 Copyright © 2020 by ITE Transactions on Media Technology and Applications (MTA)

Paper

Multimodal Important Scene Detection in Far-view Soccer Videos Based on Single Deep Neural Architecture

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

Abstract The details of the matches of soccer can be estimated from visual and audio sequences, and they correspond to the occurrence of important scenes. Therefore, the use of these sequences is suitable for important scene detection. In this paper, a new multimodal method for important scene detection from visual and audio sequences in far-view soccer videos based on a single deep neural architecture is presented. A unique point of our method is that multiple classifiers can be realized by a single deep neural architecture that includes a Convolutional Neural Network-based feature extractor and a Support Vector Machine-based classifier. This approach provides a solution to the problem of not being able to simultaneously optimize different multiple deep neural architectures from a small amount of training data. Then we monitor confidence measures output from this architecture for the multimodal data and enable their integration to obtain the final classification result.

Key words: Semantic video analysis, sports video, deep learning, convolutional neural network, support vector machine.

1. Introduction

Opportunities to watch various channels via various devices have been increasing to the development of network technologies and the spread of various video distribution services [1, 2]. Among the various channels, the demand for distribution of sports videos is high, and there are channels such as DAZN* that specializes in sports. DAZN provides various contents such as live videos of matches, their details and highlight videos, and these contents are used not only by viewers but also by training staff and players. These contents are generally obtained manually by training staff, and the procedure is very time-consuming. Therefore, techniques for sports video analysis are required for automatically obtaining these contents.

Many studies on sports video analysis have been conducted in the past decade [3]. Studies have been conducted on various sports including baseball, soccer, basketball and volleyball. Since soccer is one of the most popular sports worldwide, there is a strong demand for soccer video analysis techniques. Studies have been conducted on soccer player tracking [4–8], tactics analysis [9–14], event detection [15–23] and video summarization [24–28]. A soccer match takes about 100 minutes. With the spread of video distribution services, the number of matches that can be watched has increased, and the demand for soccer video summarization is thus increasing. The main purpose of video summarization is to create a synopsis of a video that is easy to understand and give viewers a general overview of the video [29]. By summarizing the video, viewers can know the important scenes of the video [30].

Accurate detection of important scenes such as goals, shots, and fouls is needed for realizing the summarization of soccer videos. Methods for detecting events such as goal scenes, shot scenes and foul scenes have been proposed [15–23]. These methods have been proposed for broadcasting/distribution of soccer videos, and detection of events by these methods is performed by using unique characteristics of the videos. The methods make effective use of editorial characteristics such as video shot changing, zooming and dissolving. The number of far-view soccer videos, which broadly capture soccer fields and are unedited videos before broadcasting/distribution, is larger than the number of broad-

Received September 27, 2019; Revised February 5, 2020; Accepted March 5, 2020

† Graduate School of Information Science and Technology, Hokkaido University
(Kita-14 Nishi-9 Kita-ku Sapporo, 060-0814, Japan)
†† Faculty of Engineering, Hokkaido University
(Kita-13 Nishi-8 Kita-ku Sapporo, 060-8628, Japan)
††† Faculty of Information Science and Technology, Hokkaido University
(Kita-14 Nishi-9 Kita-ku Sapporo, 060-0814, Japan)

* https://www.dazn.com/

casted/distributed soccer videos. Not only professional soccer teams but also amateur soccer teams take far-view soccer videos for analyzing their matches. Therefore, if there is a method that can detect important scenes from far-view soccer videos, it can be broadly applied to not only professional soccer matches but also amateur soccer matches taken with various cameras. However, existing methods such as those in [15–23] cannot detect important scenes in far-view soccer videos since these methods need to use editorial characteristics.

In order to successfully detect important scenes from far-view soccer videos that do not have editorial characteristics, the introduction of multimodal deep neural architectures into visual and audio data of the videos and their integration should be effective. By multimodally using heterogeneous features, each modality can compensate for the shortcomings of the other modals. However, it may be difficult to simultaneously optimize different multiple deep neural architectures for heterogeneous and a limited amount of data.

In this paper, we propose a new multimodal method for the detection of important scenes. In soccer matches, the importance of scenes can be estimated from visual sequences including movements of players. The importance of scenes can also be estimated from audio sequences such as cheering by the spectators and sounds of the whistle blown by the referee. In important scenes of soccer matches, the referee blows a whistle to stop or restart the match. For example, important events such as foul scenes and set-piece scenes almost include whistle sounds. Thus, the improvement of the important scene detection performance can be expected by detecting the whistle sounds. Therefore, we extract multiple visual and audio deep neural features representing movements of players, cheering by the spectators and whistle sounds from far-view soccer videos to construct classifiers for detecting important scenes. In the proposed method, these multiple classifiers can be realized by a single deep neural architecture that includes a Convolutional Neural Network (CNN)-based feature extractor and a Support Vector Machine (SVM) [31]-based classifier. In this approach, by focusing on a report [32] showing that high-performance classification is feasible by using spectrogram images obtained from audio sequences, we create these images from cheering by the spectators and whistle sounds. Therefore, in all modalities, CNN-based features can be calculated from the images. Specifically, we extract these CNN-based features by using the pre-trained Inception-v3 model [33]. This approach provides a solution to the problem of not being able to simultaneously optimize different multiple deep neural architectures from a small amount of training data, and it is the main contribution of this paper. By using the pre-trained model, we do not have to perform retraining for each environment such as each stadium from a large amount of training data. Then, from these multimodal features, we construct SVM-based classifiers since SVM is a powerful classification method for a limited amount of data. Therefore, the combination of CNN and SVM is considered to be suitable for the detection of important scenes in far-view soccer videos. In this approach, we introduce a criterion called a "confidence measure" [34] for estimating the confidence of each classification result. Then, by monitoring these measures of different modalities, we can simply integrate their results to obtain the final detection result. By realizing this approach, various important scenes can be accurately extracted from far-view soccer videos. It should be noted that this paper is an extended version of [35].

The rest of this paper is organized as follows. In Sect. 2, we present a new method for detecting important scenes in far-view soccer videos. In Sect. 3, experimental results that verify the performance of the proposed method are presented, and Sect. 4 shows concluding remarks.

2. Multimodal Important Scene Detection via Single Deep Neural Architecture

Our method consists of two phases. Figure 1 shows an outline of the proposed method. In the first phase, we extract multiple deep neural features based on the single CNN-based architecture and enable SVM-based classification of important scenes. In the second phase, we derive confidence measures for the classification results to realize integration of the classification results obtained from the heterogeneous features. The details of the two phases are respectively shown in the following subsections.

2. 1 First Phase : Classification Based on Single Deep Neural Architecture

In the first phase, the proposed method first extracts multiple CNN-based features from visual and audio sequences of target soccer videos and then realizes SVM-based classification. In order to obtain CNN-based fea-

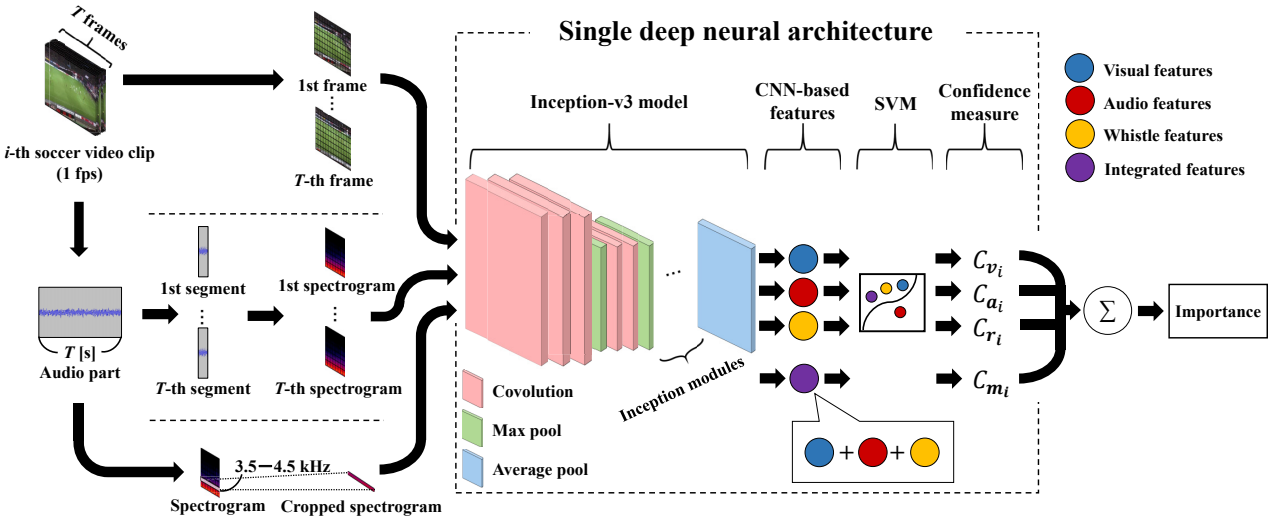


Fig. 1 Outline of the proposed method. The proposed method extracts CNN-based features from visual and audio sequences based on a single CNN architecture and enables SVM-based classification (First Phase). Then, by deriving confidence measures from the outputs of the classifiers, the proposed method enables their integration to obtain the final detection result (Second Phase).

tures via the single CNN architecture, we have to obtain images from all modalities. Since the proposed method focuses on movements of players, cheering by the spectators and sounds of the whistle blown by the referee, three kinds of images have to be calculated. In the following, we show the details of their calculations, and an overview of the calculations is shown in Fig. 2. 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.

Images for representing movements of players

The proposed method simply uses image data at the $f(=1,2,...,T)$-th frame in the $i(=1,2,...,N)$-th video clip. T and N are the time length of each video clip and the total number of video clips, respectively. Although the movements of players may be accurately analyzed by extracting players from each image, this procedure includes errors and affects the following feature extraction procedures. Therefore, we directly use the image data.

Images for representing cheering by the spectators

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.

Images for representing whistle sounds

It is known that the frequency band of sounds of the whistle blown by the referee is in the range of 3.5–4.5 kHz [36]. Thus, when a whistle is blown, it is considered that there is an energy peak in the frequency band of 3.5–4.5 kHz. Therefore, we clip the frequency band of 3.5–4.5 kHz from the spectrogram of each video clip to obtain the image representing whistle sounds by the referee. Note that since only the existence/non-existence of whistle sounds is important, we calculate the spectrogram not for each segment but for each video clip.

In this way, the proposed method obtains the three kinds of images for each segment or each video clip. The proposed method inputs these images into the Inception-v3 model [33] that is pre-trained with ImageNet dataset [37] and outputs their D-dimensional CNN-based features from its final pooling layer. Then, by aligning these features for all T segments included in each i-th video clip, the three kinds of CNN-based features representing movements of players, cheering by the spectators and sounds of the whistle blown by the referee can be obtained as visual features $\boldsymbol{v}_i \in \mathbb{R}^{DT}$, audio features $\boldsymbol{a}_i \in \mathbb{R}^{DT}$ and whistle features $\boldsymbol{r}_i \in \mathbb{R}^{D}$, respectively. Note that $\boldsymbol{r}_i$ is calculated from each video clip and its dimension becomes D.

By using the above approach, the proposed method realizes feature extraction from multimodal data via the single CNN-based architecture. Generally, it is difficult to optimize multiple deep neural architectures from a limited number of training samples. This is particularly difficult for soccer videos since the environments of soccer video clips such as stadium conditions are quite different. Therefore, in our method, we utilize

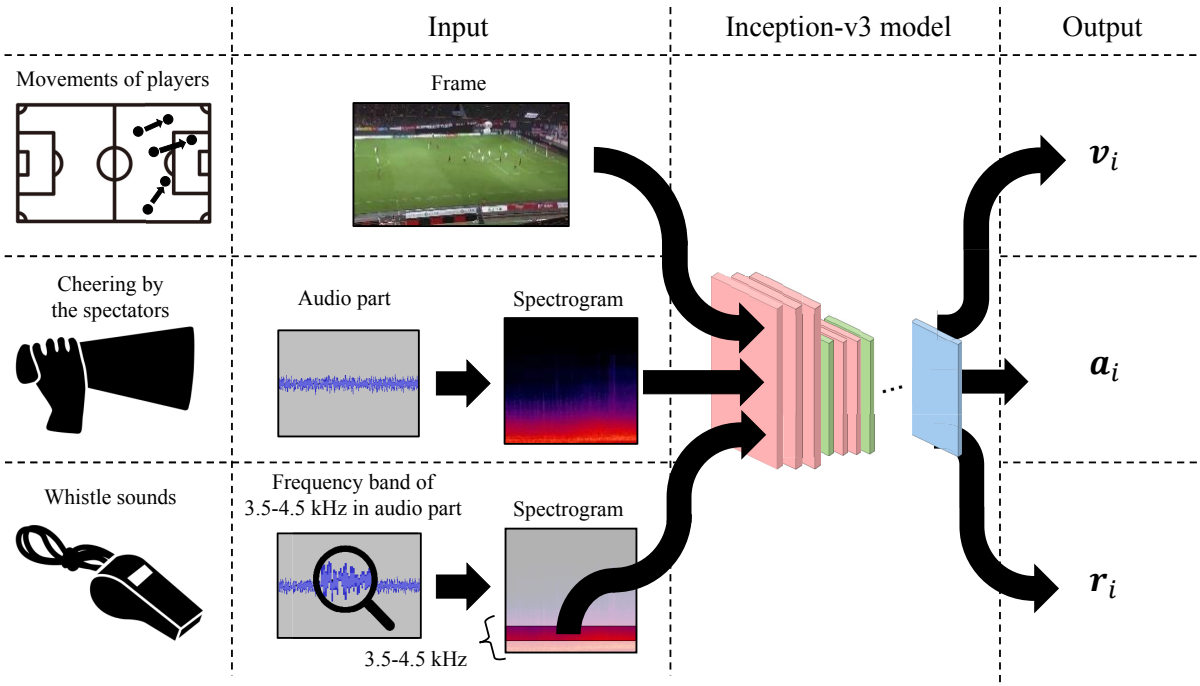


Fig. 2 Overview of the extraction of features representing movements of players ($\boldsymbol{v}_i$), cheering by the spectators ($\boldsymbol{a}_i$) and sounds of the whistle blown by the referee ($\boldsymbol{r}_i$) in the proposed method.

the pre-trained CNN architecture for effectively solving the above problem.

Next, from the obtained CNN-based features, the proposed method performs SVM-based classification of important scenes. In addition to visual features $\boldsymbol{v}_i$, audio features $\boldsymbol{a}_i$ and whistle features $\boldsymbol{r}_i$, we define integrated features $\boldsymbol{m}_i$ as multimodal features by combining visual features and audio features based on a report [38] showing that accuracy is improved by combining deep neural features obtained from multiple modalities. In the proposed method, we simply construct SVM classifiers for these four kinds of CNN-based features. In the following, we construct classifiers based on each deep neural feature obtained above. Specifically, we construct classifiers that can classify whether the target video clip is an important scene or not by using visual features $\boldsymbol{v}_i$, audio features $\boldsymbol{a}_i$ and integrated features $\boldsymbol{m}_i$. We also construct another classifier that can classify whether or not the target video clip is a scene in which the whistle is blown by the referee by using whistle features $\boldsymbol{r}_i$. First, we explain the method for construction of a classifier based on visual features $\boldsymbol{v}_i$. SVM-based training is performed by using visual features $\boldsymbol{v}_i$ calculated from video clip samples with their labels y_i that indicate whether the video clip samples are important scenes or not. By applying the generated SVM-based classifier to the target video clip for which the label is unknown, the target video clip is classified an important scene or not. Specifically, the label of the target video clip y_{target} is estimated by using visual features of the target video clip $\boldsymbol{v}_{\text{target}}$ and the classification function f^{svm} is defined as follows:

$$f^{\text{svm}}(\boldsymbol{v}_{\text{target}}) = \text{sign}(\boldsymbol{w}^\top \cdot \boldsymbol{v}_{\text{target}} + b), \tag{1}$$

$$y_{\text{target}} = \begin{cases} +1 & (f^{\text{svm}}(\boldsymbol{v}_{\text{target}}) \geqq 0) \\ -1 & (f^{\text{svm}}(\boldsymbol{v}_{\text{target}}) < 0) \end{cases}, \tag{2}$$

where $\boldsymbol{w}$ and b are a connection weight vector and a bias, respectively. When the margin in f^{svm} is κ, the training is performed to maximize the margin of the classification hyperplane $|\boldsymbol{w}^\top \cdot \boldsymbol{v}_i + b| \geqq \kappa$. Also, when $\boldsymbol{w}$ and b are normalized by κ, the hyperplane satisfies $|\boldsymbol{w}^\top \cdot \boldsymbol{v}_i + b| \geqq 1$, and $\boldsymbol{w}^\top \cdot \boldsymbol{v}_i + b$ satisfies

$$\boldsymbol{w}^\top \cdot \boldsymbol{v}_i + b \begin{cases} \geqq +1 & (y_i = +1) \\ \leqq -1 & (y_i = -1) \end{cases}. \tag{3}$$

From Eq. (3), $y_i(\boldsymbol{w}^\top \cdot \boldsymbol{v}_i + b) \geqq 1$ is established. When $\boldsymbol{v}_{\text{target}}$ is projected onto the $\boldsymbol{w}^\top$ axis, the Euclidean distance between that point and the origin is defined as $\boldsymbol{w}^\top \cdot \boldsymbol{v}_{\text{target}}/||\boldsymbol{w}||$. Therefore, the distance between the points of both labels closest to the hyperplane is $2/||\boldsymbol{w}||$. The problem of maximizing $2/||\boldsymbol{w}||$ results in the following problem:

$$\begin{aligned} &\arg\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w}||^2 \\ &\text{s.t.} \quad y_i(\boldsymbol{w}^\top \cdot \boldsymbol{v}_i + b) \geqq 1, \quad i = 1, \cdots, N. \end{aligned} \tag{4}$$

Also, by defining the Lagrange multipliers $\boldsymbol{a} = (a_1, \cdots, a_N)^\top$ and using the Lagrange multiplier method, Eq. (4) reduces to the problem of maximizing the following objective function for $\boldsymbol{a}$:

$$\tilde{L}(\boldsymbol{a}) = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i,j=1}^{N} a_i a_j y_i y_j {\boldsymbol{v}_i}^\top \boldsymbol{v}_j, \tag{5}$$

where $\boldsymbol{a}$ satisfies the following equation:

$$\begin{aligned} &\sum_{i=1}^{N} a_i y_i = 0 \\ &\text{s.t.} \quad a_i \leqq 0, \quad i = 1, \cdots, N. \end{aligned} \tag{6}$$

The classification hyperplane $f^{\text{svm}}(\boldsymbol{v}_{\text{target}})$ is expressed by the following equation.

$$f^{\text{svm}}(\boldsymbol{v}_{\text{target}}) = \sum_{i=1}^{N} a_i y_i \boldsymbol{v}_i \cdot \boldsymbol{v}_{\text{target}} + b. \tag{7}$$

Furthermore, $\boldsymbol{v}_i$ is transformed by the non-linear mapping $\psi(\boldsymbol{v}_i)$, and by applying the SVM on the space, non-linear classification is possible by the following equation.

$$f^{\text{svm}}(\boldsymbol{v}_{\text{target}}) = \sum_{i=1}^{N} a_i y_i K^{\text{svm}}(\boldsymbol{v}_i, \boldsymbol{v}_{\text{target}}) + b, \tag{8}$$

where we define a kernel function $K^{\text{svm}}(\boldsymbol{v}_i, \boldsymbol{v}_{\text{target}}) = \psi(\boldsymbol{v}_i) \cdot \psi(\boldsymbol{v}_{\text{target}})$. The output of SVM defined by Eq. (1) is calculated on the basis of $\psi(\boldsymbol{v}_{\text{target}})$ and the distance of the classification hyperplane. As described above, visual features are used to classify target video clips for which labels are unknown as important scenes. As well as the above processing, SVM is used to learn audio features and integrated features, and each feature is used to classify whether the target video clip for which the label is unknown is an important scene or not in the same manner.

Next, the whistle feature is used to perform training by SVM to classify whether the target video clip for which the label is unknown is a scene in which the a whistle is blown or not.

2.2 Second Phase : Multimodal Fusion Based on Confidence Measures

In the second phase, we perform multimodal fusion of the classification results obtained in the first phase based on their confidence measures to obtain the final classification result. In the following, we explain the derivation of the confidence measure of the visual feature-based classification result in the target video clip $C_{\boldsymbol{v}_{\text{target}}}$ as an example. Given the attribute probabilities of being an important scene and a non-important scene obtained as $P^{\text{pos}}_{\boldsymbol{v}_{\text{target}}}$ and $P^{\text{neg}}_{\boldsymbol{v}_{\text{target}}}$, respectively, the confidence measure of the visual feature-based classification result is derived as

$$C_{\boldsymbol{v}_{\text{target}}} = P^{\text{pos}}_{\boldsymbol{v}_{\text{target}}} - P^{\text{neg}}_{\boldsymbol{v}_{\text{target}}}. \tag{9}$$

Note that $P^{\text{pos}}_{\boldsymbol{v}_{\text{target}}}$ and $P^{\text{neg}}_{\boldsymbol{v}_{\text{target}}}$ are calculated by utilizing the output value of Eq. (8). Specifically, the sigmoid function is used as an activation function for $f^{\text{svm}}(\boldsymbol{v}_{\text{target}})$ according to [39]. When c is a random variable with class values to be classified and $t_{\text{target}} = (y_{\text{target}} + 1)/2 \in \{1, 0\}$, the attribute probability when $t_{\text{target}} = 1$ is calculated as follows:

$$\begin{aligned} P(c = 1 | f^{\text{svm}}(\boldsymbol{v}_{\text{target}})) &= \frac{1}{1 + \exp(\alpha f^{\text{svm}}(\boldsymbol{v}_{\text{target}}) + \beta)} \\ &= p_{\text{target}}, \end{aligned} \tag{10}$$

where the likelihood function L for maximum likelihood estimation of α and β is expressed by the following equation:

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{N} P(c = t_i | \boldsymbol{v}_i; \alpha, \beta) \\ &= \prod_{i=1}^{N} p_i^{t_i} (1 - p_i)^{1 - t_i}. \end{aligned} \tag{11}$$

Next, we calculate the negative log likelihood, that is, the cross entropy error function E, using Eq. (11) as follows:

$$\begin{aligned} E(\alpha, \beta) &= -\ln L(\alpha, \beta) \\ &= -\sum_{i=1}^{N} \{t_i \ln p_i + (1 - t_i) \ln (1 - p_i)\}. \end{aligned} \tag{12}$$

When α and β that minimize E are calculated by the gradient descent method, the class attribute probability can be calculated from Eq. (10). From the above, it becomes possible to convert the result of the binary classification problem based on $f^{\text{svm}}(\boldsymbol{v}_{\text{target}})$ in (-∞,∞) into the attribute probability in [0, 1], and it becomes possible to introduce the confidence measure $C_{\boldsymbol{v}_{\text{target}}}$. Furthermore, the proposed method standardizes these confidence measures $C_{\boldsymbol{v}_{\text{target}}}$ in such a way that their average and variance become 0 and 1, respectively, and $C^{\text{sta}}_{\boldsymbol{v}_{\text{target}}}$ is obtained. The same procedures are performed for the other features to obtain $C^{\text{sta}}_{\boldsymbol{a}_{\text{target}}}$, $C^{\text{sta}}_{\boldsymbol{r}_{\text{target}}}$ and $C^{\text{sta}}_{\boldsymbol{m}_{\text{target}}}$. Finally, the proposed method performs multimodal fusion of these results by a simple sum of

confidence measures to obtain C^{all} as follows:

$$C^{\mathrm{all}} = C^{\mathrm{sta}}_{\boldsymbol{v}_{\mathrm{target}}} + C^{\mathrm{sta}}_{\boldsymbol{a}_{\mathrm{target}}} + C^{\mathrm{sta}}_{\boldsymbol{m}_{\mathrm{target}}} + C^{\mathrm{sta}}_{\boldsymbol{r}_{\mathrm{target}}}. \quad (13)$$

If C^{all} is more than a pre-defined threshold τ, we regard the target video clip as an important scene, and if C^{all} is not more than τ, we regard it as a non-important scene.

3. Experiments

The experiments we performed are described in this section. We conducted two types of experiments: a quantitative experiment and a qualitative experiment. In the experiments, we detected the important scenes using far-view soccer videos. In Subsect. 3.1, we present the dataset that we used in the experiments, and we present the experimental results in Subsect. 3.2.

3.1 Dataset

(1) Quantitative Experiment

In this subsection, we describe the dataset used in the quantitative experiment. When we constructed SVM-based classifiers using visual features $\boldsymbol{v}_i$, audio features $\boldsymbol{a}_i$ and integrated features $\boldsymbol{m}_i$, we used a dataset (Dataset1) consisting of 373 video clips obtained from far-view videos of 6 soccer matches, with the duration of each video clip being 10 seconds. Table 1 shows the important and non-important scenes in Dataset1. Among 373 video clips, 193 video clips were obtained as important scenes, which were scenes replayed by sports broadcasting engineer in broadcasted/distributed videos by SKY PerfecTV!* or DAZN, and the remaining 180 video clips were obtained as non-important scenes by randomly selecting 30 video clips from each match. Also, when we constructed an SVM-based classifier using whistle features $\boldsymbol{r}_i$, we used a dataset (Dataset2) consisting of 360 video clips obtained from the same six far-view soccer videos, with duration of each video clip being 10 seconds. Table 2 shows scenes that include sounds of the whistle blown by the referee (whistle scenes) and scenes that do not include sounds of the whistle blown by the referee (non-whistle scenes) in Dataset 2. We manually extract 30 whistle scenes and 30 non-whistle scenes from each match. Furthermore, we divided each video clip into 10 segments, *i.e.*, 10 video frames and 10 audio segments were used for important scene detection of each scene.

(2) Qualitative Experiment

In this section, we describe the dataset used in the qualitative experiment. We used all of the video clips in Dataset1 and Dataset2 as training data. Specifically, when we constructed SVM-based classifiers using visual features $\boldsymbol{v}_i$, audio features $\boldsymbol{a}_i$ and integrated features $\boldsymbol{m}_i$, we used Dataset1, and we constructed an SVM-based classifier using whistle features $\boldsymbol{r}_i$, we used Dataset2. In addition, we used the first half of three other matches (about 45 to 50 minutes each) as test data.

Table 1 Important and non-important scenes in Dataset1.

Match	Important	Non-important
1	28	30
2	30	30
3	35	30
4	37	30
5	27	30
6	36	30
Sum	193	180

Table 2 Whistle and non-whistle scenes in Dataset2.

Match	Whistle	Non-whistle
1	30	30
2	30	30
3	30	30
4	30	30
5	30	30
6	30	30
Sum	180	180

3.2 Experimental results

(1) Quantitative Experiment

In this subsection, we present the results of the quantitative experiment. In this experiment, we conducted 5-fold cross-validation, a leave-match-out scheme, *i.e.*, we used video clips of 5 matches as training data and used video clips of the other match as test data for every validation. Also, the kernel function used in SVM was the Gaussian kernel, the threshold used in the second phase was set as $\tau = 0.4$ experimentally, and the other parameters were determined by a grid search [40].

In order to confirm the effectiveness of the proposed method, we used the comparative methods shown in Table 3. In this table, PM and CM represent "proposed method" and "comparative method", respectively. CM1–CM4 are methods using only visual features $\boldsymbol{v}_i$, audio features $\boldsymbol{a}_i$, whistle features $\boldsymbol{r}_i$ and integrated features $\boldsymbol{m}_i$, respectively. In other words, CM1–CM4 respectively performed SVM-based classification using each feature in the first phase. CM5 is

* https://www.skyperfectv.co.jp/

Table 3 Features used in PM and CM1–CM4.

	PM	CM1	CM2	CM3	CM4
$\boldsymbol{v}_i$	✓	✓			
$\boldsymbol{a}_i$	✓		✓		
$\boldsymbol{r}_i$	✓			✓	
$\boldsymbol{m}_i$	✓				✓

Table 4 Accuracy of the important scene detection results.

Match	PM	CM1	CM2	CM3	CM4	CM5	CM6
1	**0.97**	0.88	0.95	0.59	0.95	**0.97**	0.78
2	**0.95**	0.83	0.93	0.65	0.93	0.93	0.60
3	**0.92**	0.82	0.85	0.66	0.88	0.88	0.78
4	**0.90**	0.76	0.81	0.66	0.81	0.81	0.70
5	**0.91**	**0.91**	0.88	0.58	0.88	0.88	0.84
6	**0.91**	0.79	0.77	0.80	0.88	0.88	0.68
Average	**0.93**	0.83	0.87	0.66	0.89	0.89	0.73



Fig. 3 Examples of important scenes detected by the proposed method from far-view soccer videos.

our previously reported method [35], and the main improvement in the proposed method is the introduction of whistle features. Furthermore, we introduce CM6, a method based on a CNN model that is pre-trained with VGG16 [41]. In CM6, we input each frame of a soccer video clip to CNN and classified directly whether a target video clip was an important scene or not from its output layer. We evaluated the proposed method and the comparative methods by accuracy defined as follows:

$$\text{Accuracy} = \frac{\text{Num. of correctly detected video clips}}{\text{Num. of all video clips}}. \tag{14}$$

Table 4 shows the experimental results, and Fig. 3 shows examples of important scene detection by our method. As shown in this table, the proposed method outperforms the other comparative methods. From these experimental results, we can see that the detection performance obtained by each modality (*e.g.*, CM1 and CM2) tends to output sufficient results. Therefore, the introduction of the single deep neural architecture provides sufficient performance even though it is commonly applied to different modalities. Next, since the performance of our method is better than the performance of CM1–CM4, multimodal fusion based on confidence measures is effective for improving detection performance. Also, since the performance of our method is better than that of CM5, the introduction of whistle features is effective for improving detection performance. Specifically, our method can detect more important scenes such as foul scenes where the referee blows a whistle and presenting a card and set-piece scenes where the referee blows a whistle before the kick than CM5 can. Furthermore, since the detection performance of our method is better than that of CM6, we can verify that the proposed method enables accurate detection from only a limited number of training data. From the results of the experiment, we quantitatively confirmed the effectiveness of our method.

(2) Qualitative Experiment

In this subsection, we present the results of the qualitative experiment. Figure 4 shows the method used in the qualitative experiment. Video clips were extracted at equal intervals from the input video every second, *i.e.*, video clips were extracted from the start time of each of the far-view soccer videos until the end time with a slide width of one second, and C^{all} was calculated in each video clip by using the proposed method. Parts in which C^{all} continuously exceeded threshold τ for ten or more video clips were regarded as important scenes. As in the quantitative experiment, the threshold was experimentally set as $\tau = 0.4$.

Figure 5 shows the results of important scene detection by our method and scenes included in a broadcasted highlight video of the match. Note the broadcast highlight video is not a video of replay scenes broadcasted during the match but a video broadcasted after the match as a summary of the match. As shown in Fig. 5, our method could detect important scenes such

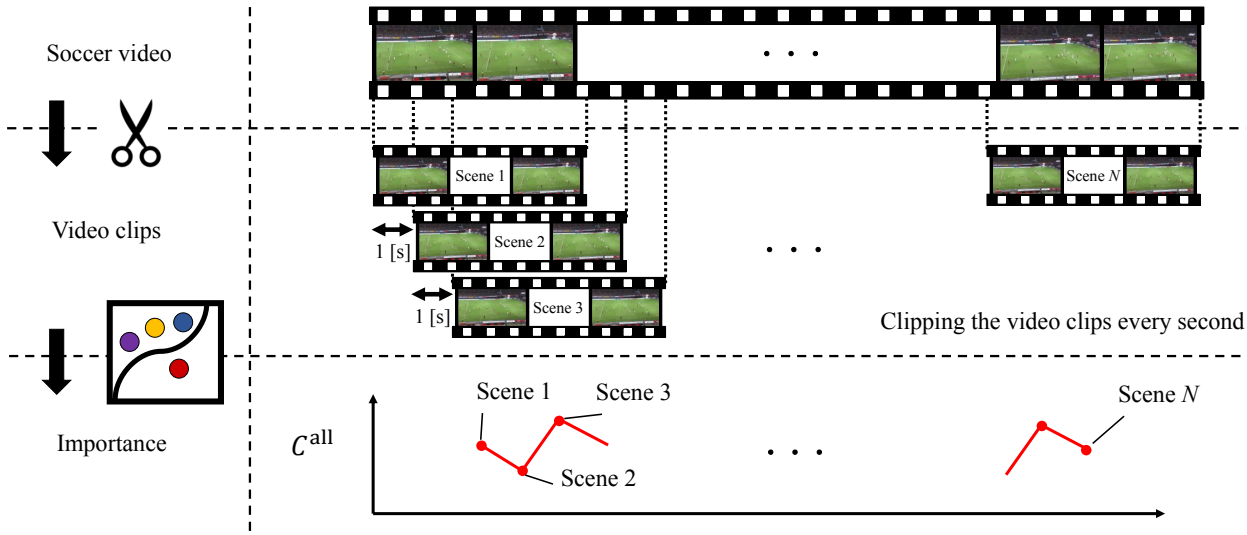


Fig. 4 Method used in the qualitative experiment. First, we divided a far-view soccer video into video clips, and then we calculated C^{all} for each video clip.

as goal scenes, shot scenes, foul scenes, set-piece scenes and other scenes, and those scenes almost matched the broadcast highlight scenes included in the broadcast highlight video. Only one of the 17 scenes included in the highlight videos was not detected by our method. However, it is true that our method detects more scenes than broadcast highlight scenes. It is difficult to include all important scenes in the broadcast highlight video due to the limitation of broadcast time. On the other hand, since our method does not consider time limitation, it detected more scenes. Specifically, in the detected scenes by the proposed method other than the broadcast highlight scenes, there are shot scenes with a counter-attack, free kick scenes unfortunately missed, foul scenes with a furious tackle and quarrel scenes. Thus, the detected scenes by the proposed method other than the broadcast highlight scenes are considered enough to be important scenes and candidate scenes of the broadcast highlight video. Therefore, it is also possible to reduce watching time by about 10-times for the TV editor by using scenes detected by our method as candidate scenes for highlight videos since the length of the detected scenes is 10-times smaller than the total length of the video. From the above results, the effectiveness of the proposed method was confirmed qualitatively.

4. Conclusions

In this paper, we have proposed a new method that can detect important scenes in far-view soccer videos. In the proposed method, multimodal deep neural features and their integrated features are extracted on the basis of a single CNN architecture. Then we construct classifiers that can classify whether a soccer video clip is an important scene or not by using each feature via SVM. Furthermore, we introduce a criterion called "confidence measure" for estimating the confidence of each classification result. By utilizing the confidence measure for each classifier and integrating the confidence measures, a method that can detect important scenes in far-view soccer videos is realized. The effectiveness of the proposed method was verified from experimental results. We have solved the problem of not being be able to simultaneously optimize different multiple deep neural architectures from a small amount of training data.

In future work, it is necessary to develop a method for important scene detection that takes into consideration the background of the soccer match and information on the players. Specifically, it is possible to obtain external information that provides consistent semantic-rich descriptions, which audiovisual signals lack, from information published on the Internet. The realization of further semantically understanding of important scenes detected from a far-view soccer video are expected by combining the information with the proposed method. Furthermore, for the reason that it uses features that represent the movement of players, the cheering from the spectators and the whistle sounds using a single architecture, it is considered that the proposed method is effective for sports, especially for soccer. In the same way, the proposed method is applicable to general video

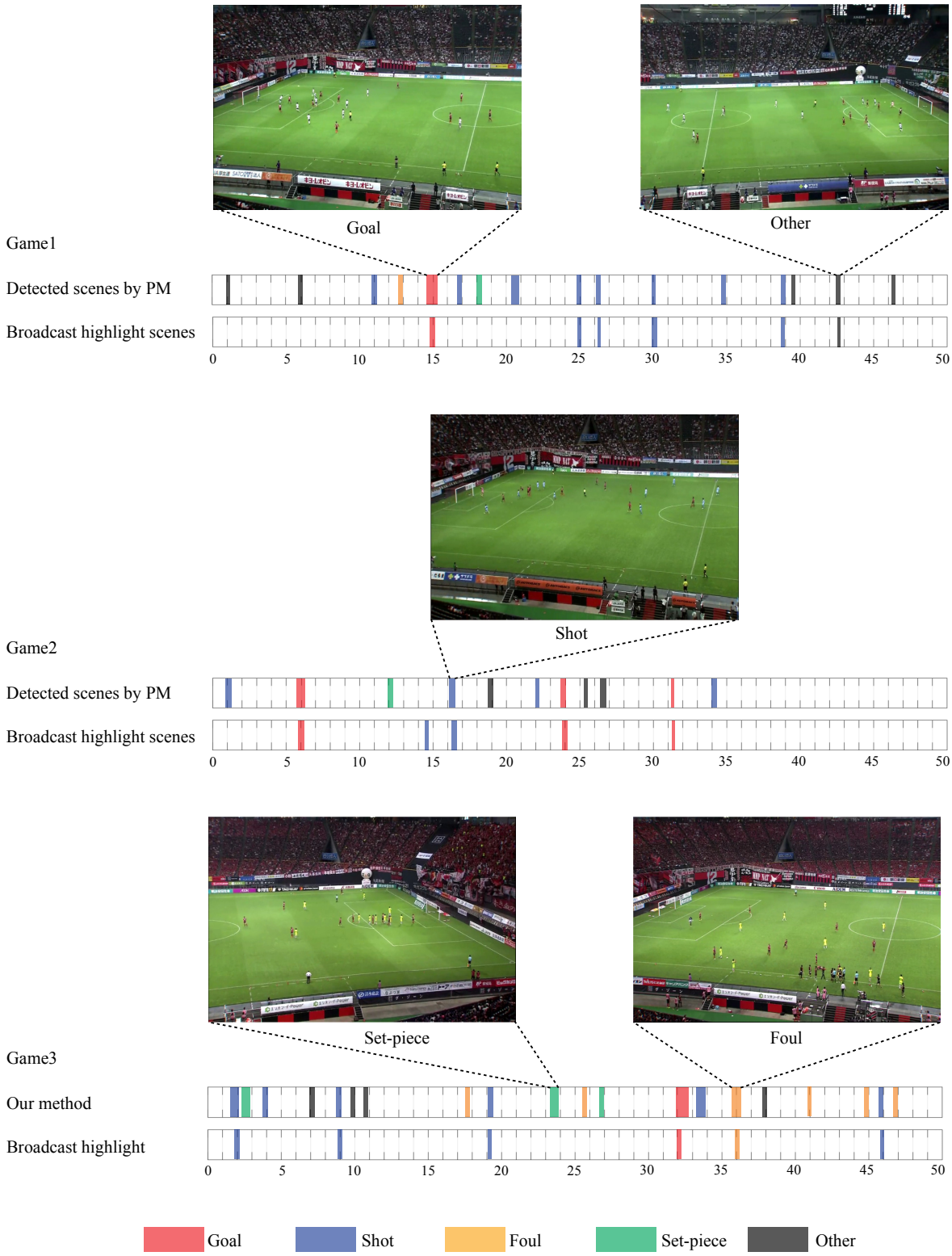


Fig. 5 Results of the qualitative experiment and examples of scenes detected by our method.

processing, and it needs to be verified in future work. Finally, since the number of feature dimensions of the integrated features is larger than the original feature, the computational cost of the integrated features is high, especially in the training phase. Therefore, we would like to reduce the computational costs of the training phase based on the integrated features in future work.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Numbers JP17H01744, JP17K00148, the MIC/SCOPE #181601001 and the Photo-excitonix Project in Hokkaido University.

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) A. R. Lauricella, E. Wartella and V. J. Rideout : "Young children's screen time: The complex role of parent and child factors," Journal of Applied Developmental Psychology (JADP), vol. 36, pp. 11–17, (2015)
3) H.-C. Shih : "A survey of content-aware video analysis for sports," IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), vol. 28, no. 5, pp. 1212–1231, (2018)
4) Y. Seo, S. Choi, H. Kim and K.-S. Hong : "Where are the ball and players? soccer game analysis with color-based tracking and image mosaick," International Conference on Image Analysis and Processing (ICIAP), pp. 196–203, (1997)
5) S. Iwase and H. Saito : "Parallel tracking of all soccer players by integrating detected positions in multiple view images," IEEE International Conference on Pattern Recognition (ICPR), vol. 4, pp. 751–754, (2004)
6) S. Takahashi, W. Lim and M. Haseyama : "Players tracking approach using level-set method based on color components of players in soccer videos," Workshop on Picture Coding and Image Processing (WPCIP), pp. 67–68, (2010)
7) S. Baysal and P. Duygulu : "Sentioscope: A soccer player tracking system using model field particles," IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), vol. 26, no. 7, pp. 1350–1362, (2016)
8) M. Manafifard, H. Ebadi and H. A. Moghaddam : "Multi-player detection in soccer broadcast videos using a blob-guided particle swarm optimization method," Multimedia Tools and Applications, vol. 76, no. 10, pp. 12251–12280, (2017)
9) J. Perl, A. Grunz and D. Memmert : "Tactics analysis in soccer-an advanced approach," International Journal of Computer Science in Sport (IJCSS), vol. 12, no. 1, pp. 33–44, (2013)
10) X. Wei, L. Sha, P. Lucey, S. Morgan and S. Sridharan : "Large-scale analysis of formations in soccer," IEEE International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8, (2013)
11) A. Bialkowski, P. Lucey, P. Carr, Y. Yue and I. Matthews : "Win at home and draw away: Automatic formation analysis highlighting the differences in home and away team behaviors," MIT Sloan Sports Analytics Conference (MIT SSAC), pp. 1–7, (2014)
12) S. Gondo, T. Inoue, K. Tarukawa and K. Okada : "Soccer tactics analysis supporting system displaying the player's actions in virtual space," IEEE International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 581–586, (2014)
13) Y. Wu, X. Xie, J. Wang, D. Deng, H. Liang, H. Zhang, S. Cheng and W. Chen : "Forvizor: Visualizing spatio-temporal team for mations insoccer," IEEE Transactions on Visualization and Computer Graphics (TVCG), vol. 25, no. 1, pp. 65–75, (2018)
14) G. Suzuki, S. Takahashi, T. Ogawa and M. Haseyama : "Team tactics estimation in soccer videos via deep extreme learning machine based on players formation," IEEE Global Conference on Consumer Electronics (GCCE), pp. 116–117, (2018)
15) Y.-L. Kang, J.-H. Lim, Q. Tian and M. S. Kankanhalli, : "Soccer video event detection with visual keywords," IEEE International Conference on Information, Communications and Signal Processing (ICICSP), vol. 3, pp. 1796-1800, (2003)
16) Q. Ye, Q. Huang, W. Gao and S. Jiang : "Exciting event detection in broad-cast soccer video with mid-level description and incremental learning," ACM International Conference on Multimedia, pp.455-458, (2005)
17) D. A. Sadlier and N. E. O'Connor : "Event detection in field sports video using audio-visual features and a support vector machine," IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), vol. 15, no. 10, pp. 1225-1233, (2005)
18) D. W. Tjondronegoro and Y.-P. P. Chen : "Knowledge-discounted event detection in sports video," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 40, no. 5, pp. 1009-1024, (2010)
19) C. Xu, J. Wang, K. Wan, Y. Li and L. Duan : "Live sports event detection based on broadcast video and web-casting text," ACM International Conference on Multimedia, pp. 221-230, (2006)
20) M. Tavassolipour, M. Karimian and S. Kasaei : "Event detection and summarization in soccer videos using bayesian network and copula," IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), vol.24, no. 2, pp. 291-304, (2014)
21) H. Saraogi, R. A. Sharma and V. Kumar : "Event recognition in broadcast soccer videos," Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), article. 14, pp. 1–7, (2016)
22) 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 (ICONIP), pp. 440-449, (2017)
23) N. Ueda and M. Izumi : "Detecting soccer goal scenes from broadcast video using telop region," Information Engineering Express, vol. 3, no. 2, pp. 25-34, (2017)
24) M. Y. Eldib, B. S. A. Zaid, H. M. Zawbaa, M. El-Zahar and M. El-Saban : "Soccer video summarization using enhanced logo detection," IEEE International Conference on Image Processing (ICIP), pp. 4345–4348, (2009)
25) F. Chen, C. De Vleeschouwer, H. Duxans Barrobs, 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)
26) H. M. Zawbaa, N. El-Bendary, A. E. Hassanien and T.-h. Kim : "Machine learning-based soccer video summarization system," IEEE International Conference on Multimediam Computer Graphics, and Broadcasting (MCGB), pp. 19–28, (2011)
27) A. Raventos, R. Quijada, L. Torres and F. Tarrs : "Automatic summarization of soccer highlights using audio-visual descriptors," SpringerPlus, vol. 4, no. 1, article. 301, pp. 1–19, (2015)
28) 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)
29) G. Fio, T. Romo, N. Correia, P. Centieiro and A. E. Dias : "Automatic generation of sport video highlights based on fan's emotions and content," ACM International Conference on Advances in Computer Entertainment Technology (ACE), article. 29, pp. 1–6 (2016)
30) A. Rehman and T. Saba : "Features extraction for soccer video semantic analysis: Current achievements and remaining issues," Artificial Intelligence Review, vol. 41, no. 3, pp. 451–461, (2014)
31) C. Cortes and V. Vapnik : "Support-vector networks" Springer, vol. 20, no. 3, pp. 273–297, (1995)
32) I. Paraskevas and E. Chilton : "Audio classification using acoustic images for retrieval from multimedia databases," EURASIP Conference focused on Video/Image Processing and Multimedia Communications (EC-VIP-MC), vol. 1, pp 187–192, (2003)
33) C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna : "Rethinking the inception architecture for computer vision," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826, (2016)
34) T. Maeda, T. Yamasaki and K. Aizawa : "Multi-stage object classification featuring confidence analysis of classifier and inclined local naive bayes nearest neighbor," IEEE International Conference on Image Processing (ICIP), pp. 5177–5181, (2014)
35) T. Haruyama, S. Takahashi, T. Ogawa and M. Haseyama : "Es-

timation of Important Scenes in Soccer Videos Based on Collaborative Use of Audio-Visual CNN Features", IEEE Global Conference on Consumer Electronics (GCCE), pp. 710–711, (2018)
36) A. Raventos, R. Quijada, Raul L. Torres, F. Tarres, E. Carasusan and D. Giribet : "The importance of audio descriptors in automatic soccer highlights generation," IEEE International Multi-Conference on Systems, Signals & Devices, pp. 1–6, (2014)
37) A. Krizhevsky, I. Sutskever and G. Hinton : "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, pp. 1097–1105, (2012)
38) H. P. Martinez and G. N. Yannakakis : "Deep multimodal fusion: Combining discrete events and continuous signals," International Conference on Multimodal Interaction (ICMI), pp. 34–41, (2014)
39) J. C. Platt : "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," Advances in Large Margin Classifiers, vol. 10, no. 3, pp. 61–74, (1999)
40) C. Hsu, C. Chang and C. Lin : "A practical guide to support vector classification,", Technical Report, Department of Computer Science and Information Engineer, National Taiwan University, (2014)
41) K. Simonyan and Z. Andrew : "Very deep convolutional networks for large-scale image recognition", arXiv:1409.1556, (2014)



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 the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan in 2003, 2005, and 2007, respectively, all in electronics and information engineering. He is currently an Associate Professor with the Faculty of Information Science and Technology, Hokkaido University. His research interests are multimedia signal processing and its applications. He has been an Associate Editor of the ITE Transactions on Media Technology and Applications. He is a member of the EURASIP, IEICE, and Institute of Image Information and Television Engineers.



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