T = 10.0\), \(\tau = 2.2\) and \(T' = 8.0\).

Also, the experimental procedure is as follows. Subjects first watch the target match. Next, subjects choose the events and the length of time they want for the summary. In addition, subjects watch the summary generated by the proposed method based on the user’s selections and the broadcast highlight (BH) made available to the public by DAZN. Finally, we ask the following two questions to each of the summary by the proposed method and BH.

**Q1:** Please rate the validity of a summary according to 7 grades (1–7).

**Q2:** Please rate whether a summary is adapted to your selections according to 7 grades (1–7).

\(^*\) https://www.jleague.jp/
Table 2: Details of each method and F-measures of event classification results of the comparative methods (CM1–CM11) and the proposed method (PM).

| Methods | Algorithm | Features | Event classification | Average |
|---------|-----------|----------|----------------------|---------|
|         |           | $X^{\text{visual}}$ | $X^{\text{player}}$ | $X^{\text{audio}}$ | Shot | Corner kick | Free kick | Foul | Micro-F1 | Macro-F1 |
| CM1     | SDS       | ✓         | ✓                    | ✓        | 0.67 | 0.43 | 0.33 | 0.32 | 0.55 | 0.44 |
| CM2     | SDS       | ✓         | ✓                    | ✓        | 0.69 | 0.61 | 0.29 | 0.07 | 0.56 | 0.42 |
| CM3     | SDS       | ✓         | ✓                    | ✓        | 0.65 | 0.25 | 0.09 | 0.04 | 0.48 | 0.26 |
| CM4     | SDS       | ✓         | ✓                    | ✓        | 0.68 | 0.54 | 0.37 | 0.27 | 0.57 | 0.47 |
| CM5     | f-LSTM    | ✓         | ✓                    | ✓        | 0.55 | 0.55 | 0.47 | 0.45 | 0.54 | 0.51 |
| CM6     | f-LSTM    | ✓         | ✓                    | ✓        | 0.58 | 0.73 | 0.44 | 0.54 | 0.60 | 0.57 |
| CM7     | f-LSTM    | ✓         | ✓                    | ✓        | 0.42 | 0.22 | 0.11 | 0.31 | 0.40 | 0.27 |
| CM8     | f-LSTM    | ✓         | ✓                    | ✓        | 0.65 | 0.71 | 0.45 | 0.53 | 0.63 | 0.59 |
| CM9     | b-LSTM    | ✓         | ✓                    | ✓        | 0.35 | 0.46 | 0.40 | 0.56 | 0.48 | 0.44 |
| CM10    | b-LSTM    | ✓         | ✓                    | ✓        | 0.58 | 0.17 | 0.11 | 0.42 | 0.44 | 0.32 |
| CM11    | b-LSTM    | ✓         | ✓                    | ✓        | 0.57 | 0.75 | 0.40 | 0.51 | 0.57 | 0.56 |
| CM12    | b-LSTM    | ✓         | ✓                    | ✓        | 0.62 | 0.68 | 0.48 | 0.56 | 0.59 | 0.58 |
| CM13    | BiLSTM    | ✓         | ✓                    | ✓        | 0.74 | 0.83 | 0.69 | 0.63 | 0.72 | 0.72 |
| CM14    | BiLSTM    | ✓         | ✓                    | ✓        | 0.75 | 0.80 | 0.66 | 0.57 | 0.71 | 0.70 |
| CM15    | BiLSTM    | ✓         | ✓                    | ✓        | 0.66 | 0.40 | 0.24 | 0.50 | 0.54 | 0.45 |
| PM      | BiLSTM    | ✓         | ✓                    | ✓        | 0.81 | 0.87 | 0.75 | 0.68 | 0.79 | 0.78 |

Q1 can monitor the quality of the summary, and Q2 can monitor the reflection of the user’s selections. In this experiment, grade 4 indicated an acceptable rating and grade 7 indicated an ideal rating.

Table 3 shows the user selections and their ratings. In this experiment, among the 10 subjects (S1–S10), S1–S5 and S6–S10 evaluated the first match and the second match, respectively. The average ratings of Q1 and Q2 for PM were 4.6 and 4.9, respectively. Almost all of the ratings are equal to or greater than the acceptable rating. Although it is not as good as the rating of Q1 for BH, which can be defined as an ideal summary, the difference is about 1 point, which is a sufficient value. As for Q2, BH is below the acceptable rating, but PM is much higher. From the above results, we confirm the effectiveness of the proposed method. Examples of results of the experiment are shown in Fig. 7. In this figure, “PM” and “GT” represent “proposed method” and “ground truth”, respectively. GT represents the correct class of events included in the summary generated by PM. Also, the parts surrounded by black lines in GT are areas included in the actual broadcast highlight video by DAZN. In these examples, since many of the parts in the PM overlap with those in the GT, the effectiveness of importance calculation (Section 2.1) in the proposed method was confirmed. In addition, since many of the event classification results are correct, event classification (Section 2.2) in the proposed method was confirmed to be effective. However, the comparison between S1 and S2 suggests that a smaller number of user-selected events results in lower ratings for Q2. This is because the mistaken event classification of candidate scenes greatly affected generated summaries when the number of user-selected events was smaller. Therefore, it is necessary to improve the accuracy of event classification to solve this problem. In order to improve the accuracy of event classification, the use of player position data, which will be commonly captured by various teams in the future, would be effective. Also, the results for S7 and S10 suggest that the evaluation of Q1 is lowered when the time length of the summary requested by the user increases. This is probably because relatively non-important scenes are included in the generated summary in such a situation. Since the number of important scenes varies from match to match, we believe that it is useful to show the user the length of time necessary to ensure a certain level of importance. These issues will be our future work.

4. Conclusions

In this paper, we have proposed a new method that generates user-selectable event summaries from
unedited raw soccer videos. The proposed method introduces a multimodal CNN-BiLSTM architecture for analyzing unedited raw soccer videos. This architecture extracts candidate scenes for event summarization and performs event classification with a high level of accuracy and finally enables the generation of user-selectable event summaries. Experimental results using real unedited raw soccer videos showed the effectiveness of the proposed method. A method to show the user the length of time needed to ensure a certain level of importance, and a method that changes the importance for each user are our future works.

Acknowledgment

This work was partly supported by the MIC/SCOPE #181601001.

References

1) L. Zhou : “Mobile device-to-device video distribution: Theory and application,” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 12, no. 3, article. 38, pp.1–23, (2016)
2) B. Hutchins, B. Li and D. Rowe : “Over-the-top sport: Live streaming services, changing coverage markets and the growth of media sport portals,” Media, Culture & Society, vol. 41, no. 7, pp. 975–994, (2019)
3) H. Ohmata, M. Takechi, S. Mituya, K. Otsuki, A. Baba, K. Matsumura, K. Majima and S. Sunasaki : “Hybridcast: A new media experience by integration of broadcasting and broadband,” IEEE ITU Kaleidoscope: Building Sustainable Communities, pp. 1–8, (2013)
4) J. Hannon, K. McCarthy, J. Lynch and B. Smyth : “Personalized and automatic social summarization of events in video,” ACM International Conference on Intelligent User Interfaces (IUI), pp. 335–338, (2011)
5) A. Ekin, A. M. Takalp and R. Mehrotra : “Automatic soccer video analysis and summarization,” IEEE Transactions on Image Processing, vol. 12, no. 7, pp. 796–807, (2003)
6) M. Y. Eldih, B. S. A. Zaid, H. M. Zawbaa, M. El-Zahar and M. El-Saban : “Soccer video summarization using enhanced log detection,” IEEE International Conference on Image Processing (ICIP), pp. 4345–4348, (2009)
7) F. Chen, C. De Vleeschower, H. Duxans Barrohés, J. Gregorio Escalada and D. Conejero : “Automatic summarization of audio-visual soccer feeds,” IEEE International Conference on Multimedia and Expo (ICME), pp. 837–842, (2010)
8) H. M. Zawbaa, N. El-Bendary, A. E. Hassanien and T.-h. Kim : “Machine learning-based soccer video summarization system,” IEEE International Conference on Multimediain Computer Graphics, and Broadcasting (MCCGB), pp. 19–28, (2011)
9) A. Raventos, R. Quijada, L. Torres and F. Tarrés : “Automatic summarization of soccer highlights using audio-visual descriptors,” SpringerPlus, vol. 4, no. 1, article. 301, pp.1–19, (2015)
10) M. H. Kolekar and S. Sengupta : “Bayesian network-based customized highlight generation for broadcast soccer videos,” IEEE Transactions on Broadcasting, vol. 61, no. 2, pp. 195–209 (2015)
11) H. Jiang, Y. Lu and J. Xue : “Automatic soccer video event detection based on a deep neural network combined cnn and rnn,” IEEE International Conference on Tools with Artificial Intelligence (ICTAI), pp. 490–494, (2016)
12) K. Blanc, D. Lingrand and F. Precioso : “Singlets: Multi-resolution motion singularities for soccer video abstraction,” IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 10–19, (2017)
13) T. Liu, Y. Lu, X. Lei, L. Zhang, H. Wang, W. Huang and Z. Wang : “Soccer video event detection using 3D convolutional networks and shot boundary detection via deep feature distance,” International Conference on Neural Information Processing (NeurIPS), pp. 440–449, (2017)
14) K. Tang, Y. Bao, Z. Zhao, L. Zhu, Y. Lin and Y. Peng : “AutoHighlight: Automatic highlights detection and segmentation in soccer matches,” IEEE International Conference on Big Data (Big Data), pp. 4619–4624, (2018)
15) B. Fakhar, H. R. Kanan, A. Behrad : “Event detection in soccer videos using unsupervised learning of Spatio-temporal features based on pooled spatial pyramid model,” Multimedia Tools and Applications, vol. 78, no. 12, pp. 10995–17025, (2019)
16) A. Graves, A. Mohamed and G. Hinton : “Speech recognition with deep recurrent neural networks,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649, (2013)
17) S. Hochreiter and J. Schmidhuber : “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, (1997)
18) J. P. Chiu and E. Nichols : “Named entity recognition with bidirectional LSTMs-CCNs,” Transactions of the Association for Computational Linguistics, vol. 4, pp. 357–370, (2016)
19) T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, V. H. Albuquerque : “Cloud-assisted multiview video summarization using CNN and bidirectional LSTM,” IEEE Transactions on Industrial Informatics, vol. 16, no. 1, pp. 77–86, (2019)
20) R. Alzaidy, C. Caragea and C. L. Giles : “Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents,” The World Wide Web Conference, pp. 2551–2557, (2019)
21) C. Li, Z. Bao, L. Li and Z. Zhao : “Exploring temporal represen-
Fig. 7 Some examples of results of the experiment using subjects. In this figure, “PM” and “GT” represent “proposed method” and “ground truth”, respectively. GT represents the correct class of events included in the summarization by PM. The parts surrounded by black lines in GT are areas included in the actual highlight video broadcast by DAZN.

22) J. Redmon and A. Farhadi: Yolov3: “An incremental improvement,” arXiv preprint arXiv:1804.02767, (2018)
23) T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. DDrill and C. L. Zintnick: “Microsoft coco: Common objects in context,” European Conference on Computer Vision
24) K. Simonyan and A. Zisserman : “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, (2014)

25) A. Krizhevsky, I. Sutskever and G. Hinton : “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, pp. 1097–1105, (2012)

26) E. Al Hadhrami, M. Al Mufti, B. Taha and N. Werghi : “Ground moving radar targets classification based on spectrogram images using convolutional neural networks,” IEEE International Radar Symposium (IRS), pp. 1–9, (2018)

27) S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird and B. Schuller : “Snore sound classification using image-based deep spectrum features,” INTERSPEECH, vol. 434, pp. 3512–3516, (2017)

28) T. Haruyama, S. Takahashi, T. Ogawa and M. Haseyama : “Retrieval of similar scenes based on multimodal distance metric learning in soccer videos,” ACM International Workshop on Multimedia Content Analysis in Sports (MMSports), pp. 10–15, (2019)

29) T. Haruyama, S. Takahashi, T. Ogawa and M. Haseyama : Multimodal retrieval of similar soccer videos based on optimal combination of multiple distance measures,” IEEE Global Conference on Consumer Electronics (GCCE), pp. 691–692, (2019)

30) T. Haruyama, S. Takahashi, T. Ogawa and M. Haseyama : Multimodal important scene detection in far-view soccer videos based on single deep neural architecture,” IET Transactions on Media Technology and Applications, vol. 8, no. 2, pp. 89–99, (2020)

31) C. Cortes and V. Vapnik : Support-vector networks,” Machine Learning, vol. 20, no.3, pp. 273–297, (1995)

32) D. P. Kingma and J. Ba : Adam: “A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, (2014)

33) C. Hsu, C. Chang, C. Lin and et al. : “A practical guide to support vector classification,” Technical Report, Department of Computer Science, National Taiwan University, (2003)

Tomoki Haruyama received the B.S. degree in Electronics and Information Engineering from Hokkaido University, Japan in 2019. He is currently pursuing a M.S. degree with the Graduate School of Information Science and Technology, Hokkaido University. His research interests include sports video analysis. He is a student member of the IEEE.

Sho Takahashi received his B.S., M.S. and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2008, 2010 and 2013, respectively. He joined the Graduate School of Information Science and Technology, Hokkaido University, as an Assistant Professor in 2013. He was an Associate Professor of the Education and Research Center for Mathematical and Data Science, Hokkaido University from 2017 to 2018. He was a visiting researcher of Media Integration and Communication Center, University of Florence from 2018 to 2019. He is currently an Associate Professor in the Faculty of Engineering, Hokkaido University. His research interests include semantic analysis and visualization in various data and its application. He is a member of the IEEE, IEICE, Institute of Image Information and Television Engineers (ITE), Japan Society of Civil Engineering (JSCE), Society of Automotive Engineering of Japan (JSAE), Japan Society of Traffic Engineers (JSTE) and Japan Association for Human and Environmental Symbiosis (JAHES).

Takahiro Ogawa received his B.S., M.S. and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2003, 2005 and 2007, respectively. He joined Graduate School of Information Science and Technology, Hokkaido University in 2008. He is currently an associate professor in the Faculty of Information Science and Technology, Hokkaido University. His research interests are AI, IoT and big data analysis for multimedia signal processing and its applications. He was a special session chair of IEEE ISCE2009, a Doctoral Symposium Chair of ACM ICMR2018, an organized session chair of IEEE GCCE2017-2019, a TPC Vice Chair of IEEE GCCE2018, a Conference Chair of IEEE GCCE2019, etc. He has been also an Associate Editor of ITE Transactions on Media Technology and Applications. He is a senior member of IEEE and a member of ACM, IEICE and ITE.

Miki Haseyama received her B.S., M.S. and Ph.D. degrees in Electronics from Hokkaido University, Japan in 1986, 1988 and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University as an associate professor in 1994. She was a visiting associate professor of Washington University, USA from 1995 to 1996. She is currently a professor in the Faculty of Information Science and Technology Division of Media and Network Technologies, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She has been a Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE), an Editor-in-Chief of ITE Transactions on Media Technology and Applications, a Director, International Coordination and Publicity of The Institute of Electronics, Information and Communication Engineers (IEICE). She is a member of the IEEE, IEICE, Institute of Image Information and Television Engineers (ITE) and Acoustical Society of Japan (ASJ).