IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 10, 2021, accepted June 2, 2021, date of publication June 10, 2021, date of current version June 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3088284

Detection of Important Scenes in Baseball Videos via Bidirectional Time Lag Aware Deep Multiset Canonical Correlation Analysis

KAITO HIRASAWA[1], (Student Member, IEEE), KEISUKE MAEDA[2], (Member, IEEE), TAKAHIRO OGAWA[3], (Senior Member, IEEE), AND MIKI HASEYAMA[3], (Senior Member, IEEE)

[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan
[2]Office of Institutional Research, Hokkaido University, Sapporo 060-0808, Japan
[3]Faculty of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan

Corresponding author: Kaito Hirasawa (hirasawa@lmd.ist.hokudai.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP17H01744.

ABSTRACT A novel method for detection of important scenes in baseball videos based on correlation maximization between heterogeneous modalities via bidirectional time lag aware deep multiset canonical correlation analysis (BiTl-dMCCA) is presented in this paper. The proposed method enables detection of important scenes by collaboratively using baseball videos and their corresponding tweets. The technical contributions of this paper are twofold. First, since there are time lags between not only ‘‘tweets and corresponding multiple previous events’’ but also ‘‘events and corresponding multiple following posted tweets’’, the proposed method considers these bidirectional time lags. Specifically, the representation of such bidirectional time lags into the derivation of their covariance matrices is newly introduced. Second, the proposed method adopts textual, visual and audio features calculated from tweets and videos as multi-modal time series features. Important scenes are detected as abnormal scenes via anomaly detection based on a generative adversarial network using multi-modal features projected by BiTl-dMCCA. The proposed method does not need any training data with annotation. Experimental results obtained by applying the proposed method to actual baseball matches show the effectiveness of the proposed method.

INDEX TERMS Unsupervised important scene detection, time lag aware canonical correlation maximization, anomaly detection, generative adversarial network.

I. INTRODUCTION

Many video distribution services have recently become popular due to the development of various network technologies and devices. The number of viewable videos has therefore been increasing. There has been an increased interest in sports during the past ten years or so, and sports video distribution services such as Rakuten Sports,[1] DAZN[2] and MLB.tv[3] have therefore become very popular. Since there are many sports matches held throughout the year, it is difficult for viewers to watch all matches. Unlike basketball or soccer matches, which last for about 80 minutes, a baseball match lasts for about 180 minutes. Therefore, techniques for viewers to easily understand the match context are needed [1].

Generation of highlights would enable viewers to have an easy understanding of the match context. In conventional methods [2], [3], baseball video summarization has been realized on the basis of both metadata for large sports video archives and deep learning techniques for learning the relationship between a highlight and non-highlight video segments. Highlights are generated from many important scenes such as scenes of scoring, good play and appearance of popular players. Therefore, many methods for important scene detection based on video and Twitter analyses have been proposed [4]–[10]. There are various video analysis-based

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

[1] https://sports.rakuten.com/
[2] https://www.dazn.com/
[3] https://www.mlb.com/

methods for detection of important scenes in which a hidden Markov model [4], [5] or a maximum entropy model [6] is applied to cheers of the audience, switching of cameras and player movements. On the other hand, various methods for detection of important scenes that can consider the opinions of viewers have been realized by using Twitter[4] [7]–[9]. Twitter is one of the microblogging services, and users of the service post and receive short text messages called *tweets* [11]. Tweets have been used to report everything from daily life stories to the latest local and global news and events. Conventional methods [12], [13] detect news and events based on multiassignment clustering, n-grams cooccurrence and topic ranking. Since tweets are often posted in real time by users while they are viewing baseball matches, posted tweets may express the content of the match and the opinions of viewers. Thus, posted tweets express some information related to the important scenes of the corresponding match. On the other hand, news is summarized and its information is limited. Therefore, it is difficult for news to fully reflect the opinions of viewers. We therefore use Twitter instead of news to detect important scenes.

Since existing methods can extract semantic information from tweets, Twitter analysis has an advantage for extraction of opinions or feelings of viewers. On the other hand, video analysis is better than Twitter analysis for representation of player movements and cheers of the audiences. Therefore, since it is expected that a method combining Twitter analysis and video analysis would enable high-quality important scene detection, a method based on tweets and videos has been proposed [10]. This method combines burst detection of tweets and video content analysis using score-box and audio features. Multi-modal fusion methods were then proposed [14], [15]. A multi-modal aspect-aware latent factor model [14] realizes explainable recommendation by leveraging user reviews and item images. In addition, Flexible Multimodal Hashing [15] can adaptively generate hash codes according to specific query types, and it can deal with the modality-missing problem in multimedia retrieval. For the use of multi-modal information such as information on videos and tweets, we have to try to solve the following problem. Since there is a temporal difference between the timing of postings on Twitter and the occurrence of the corresponding previous events, the conventional method [10] focuses on the time lag between the timing of posts on Twitter and the occurrence of the corresponding previous event. However, an event that occurs affects multiple following posted tweets. Therefore, not only time lags between "tweets and corresponding multiple previous events" but also time lags between "events and corresponding multiple following posted tweets" must be considered. In other words, when events occur and tweets are posted, it is necessary to consider not only past time lags but also future time lags. A method that considers a bidirectional time series such as a Bidirectional Long Short-Term Memory (Bi-LSTM) [16] can represent the features of the time series data with high accuracy. Therefore, it is expected that highly accurate detection of important scenes can be achieved by a method that considers these bidirectional time lags.

In this paper, we propose a method for detection of important scenes in baseball videos that considers bidirectional time lags. Specifically, the proposed method derives a novel feature embedding approach considering time lags between not only "tweets and corresponding multiple previous events" but also "events and corresponding multiple following posted tweets". The proposed method newly derives bidirectional time lag aware deep multiset canonical correlation analysis (BiTl-dMCCA), which is an extended version of dMCCA [17], to consider bidirectional time lags depending on events occurring and posted tweets. dMCCA and BiTl-dMCCA learn non-linear transformations from different modalities to a shared subspace so that the representations maximize the ratio of between- and within-modality covariance of the observations. While dMCCA cannot learn non-linear transformations considering the time lags between modalities, BiTl-dMCCA can learn non-linear transformations considering the time lag. Bi-LSTM, which is an extended version of LSTM, considers bidirectionaly by using an LSTM network in which past data are recursed into the future and an LSTM network in which future data are recursed into the past. On the other hand, BiTl-dMCCA considers the bidirectionaly of time lags. Thus, it is possible to calculate the features considering bidirectionality with both Bi-LSTM and BiTl-dMCCA. However, only BiTl-dMCCA can associate the correlation of features considering bidirectionality. This is the difference between Bi-LSTM and BiTl-dMCCA. The proposed method extracts textual, visual and audio features from tweets and baseball videos. By introducing consideration of the bidirectional time lags into the derivation of covariance matrices of BiTl-dMCCA, it can correctly consider their relationships. This is the biggest novelty of this paper. Moreover, important scenes of baseball matches have more interest and excitement than other scenes. By focusing on the above characteristics, the proposed method detects important scenes as abnormal scenes via anomaly detection based on a generative adversarial network (GAN) [18] using multi-modal features projected by BiTl-dMCCA. It should be noted that both feature extraction via BiTl-dMCCA and the GAN-based anomaly detection model can be performed in a completely unsupervised fashion. The method [19] that realizes video summarization by extracting the most characteristic scenes based on cross-correlation optimization excludes similar scenes and extracts meaningful scenes. When this method is applied to baseball videos, it is possible to detect important scenes to some extent, but there is a possibility of over-detecting normal scenes such as scenes showing audiences and benches. Furthermore, other important scenes that are similar to the detected important scenes cannot be detected. Therefore, important scenes that are more exciting and interesting than other scenes are treated as abnormal scenes, and we realize detection of important scenes by applying the anomaly

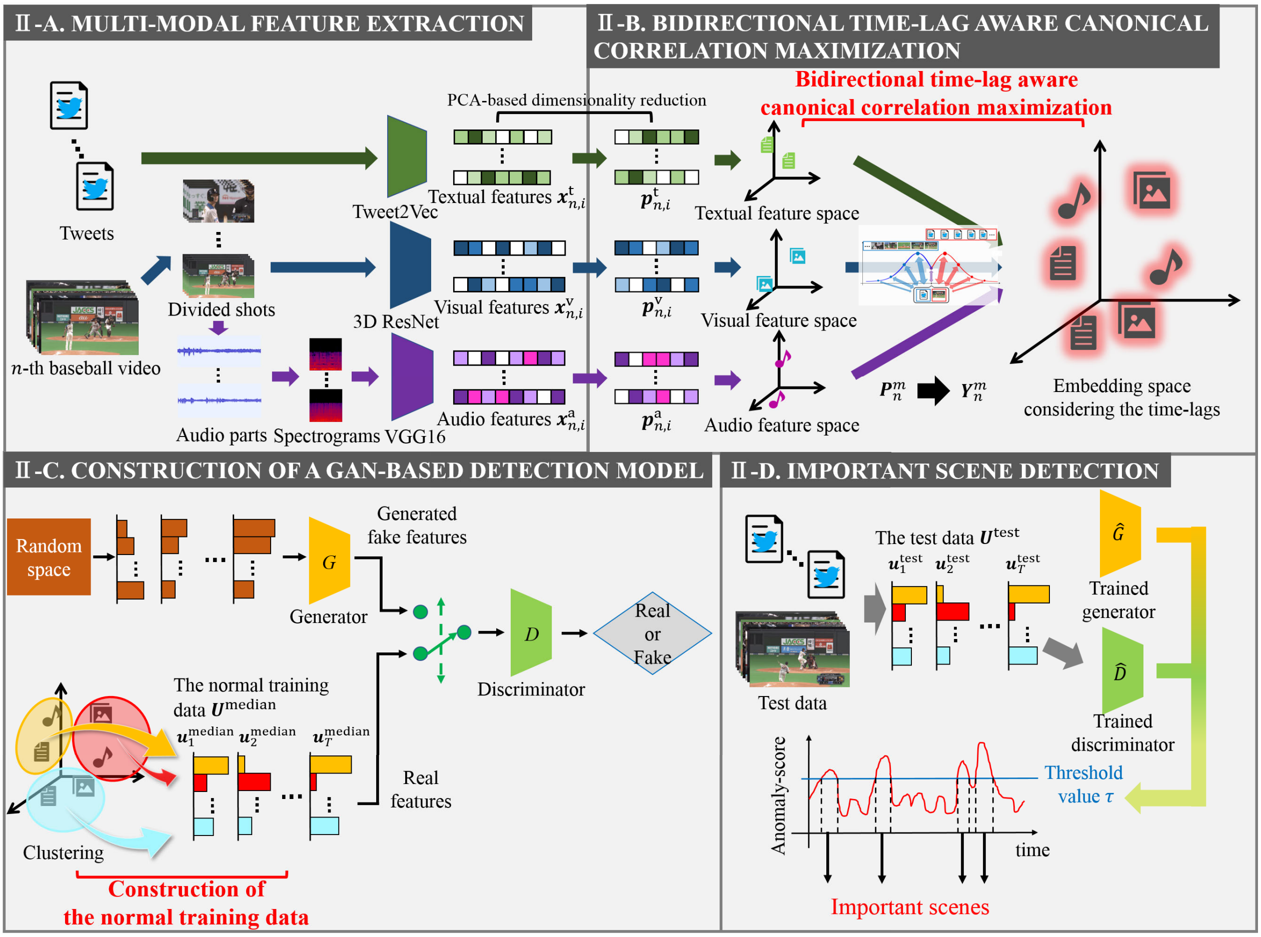

II-A. MULTI-MODAL FEATURE EXTRACTION
II-B. BIDIRECTIONAL TIME-LAG AWARE CANONICAL CORRELATION MAXIMIZATION
Bidirectional time-lag aware canonical correlation maximization
PCA-based dimensionality reduction
Tweets
Tweet2Vec
Textual features $x^{t}_{n,i}$
$p^{t}_{n,i}$
Textual feature space
Divided shots
n-th baseball video
3D ResNet
Visual features $x^{v}_{n,i}$
$p^{v}_{n,i}$
Visual feature space
Audio parts
Spectrograms
VGG16
Audio features $x^{a}_{n,i}$
$p^{a}_{n,i}$
Audio feature space
P^{m}_{n}
Y^{m}_{n}
Embedding space considering the time-lags
II-C. CONSTRUCTION OF A GAN-BASED DETECTION MODEL
Random space
G
Generator
Generated fake features
D
Discriminator
Real or Fake
The normal training data U^{median}
u^{median}_{1} u^{median}_{2} u^{median}_{T}
Real features
Clustering
Construction of the normal training data
II-D. IMPORTANT SCENE DETECTION
The test data U^{test}
u^{test}_{1} u^{test}_{2} u^{test}_{T}
Test data
$\hat{G}$
Trained generator
$\hat{D}$
Trained discriminator
Anomaly-score
Threshold value τ
time
Important scenes

FIGURE 1. Overview of the proposed method. The proposed method has four phases. Textual, visual and audio features are extracted as multi-modal features as described in Section II-A. In order to consider the bidirectional temporal difference between textual features and the other features, the proposed method newly performs bidirectional time lag aware canonical correlation maximization as described in Section II-B. Furthermore, by using the obtained features, a GAN-based model for detection of important scenes is constructed as described in Section II-C. Finally, we perform important scene detection as described in Section II-D.

detection method. This is the second contribution of this paper. This paper is an extended version of [20].

The contributions of this paper are summarized as follows. BiTl-dMCCA enables the transformation of multi-modal time series features to effective new latent features with consideration of their bidirectional time lags. Furthermore, by inputting the calculated features in the novel embedding space into the anomaly detection model, we can realize unsupervised detection of important scenes in baseball videos. This paper is an extended version of dMCCA [17]. Specifically, the proposed method introduces the consideration of bidirectional time lags into the derivation of covariance matrices of dMCCA. By non-linear transformation based on BiTl-dMCCA, the proposed method realizes flexible embedding for heterogeneous features with complex relationships.

II. DETECTION OF IMPORTANT SCENES VIA BiTl-dMCCA

In this section, we explain the novel method for detection of important scenes in baseball videos based on BiTl-dMCCA. FIGURE 1 shows an overview of the proposed method. The proposed method detects important scenes using tweets posted by viewers and baseball videos. Specifically, the proposed method extracts textual features from tweets using a language model. Visual and audio features are extracted from baseball videos using a convolutional neural network (CNN) model (Section II-A). Next, the extracted features are transformed into features maximizing their correlation with consideration of bidirectional time lags based on BiTl-dMCCA (Section II-B). Moreover, the proposed method constructs a GAN-based anomaly detection model for important scene detection from the transformed features (Section II-C). Details of the detection are explained in Section II-D.

A. MULTI-MODAL FEATURE EXTRACTION

In this subsection, we show the multi-modal feature extraction for each modality. When an n-th baseball video ($n = 1, 2, \ldots, N$; N being the number of training videos) is given, the proposed method extracts multi-modal features $\boldsymbol{x}^{m}_{n,i}(i = 1, 2, \ldots, I_n; I_n$ being the number of the tweets for the n-th video). Note that $m \in \{\mathrm{t}, \mathrm{v}, \mathrm{a}\}$ means the modality. t, v and a mean textual, visual and audio modalities, respectively.

1) TEXTUAL FEATURES

Tweets posted by viewers watching baseball matches are important elements for analysis of the content of the match and the opinions of viewers. Therefore, the proposed method extracts textual features from these tweets. Textual features $\boldsymbol{x}^{\mathrm{t}}_{n,i}$ are extracted from the main texts of tweets posted by viewers watching baseball matches based on Tweet2Vec [21], which is a representative language model. In the proposed method, the Tweet2Vec model pre-trained by using tweets related to professional baseball is used. Its scheme is shown in Section III-A. Tweet2Vec is a type of bidirectional recursive neural network and is an extended version of LSTM [22]. By using Tweet2Vec, textual features that are robust to abbreviations, typographical errors and slang unique to Twitter can be extracted.

2) VISUAL FEATURES

Since visual sequences in baseball videos are important elements to understand scene situations, visual features should be extracted from visual sequences. The proposed method extracts visual features from shots including frames of baseball videos when tweets are posted by inputting these shots into a 3D ResNet model [23]. Note that 3D ResNet is pre-trained on the Kinetics dataset [24]. The Kinetics dataset is a large-scale dataset covering a diverse range of human actions. Then visual features $\boldsymbol{x}^{\mathrm{v}}_{n,i,j}$ ($j = 1, 2, \ldots, J_{n,i}; J_{n,i}$ being the number of the divided frames within the i-th shot of the n-th match) are extracted from the global average pooling layer of 3D ResNet. The following visual features are obtained to express shots: $\boldsymbol{x}^{\mathrm{v}}_{n,i} = (1/J_{n,i}) \sum_{j=1}^{J_{n,i}} \boldsymbol{x}^{\mathrm{v}}_{n,i,j}$.

3) AUDIO FEATURES

Audio features should be extracted from audio sequences in baseball videos. Audio features are extracted from a spectrogram of each shot based on the pre-trained CNN model. It is known that the use of spectrogram-based features outputted from the pre-trained CNN model is effective for representing audio data [25], [26]. It is not common to extract audio features using VGG16 trained by using ImageNet. However, since the effectiveness of the audio feature extraction using a CNN model trained by using ImageNet has been reported in the method for detection of important scenes of other sports videos [27], the proposed method experimentally adopts VGG16 trained by using ImageNet. Thus, it is expected to be effective for our tasks. The proposed method calculates audio features $\boldsymbol{x}^{\mathrm{a}}_{n,i}$ of the i-th shot from the output of the final pooling layer of VGG16 [28]. Note that the proposed method adopts VGG16 pre-trained by using the ImageNet dataset [29].

B. BIDIRECTIONAL TIME LAG AWARE DEEP MULTISET CANONICAL CORRELATION ANALYSIS

This subsection shows the new derivation of BiTl-dMCCA. First, the proposed method applies principal component analysis (PCA) [30] to $\boldsymbol{x}^{\mathrm{t}}_{n,i}$, $\boldsymbol{x}^{\mathrm{v}}_{n,i}$ and $\boldsymbol{x}^{\mathrm{a}}_{n,i}$ to avoid overfitting in BiTl-dMCCA. PCA is applied to avoid overfitting in BiTl-dMCCA and to align dimensions of the features of each modality. Note that BiTl-dMCCA is an extended version of dMCCA, which is a method assuming that features with the same dimensions of each modality are inputted. Then the proposed method respectively obtains $\boldsymbol{p}^{\mathrm{t}}_{n,i}$, $\boldsymbol{p}^{\mathrm{v}}_{n,i}$ and $\boldsymbol{p}^{\mathrm{a}}_{n,i}$, which are d_p-dimensional vectors, and their feature matrices $\boldsymbol{P}^{m}_{n} = [\boldsymbol{p}^{m}_{n,1}, \ldots, \boldsymbol{p}^{m}_{n,i}, \ldots, \boldsymbol{p}^{m}_{n,I_n}] \in \mathbb{R}^{d_p \times I_n}$. Thus, by transforming the input features $\boldsymbol{P}^{m}_{n}$ based on a multi-layered neural network, the proposed method obtains $\boldsymbol{Y}^{m}_{n} = [\boldsymbol{y}^{m}_{n,1}, \ldots, \boldsymbol{y}^{m}_{n,i}, \ldots, \boldsymbol{y}^{m}_{n,I_n}] \in \mathbb{R}^{d_y \times I_n}$ from the last layer of the neural network.

Tweets posted by viewers are influenced by corresponding multiple previous events. Moreover, an event influences corresponding multiple following posted tweets. In order to obtain the canonical correlation between multi-modal features with consideration of such bidirectional time lags, BiTl-dMCCA calculates the covariance matrices with consideration of the bidirectional time lags. Since the influence of events on tweets tends to be gradually weakened with increase in their time intervals, we should consider such continuous influence that occurs due to the bidirectional time lag. Therefore, as shown in FIGURE 2, BiTl-dMCCA assumes that posted tweets are affected by present to past events, and their influence is determined on the basis of the Poisson distribution defined for time lags. Furthermore, BiTl-dMCCA assumes that events occurring also affect present to future tweets, and their influence is determined on the basis of the Poisson distribution defined for time lags. Since the Poisson

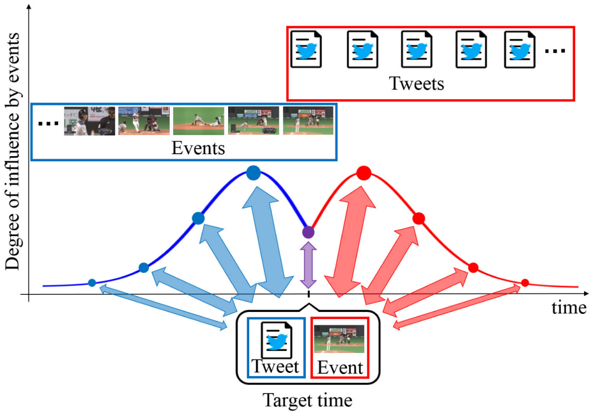


FIGURE 2. Relationships between "target tweet and corresponding multiple previous events" and "target event and corresponding multiple following tweets". The proposed method associates the tweet with events from the present to past as shown in the blue rectangles. Furthermore, the proposed method associates the event with tweets from the present to future as shown in the red rectangles. Then they are weighted according to the degree of influence defined by the Poisson distribution.

distribution expresses the number of events occurring in a fixed interval of time, we can regard it as the degree of influence from tweets and baseball events.

In order to calculate $\boldsymbol{Y}_n^m$ considering the influence defined by these bidirectional time lags with optimization of parameters of the multi-layered neural network based on BiTl-dMCCA, we maximize the average inter-set correlation (ISC) [31] defined as

$$\rho = \frac{1}{d_y}\sum_{d_h=1}^{d_y}\rho_{d_h}, \tag{1}$$

where

$$\rho_{d_h} = \frac{1}{M-1}\frac{\boldsymbol{\psi}_{d_h}^\top \boldsymbol{R}_B \boldsymbol{\psi}_{d_h}}{\boldsymbol{\psi}_{d_h}^\top \boldsymbol{R}_W \boldsymbol{\psi}_{d_h}}. \tag{2}$$

Note that $\boldsymbol{\psi}_{d_h} \in \mathbb{R}^{d_y}(d_h = 1, 2, \ldots, d_y)$ is the optimal projection common to all modalities, and $M(= 3)$ is the number of modalities. Then $\boldsymbol{R}_W$ and $\boldsymbol{R}_B$ are the within-set covariance matrix and the between-set covariance matrix, respectively, with consideration of the bidirectional time lags defined as

$$\boldsymbol{R}_W = \sum_{n=1}^{N}\sum_{m\in\{\mathrm{t,v,a}\}} \underline{\boldsymbol{C}}_n^{m,m}, \tag{3}$$

$$\boldsymbol{R}_B = \sum_{n=1}^{N}\sum_{m_1\in\{\mathrm{t,v,a}\}}\sum_{m_2\in\{\mathrm{t,v,a}\},m_2\neq m_1} \overline{\boldsymbol{C}}_n^{m_1,m_2}. \tag{4}$$

The same scaling value $(I_n - 1)^{-1}M^{-1}$ is omitted. Equation (3) represents the covariance matrix within each of the textual, visual and audio modalities. For simplicity of the calculation of within-set covariance, different types of features are differentiated in Eq. (3). The details of the covariance matrices are defined as follows:

$$\underline{\boldsymbol{C}}_n^{m,m} = \widehat{\boldsymbol{Y}}^m_{n,0}\widehat{\boldsymbol{Y}}^{m\top}_{n,0} + \widetilde{\boldsymbol{Y}}^m_{n,0}\widetilde{\boldsymbol{Y}}^{m\top}_{n,0}, \tag{5}$$

$$\overline{\boldsymbol{C}}_n^{m_1,m_2} = \begin{cases} \dfrac{\sum_{l=0}^{L-1}\frac{e^{-\lambda}\lambda^l}{l!}\left(\widehat{\boldsymbol{Y}}^{m_1}_{n,0}\widehat{\boldsymbol{Y}}^{m_2\top}_{n,l} + \widetilde{\boldsymbol{Y}}^{m_1}_{n,l}\widetilde{\boldsymbol{Y}}^{m_2\top}_{n,0}\right)}{\sum_{l=0}^{L-1}\frac{e^{-\lambda}\lambda^l}{l!}} \\ \quad \text{if } (m_1 \in \{\mathrm{t}\}, m_2 \in \{\mathrm{v,a}\}) \\ \dfrac{\sum_{l=0}^{L-1}\frac{e^{-\lambda}\lambda^l}{l!}\left(\widehat{\boldsymbol{Y}}^{m_1}_{n,l}\widehat{\boldsymbol{Y}}^{m_2\top}_{n,0} + \widetilde{\boldsymbol{Y}}^{m_1}_{n,0}\widetilde{\boldsymbol{Y}}^{m_2\top}_{n,l}\right)}{\sum_{l=0}^{L-1}\frac{e^{-\lambda}\lambda^l}{l!}} \\ \quad \text{if } (m_1 \in \{\mathrm{v,a}\}, m_2 \in \{\mathrm{t}\}) \\ \widehat{\boldsymbol{Y}}^m_{n,0}\widehat{\boldsymbol{Y}}^{m\top}_{n,0} + \widetilde{\boldsymbol{Y}}^m_{n,0}\widetilde{\boldsymbol{Y}}^{m\top}_{n,0} \qquad \text{(otherwise)}, \end{cases} \tag{6}$$

where L determines how many previous events affect the posted tweet and how many following tweets are affected by the occurring event. Furthermore, λ is a parameter of the Poisson distribution and controls the peak of the distribution. Note that λ corresponds to the mean and variance of the distribution. Feature matrices $\widehat{\boldsymbol{Y}}^m_{n,l} = [\boldsymbol{y}^m_{n,L-l}, \ldots, \boldsymbol{y}^m_{n,I_n-L-l}]$ $(l = 0, \ldots, L-1)$ and $\widetilde{\boldsymbol{Y}}^m_{n,l} = [\boldsymbol{y}^m_{n,L+l}, \ldots, \boldsymbol{y}^m_{n,I_n-L+l}]$ $(l = 0, \ldots, L-1)$ are mean-normalized.

BiTl-dMCCA maximizes the average ISC by solving the following generalized eigenvalue problem:

$$\boldsymbol{R}_B\boldsymbol{\psi}_{d_h} = \rho_{d_h}\boldsymbol{R}_W\boldsymbol{\psi}_{d_h}. \tag{7}$$

By applying the Lagrange multiplier method to the maximization problem of Eq. (1), we obtain the general eigenvalue problem of Eq. (7) [32], [33]. BiTl-dMCCA is an extended version of dMCCA that solves the problem. The partial derivation of the eigenvalue ρ_{d_h} with respect to $\boldsymbol{Y}_n^m$ can be written as follows [34]:

$$\frac{\partial\rho_{d_h}}{\partial\boldsymbol{Y}_n^m} = \boldsymbol{\psi}_{d_h}^\top\left(\frac{\partial\boldsymbol{R}_B}{\partial\boldsymbol{Y}_n^m} - \rho_{d_h}\frac{\partial\boldsymbol{R}_W}{\partial\boldsymbol{Y}_n^m}\right)\boldsymbol{\psi}_{d_h}. \tag{8}$$

Gradients of $\boldsymbol{R}_W$ and $\boldsymbol{R}_B$ are derived by referring to [35]. Consequently, the proposed method obtains an effective feature matrix $\boldsymbol{Y}_n^m$ considering the bidirectional time lags.

The main novelty of this paper is the introduction of bidirectional time lags into the derivation of covariance matrices of dMCCA [17]. By non-linear transformation based on BiTl-dMCCA, the proposed method realizes flexible embedding for heterogeneous features with complex relationships. Since general canonical correlation maximization approaches often use the simple between-set covariance matrix, the bidirectional temporal difference between different modalities cannot be considered. On the other hand, BiTl-dMCCA assumes that textual features are related to visual and audio features before the tweets are posted. Also, it is assumed that visual and audio features are related to textual features after events have occurred. The covariance matrices in Eq. (6) reflect these characteristics. This is the main contribution for improving the embedding performance of heterogeneous features in the proposed method.

C. CONSTRUCTION OF A GAN-BASED DETECTION MODEL

Important scenes in baseball matches have more excitement and interest than those in other scenes. By focusing on these characteristics, we introduce a new approach for detecting important scenes as abnormal scenes. The proposed method, which adopts an extended version of the multivariate anomaly detection strategy with a GAN (MAD-GAN) [36], detects important scenes based on an anomaly score that indicates the degree of abnormality.

In order to construct the MAD-GAN, we have to prepare a set of normal data not including important scenes. However, the training data obtained by the procedures described in the previous subsection may include some important scenes. For realizing an unsupervised approach, we obtain training data by the method shown in FIGURE 3. The details of the method are as follows. First, features in the embedding space are transformed into data considering the multi-modal time series data. Given $\boldsymbol{Y}_n^m = [\boldsymbol{y}^m_{n,1}, \ldots, \boldsymbol{y}^m_{n,i}, \ldots, \boldsymbol{y}^m_{n,I_n}]$ obtained by BiTl-dMCCA, we perform clustering of $\boldsymbol{y}^m_{n,i}$ and assign them to C clusters using the k-means algorithm [37]. Then the proposed method calculates a Bag-of-Feature (BoF)-based feature vector $\boldsymbol{u}_{n,t}$, for which the elements are the numbers of vectors assigned to each cluster, for the n-th baseball match at

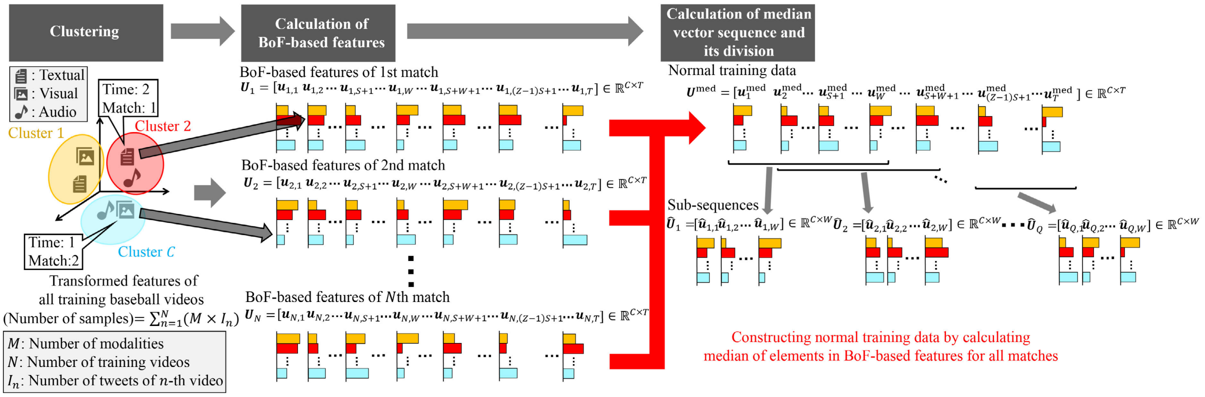

Clustering
Calculation of BoF-based features
Calculation of median vector sequence and its division
: Textual
: Visual
: Audio
Time: 2
Match: 1
Cluster 1
Cluster 2
Cluster C
Time: 1
Match:2
Transformed features of all training baseball videos
(Number of samples)= $\sum_{n=1}^{N}(M \times I_n)$
M: Number of modalities
N: Number of training videos
I_n: Number of tweets of n-th video
BoF-based features of 1st match
$\boldsymbol{U}_1 = [\boldsymbol{u}_{1,1}\ \boldsymbol{u}_{1,2} \cdots \boldsymbol{u}_{1,S+1} \cdots \boldsymbol{u}_{1,W} \cdots \boldsymbol{u}_{1,S+W+1} \cdots \boldsymbol{u}_{1,(Z-1)S+1} \cdots \boldsymbol{u}_{1,T}] \in \mathbb{R}^{C\times T}$
BoF-based features of 2nd match
$\boldsymbol{U}_2 = [\boldsymbol{u}_{2,1}\ \boldsymbol{u}_{2,2} \cdots \boldsymbol{u}_{2,S+1} \cdots \boldsymbol{u}_{2,W} \cdots \boldsymbol{u}_{2,S+W+1} \cdots \boldsymbol{u}_{2,(Z-1)S+1} \cdots \boldsymbol{u}_{2,T}] \in \mathbb{R}^{C\times T}$
BoF-based features of Nth match
$\boldsymbol{U}_N = [\boldsymbol{u}_{N,1}\boldsymbol{u}_{N,2} \cdots \boldsymbol{u}_{N,S+1} \cdots \boldsymbol{u}_{N,W} \cdots \boldsymbol{u}_{N,S+W+1} \cdots \boldsymbol{u}_{N,(Z-1)S+1} \cdots \boldsymbol{u}_{N,T}] \in \mathbb{R}^{C\times T}$
Normal training data
$\boldsymbol{U}^{\mathrm{med}} = [\boldsymbol{u}_1^{\mathrm{med}}\ \boldsymbol{u}_2^{\mathrm{med}} \cdots \boldsymbol{u}_{S+1}^{\mathrm{med}} \cdots \boldsymbol{u}_W^{\mathrm{med}} \cdots \boldsymbol{u}_{S+W+1}^{\mathrm{med}} \cdots \boldsymbol{u}_{(Z-1)S+1}^{\mathrm{med}} \cdots \boldsymbol{u}_T^{\mathrm{med}}] \in \mathbb{R}^{C\times T}$
Sub-sequences
$\widehat{\boldsymbol{U}}_1 = [\widehat{\boldsymbol{u}}_{1,1}\widehat{\boldsymbol{u}}_{1,2} \cdots \widehat{\boldsymbol{u}}_{1,W}] \in \mathbb{R}^{C\times W}$
$\widehat{\boldsymbol{U}}_2 = [\widehat{\boldsymbol{u}}_{2,1}\widehat{\boldsymbol{u}}_{2,2} \cdots \widehat{\boldsymbol{u}}_{2,W}] \in \mathbb{R}^{C\times W}$
$\widehat{\boldsymbol{U}}_Q = [\widehat{\boldsymbol{u}}_{Q,1}\widehat{\boldsymbol{u}}_{Q,2} \cdots \widehat{\boldsymbol{u}}_{Q,W}] \in \mathbb{R}^{C\times W}$
Constructing normal training data by calculating median of elements in BoF-based features for all matches

FIGURE 3. Method for obtaining normal training data without outliers. First, we perform clustering of the multi-modal features obtained by BiTl-dMCCA. Next, the proposed method calculates the BoF-based features for *N* matches of each cluster. Then the normal training data U^{med} are obtained by calculating the median of the elements of the BoF-based features. Finally, in order to realize efficient training of the important scene detection model, we divide the normal training data U^{med} into sub-sequences.

time t $(= 1, 2, \ldots, T$; T being the number of match time divisions). Then we define $\boldsymbol{U}_n = [\boldsymbol{u}_{n,1}, \ldots, \boldsymbol{u}_{n,t}, \ldots, \boldsymbol{u}_{n,T}] \in \mathbb{R}^{C\times T}$. Note that the number of match time divisions T is normalized to be the same size in all matches. Next, our method calculates the median of the corresponding elements in $\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_N$ to obtain the normal training data $\boldsymbol{U}^{\mathrm{med}} = [\boldsymbol{u}_1^{\mathrm{med}}, \boldsymbol{u}_2^{\mathrm{med}}, \ldots, \boldsymbol{u}_T^{\mathrm{med}}] \in \mathbb{R}^{C\times T}$. Note that we normalize $\boldsymbol{U}^{\mathrm{med}}$ so that the sum of the elements belonging to each cluster c becomes 1. In our method, since we regard important scenes as outliers, the synthesized training data $\boldsymbol{U}^{\mathrm{med}}$ including only normal data can be obtained by removing outliers by adopting the median. This scheme is the key procedure of the second contribution explained below.

Next, in order to handle the multi-modal time series data, the proposed method constructs a generator (G) and a discriminator (D) by two LSTMs according to [36]. As with the training of the general GAN model, fake features are generated from a random latent space by G and the generated features are inputted into D. On the other hand, D tries to distinguish the generated fake features and the original training data. In order to realize efficient training of the important scene detection model, the normal training data $\boldsymbol{U}^{\mathrm{med}}$ are divided into sub-sequences. Specifically, the proposed method utilizes a step size S and a window size W to divide the normal training data $\boldsymbol{U}^{\mathrm{med}}$ into a set of sub-sequences and calculates the sub-sequence $\widehat{\boldsymbol{U}}_q = [\widehat{\boldsymbol{u}}_{q,1}, \ldots, \widehat{\boldsymbol{u}}_{q,k}, \ldots \widehat{\boldsymbol{u}}_{q,W}] \in \mathbb{R}^{C\times W} (q = 1, 2, \ldots, Q; k = 1, 2, \ldots, W)$, where $Q = \frac{T-W}{S}$ is the number of sub-sequences. Then a set of sub-sequences $\widehat{\boldsymbol{U}} = [\widehat{\boldsymbol{U}}_1, \widehat{\boldsymbol{U}}_2, \ldots, \widehat{\boldsymbol{U}}_Q] \in \mathbb{R}^{Q\times C\times W}$ is obtained from the calculated sub-sequences. Furthermore, $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_Q] \in \mathbb{R}^{Q\times C\times W}$ is a set of sub-sequences taken from a random space. By respectively inputting $\widehat{\boldsymbol{U}}$ and $\boldsymbol{Z}$ into the important scene detection model, G and D are trained by solving the following two-player minimax problem:

$$\min_G \max_D V(G, D) = \mathbb{E}_{\widehat{\boldsymbol{U}}\sim p_{\mathrm{data}}}\left[\log D(\widehat{\boldsymbol{U}})\right] + \mathbb{E}_{\boldsymbol{Z}\sim p_{\mathrm{fake}}}\left[\log\left(1 - D\left(G(\boldsymbol{Z})\right)\right)\right], \quad (9)$$

where $\widehat{\boldsymbol{U}}$ is the variable following the prior distribution p_{data} of real data $\widehat{\mathcal{U}}$. Similarly, $\boldsymbol{Z}$ is the variable following the latent distribution p_{fake} of fake data $\mathcal{Z}$. Consequently, the trained models $\widehat{G}$ and $\widehat{D}$ are obtained by performing sufficient rounds of iterations.

The second contribution of our method is the construction of an unsupervised important scene detection model based on an anomaly detection scheme. Our anomaly detection model requires normal data without outliers. However, we realize the model construction without provision of the label (normal/outlier) by calculating the median of the BoF-based features included in all data. Thus, important scenes can be detected in a completely unsupervised fashion, and its details are shown in the following subsection.

D. IMPORTANT SCENE DETECTION

This subsection shows the detection of important scenes as a test phase. The proposed method extracts features $\boldsymbol{p}_i^{\mathrm{t}}$, $\boldsymbol{p}_i^{\mathrm{v}}$ and $\boldsymbol{p}_i^{\mathrm{a}}$ from a test baseball match to obtain $\boldsymbol{P}^m$. Then $\boldsymbol{Y}^m = [\boldsymbol{y}_1^m, \ldots, \boldsymbol{y}_i^m, \ldots, \boldsymbol{y}_I^m]$ is obtained from the last layer of the neural network of BiTl-dMCCA. Note that I is the number of tweets of the test baseball video. Next, the proposed method assigns $\boldsymbol{y}_i^m$ to C clusters in the same manner as $\boldsymbol{y}_{n,i}^m$. Then the test data sequence $\boldsymbol{U}^{\mathrm{test}} = [\boldsymbol{u}_1^{\mathrm{test}}, \ldots, \boldsymbol{u}_t^{\mathrm{test}}, \ldots, \boldsymbol{u}_T^{\mathrm{test}}] \in \mathbb{R}^{C\times T}$ is obtained. Note that the length of $\boldsymbol{U}^{\mathrm{test}}$ is the same as $\boldsymbol{U}$, and we normalize $\boldsymbol{U}^{\mathrm{test}}$ in the same manner as $\boldsymbol{U}$. $\boldsymbol{U}^{\mathrm{test}}$ is divided into a set of sub-sequences to calculate the anomaly score, and we calculate the sub-sequence $\widehat{\boldsymbol{U}}_q^{\mathrm{test}} = [\widehat{\boldsymbol{u}}_{q,1}^{\mathrm{test}}, \ldots \widehat{\boldsymbol{u}}_{q,k}^{\mathrm{test}}, \ldots \widehat{\boldsymbol{u}}_{q,W}^{\mathrm{test}}] \in \mathbb{R}^{C\times W}$.

$\widehat{G}$ and $\widehat{D}$ are utilized to calculate the anomaly score. Note that $\widehat{G}$ can represent a model based on distributions from the normal data, and $\widehat{D}$ can distinguish anomaly data and normal data. Thus, by using $\widehat{D}$ and the residuals between $\widehat{G}$ and the test data, the anomaly score at time t is defined as follows:

$$AS_t = \frac{1}{O_t} \sum_{q,k\in\{q\times S+k=t\}} \left\{\alpha\left|\widehat{\boldsymbol{u}}_{q,k}^{\mathrm{test}} - \widehat{G}(z_{q,k})\right| + (1-\alpha)\widehat{D}(\widehat{\boldsymbol{u}}_{q,k}^{\mathrm{test}})\right\}, \quad (10)$$

where α is a parameter, and O_t is the number of combinations of q and k that satisfy $q \times S + k = t$. Moreover, the proposed method finds the optimal sample $z_{q,w}$ representing the test data from the latent space by referring to [36]. Note that $\widehat{G}(z_{q,w})$ outputs data similar to the normal data generated from the latent space. Therefore, if $\hat{\boldsymbol{u}}_{q,k}^{\text{test}}$ becomes more abnormal, $|\hat{\boldsymbol{u}}_{q,k}^{\text{test}} - \widehat{G}(z_{q,k})|$ becomes a larger value. Moreover, if $\hat{\boldsymbol{u}}_{q,w}^{\text{test}}$ is more abnormal, $\widehat{D}(\hat{\boldsymbol{u}}_{q,w}^{\text{test}})$ outputs a larger value. When $E(AS_t, 1) > \tau$, the proposed method determines the scene at time t as an important scene. Note that τ is a predetermined threshold value, and $E(\cdot, \cdot)$ is the cross entropy error.

III. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETTING

(**Evaluation dataset**) We carried out an experiment to verify the effectiveness of the proposed method. We used 28 baseball videos (30 fps) and their corresponding tweets. We collected videos that were broadcasted from June 13th to September 27th in 2019 by Pacific League TV[5] and we collected tweets by using the query "#lovefighters", which is an official hashtag of the professional baseball team. We used 23 randomly selected matches as training data and the other five matches as test data. In the field of computer vision, it is common to use public datasets. However, private datasets collected by authors are generally used in experiments using tweets and videos [10]. Therefore, we cannot verify the time lags, our biggest novelty, by using public datasets. We therefore need to use a private dataset. Since previous works [4]–[6] used 24, 6 and 10 matches for experiments, the number of baseball matches used in our experiments is sufficient for verifying the performance of our method. For training Tweet2Vec, 27 hashtags related to professional baseball teams were used as queries.

(**Ablation studies**) For confirming the validity of the proposed method, we used the following comparative methods (Comps. 1-15).

TABLE 1. Features used in Comps. 1-6.

Features	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6
Textual	✓			✓	✓	
Visual		✓		✓		✓
Audio			✓		✓	✓

Comps. 1-3: These are methods using a unimodal feature shown in TABLE 1. Since these methods use a unimodal feature, an embedding scheme is not used. These methods detect important scenes based on the same detection model as that used in the proposed method. By using Comps. 1-3 in the experiment, we could evaluate the effectiveness of introducing multi-modal features.

Comps. 4-6: These are methods using two types of features shown in TABLE 1. These methods use CCA [38], which is the simplest embedding scheme. Comps. 4 and 5 consider time lags. Comp. 6 is a method using visual and audio features. Since both visual and audio features are extracted from the videos, there is no time lag. Therefore, Comp.6 does not need to consider time lags. These methods detect important scenes based on the same detection model as that used in the proposed method. Comps. 4-6 were used to evaluate the effectiveness of introducing multi-modal features by comparing with Comps. 1-3.

Comp. 7: This is a method that simply integrates detection models constructed for each modality. Specifically, this method determines important scenes by majority voting of detection results using Comps. 1-3. We could evaluate the effectiveness of introducing a feature embedding scheme by using Comp. 7.

Comp. 8: This is a method based on [10]. This approach considers time lags between the timing of posts on Twitter and the occurrence of only one corresponding event. By comparing the proposed method and Comp. 8, we could determine whether consideration of tweets and multiple corresponding events is effective.

Comp. 9: This is a method using dMCCA [17] that does not consider time lags. This method detects important scenes based on the same detection model as that used in the proposed method. We could evaluate the consideration of time lags by using Comp. 10.

Comp. 10: This is a method considering only time lags between the tweets and their corresponding multiple previous events. This method detects important scenes based on the same detection model as that used in the proposed method. By comparing the proposed method and Comp. 11, we could determine whether consideration of bidirectional time lags is effective.

Comp. 11: This is a method considering only time lags between the events and their corresponding multiple following tweets. This method detects important scenes based on the same detection model as that used in the proposed method. As with Comp. 11, by using Comp. 12, we could also determine whether consideration of bidirectional time lags is effective.

Comp. 12: This is a method based on [27] using a support vector machine (SVM) [39] for visual and audio features. In order to provide a fair comparison, a one-class SVM [40], which is an unsupervised method, was used instead of a general SVM for Comp. 9 in this experiment. By using Comp. 9, we could evaluate the effectiveness of the use of a GAN for important scene detection.

Comp. 13: This is a method using LSTM as an anomaly detection method. This method uses BiTl-dMCCA as the embedding scheme. Comp. 13 was used to evaluate the effectiveness of utilization of a GAN for important scene detection.

Comp. 14: This is a method based on another video summarization method [41]. This method extracts visual features from shots based on a CNN and generates a summary based on Bi-LSTM [16] using visual features.

Comp. 15: This is a method using textual, visual and audio features based on [41].

(**Comparison results**) Comparison results are shown in TABLE 2.

[5] https://tv.pacificleague.jp/ptv/pc/

TABLE 2. Specificity of important scene detection in Ours and Comps. 1–15.

Matches	Ours	Comp. 1	Comp. 2	Comp.3	Comp. 4	Comp. 5	Comp. 6	Comp. 7
1	**0.376**	0.345	0.345	0.350	0.345	0.345	0.345	0.345
2	0.417	0.390	0.390	0.297	0.390	0.390	0.390	0.390
3	**0.467**	0.449	0.449	0.461	0.449	0.449	0.449	0.449
4	**0.328**	0.301	0.301	0.301	0.301	0.301	0.301	0.301
5	**0.421**	0.337	0.336	0.336	0.411	0.409	0.402	0.337
Average	**0.402**	0.365	0.364	0.369	0.379	0.379	0.378	0.365

Matches	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12	Comp. 13	Comp. 14	Comp. 15
1	0.360	0.352	0.360	0.360	0.361	0.370	0.350	0.360
2	0.406	0.390	0.406	0.412	0.407	**0.418**	0.390	0.406
3	0.462	0.457	0.462	0.462	0.463	0.463	0.449	0.462
4	0.295	0.317	0.308	0.295	0.294	0.302	0.295	0.301
5	0.333	0.409	**0.421**	0.412	0.325	0.413	0.325	0.373
Average	0.371	0.385	0.391	0.388	0.370	0.393	0.362	0.380

(**Baselines**) In this experiment, we regarded each at-bat as important when at least 80% of its length was detected as an important scene. In Comps., the above percentage was set to the optimal value in such a way that their performance became the highest. Note that we define highlights of matches as ground truth.

Also, d_p, d_y, L, λ, C, W, S, α and τ were empirically set to 500, 50, 7, 3, 5, 60, 20, 0.5 and 0.7, respectively. BiTl-dMCCA was used for projecting features into the same space. In this experiment, the dimension was experimentally reduced from 500 to 50. In the test phase, the proposed method takes 10,347 seconds for a total of 56,760 seconds of test data. Specifically, it takes 9,777 seconds for multi-modal feature extraction, 2 seconds for BiTl-dMCCA, and 568 seconds for detection of important scenes.

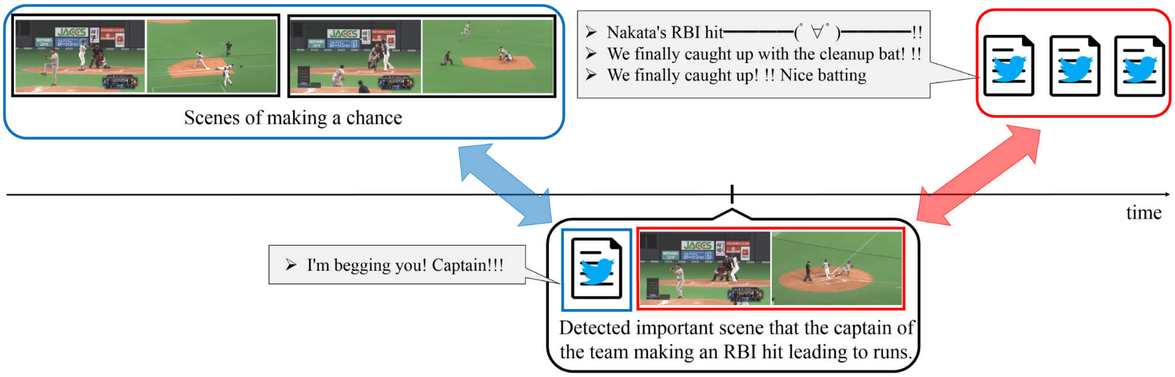
Comps. 1-11 were used to evaluate the first novel idea of introducing consideration of the bidirectional time lags between tweets and events into the derivation of covariance matrices of BiTl-dMCCA. Comps. 12 and 13 were used to evaluate the second novel idea of performing detection in a completely unsupervised fashion based on GAN-based anomaly detection.

For quantitative evaluation, we compared the proposed method with the comparative methods by using the specificity when maximizing the sensitivity (i.e., sensitivity being almost 1.0). Focusing on the specificity when maximizing the sensitivity means how much over-detection of normal scenes can be reduced when important scenes are detected.

B. PERFORMANCE EVALUATION

The specificity when maximizing the sensitivity of important scene detection in our method (Ours) and Comps. 1-15 are shown in TABLE 2. In order to maximize the sensitivity of important scene detection, all important scenes must be detected. Thus, it is difficult to obtain high specificity when maximizing the sensitivity. Since Ours has a higher average specificity than those of Comps. 1-6, the effectiveness of introducing multi-modal features, i.e., textual, visual and audio features, can be confirmed. Therefore, we can confirm that utilizing both tweets and videos is effective for detection of important scenes. Comp. 9 has a higher average specificity than those of Comps. 4-6, but those of Comps. 7 and 8 are lower than those of Comps. 4-6. Therefore, the specificity is not necessarily improved by simply utilizing multi-modal features. On the other hand, by comparing Ours, Comps. 11 and 13 with Comp. 9, which is a method using dMCCA, we can confirm the effectiveness of introducing multi-modal features with consideration of time lags. Since the average specificity of Ours is higher than those of Comps. 10 and 11, which are methods considering unidirectional time lags, the effectiveness of introducing covariance matrices considering bidirectional time lags can be confirmed. Furthermore, since the average specificity of Ours is higher than that of Comp. 12, which is a traditional important scene detection method based on an SVM, and that of Comp. 13 based on LSTM, we can confirm the effectiveness of introducing a GAN to detect important scenes. Ours has higher specificity than the specificity of Comp. 14. The specificity of Comps. 2 and 14 are about the same. Thus, the specificity of the methods using visual features are limited in important scene detection, and the effectiveness of Ours using textual, visual and audio features has been confirmed. Moreover, since the specificity of Comps. 14 and 15 are lower than those of Comps. 2 and 9, respectively, we can confirm that the methods based on Ours are more effective than the methods based on [41] for detection of important scenes. Consequently, accurate detection of important scenes in baseball videos by anomaly detection via BiTl-dMCCA considering bidirectional time lags between multi-modal features based on a GAN has been realized.

An example of a correctly detected important scene by Ours and its tweets corresponding to the scene is shown in FIGURE 4. Scenes surrounded by the blue rectangle represent batters making chances. The main text of the tweet surrounded by the blue rectangle includes expectations of the viewers who saw these scenes of making chances. Furthermore, the important scene detected by Ours surrounded by the red rectangle represents the captain of the team making an RBI hit. The main text of the tweets surrounded by the


Scenes of making a chance
time
Detected important scene that the captain of the team making an RBI hit leading to runs.

- ➢ Nakata's RBI hit———(ﾟ∀ﾟ)———!!
- ➢ We finally caught up with the cleanup bat! !!
- ➢ We finally caught up! !! Nice batting
- ➢ I'm begging you! Captain!!!

FIGURE 4. An example of an important scene detected by Ours and its tweets corresponding to the scene. The horizontal axis represents time. Note that "Nakata" in the main text of tweets is the batter's name and he is the captain of the team. A run batted in (RBI) hit is a hit by the batter who scored a run.

TABLE 3. Specificity of important scene detection in Ours for each parameter λ of the Poisson distribution.

	A parameter λ of the Poisson distribution			
Matches	1	3	5	7
1	0.360	**0.376**	**0.376**	0.360
2	0.406	**0.417**	0.406	0.406
3	0.462	**0.467**	0.462	0.462
4	0.304	**0.328**	0.307	0.295
5	0.412	**0.421**	**0.421**	0.412
Average	0.389	**0.402**	0.394	0.387

TABLE 4. Specificity of important scene detection in Ours for each parameter *L* of the Poisson distribution.

	A parameter L of the Poisson distribution			
Matches	3	5	7	9
1	0.360	0.360	**0.376**	**0.376**
2	0.390	0.406	**0.417**	0.412
3	0.462	0.457	**0.467**	0.462
4	0.295	0.301	**0.328**	0.317
5	0.402	**0.421**	**0.421**	0.412
Average	0.382	0.389	**0.402**	0.396

red rectangle includes the delights of viewers who saw this important scene. Therefore, there obviously exists bidirectional time lags between posted tweets and corresponding events. From the qualitative evaluation, we confirmed that Ours can accurately detect these important scenes by BiTl-dMCCA.

The results of Ours for each parameter λ of the Poisson distribution are shown in TABLE 3. It is shown in this table how much the peak of the distribution should be slid. In other words, we can understand the relationship of the time lags between the posted tweets and events that occurred. By comparing the specificity when changing the parameter λ, we can confirm that the highest specificity is achieved when λ is three. Moreover, the results of Ours for each parameter λ of the Poisson distribution are shown in TABLE 4. It is shown in this table how much previous events affect the tweets. By comparing the specificity when changing the parameter L, we can confirm that the highest specificity is achieved when L is seven. Since the viewers post tweets of test data about every 24 seconds on average, the results suggest that the bidirectional time lags between tweets and events are about 72 seconds and that events up to 168 seconds in the past affect the tweets. From the above discussion, we can consider that calculation using the parameters λ and L of the Poisson distribution can be an effective indicator to reveal bidirectional time lags.

IV. CONCLUSION

In this paper, we have presented a new method for detection of important scenes in baseball videos based on canonical correlation maximization with consideration of bidirectional time lags via BiTl-dMCCA. By introducing consideration of the bidirectional time lags between tweets and events into the derivation of covariance matrices of BiTl-dMCCA, it can correctly consider their relationships. Furthermore, the proposed method can be performed in a completely unsupervised fashion based on GAN-based anomaly detection. Experimental results verified that the proposed method can correctly detect important scenes.

In a future work, we will automatically decide the parameter λ according to each event. Specifically, events such as a home run and an RBI hit strongly affect tweets that immediately follow. On the other hand, events such as the appearance of popular players have a long-term effect on tweets. Therefore, the construction of covariance matrices reflecting these differences should be considered in a future work.

REFERENCES

[1] P. Shukla, H. Sadana, A. Bansal, D. Verma, C. Elmadjian, B. Raman, and M. Turk, "Automatic cricket highlight generation using event-driven and excitement-based features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1800–1808.

[2] Y. Takahashi, N. Nitta, and N. Babaguchi, "Video summarization for large sports video archives," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2005, pp. 1170–1173.

[3] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 982–990.

[4] C.-C. Cheng and C.-T. Hsu, "Fusion of audio and motion information on HMM-based highlight extraction for baseball games," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 585–599, Jun. 2006.

[5] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden Markov models," in *Proc. Int. Conf. Image Process.*, vol. 1, 2002, p. 1.

[6] Y. Gong, M. Han, W. Hua, and W. Xu, "Maximum entropy model-based baseball highlight detection and classification," *Comput. Vis. Image Understand.*, vol. 96, no. 2, pp. 181–199, Nov. 2004.

[7] M. Nakazawa, M. Erdmann, K. Hoashi, and C. Ono, "Social indexing of TV programs: Detection and labeling of significant TV scenes by Twitter analysis," in *Proc. 26th Int. Conf. Adv. Inf. Netw. Appl. Workshops*, Mar. 2012, pp. 141–146.

[8] K. Doman, T. Tomita, I. Ide, D. Deguchi, and H. Murase, "Event detection based on Twitter enthusiasm degree for generating a sports highlight video," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 949–952.

[9] L.-C. Hsieh, C.-W. Lee, T.-H. Chiu, and W. Hsu, "Live semantic sport highlight detection based on analyzing tweets of Twitter," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 949–954.

[10] S. Jai-Andaloussi, A. Mohamed, N. Madrane, and A. Sekkaki, "Soccer video summarization using video content analysis and social media streams," in *Proc. IEEE/ACM Int. Symp. Big Data Comput.*, Dec. 2014, pp. 1–7.

[11] E. K. Seltzer, N. S. Jean, E. Kramer-Golinkoff, D. A. Asch, and R. M. Merchant, "The content of social media's shared images about ebola: A retrospective study," *Public Health*, vol. 129, no. 9, pp. 1273–1277, Sep. 2015.

[12] N. D. Doulamis, A. D. Doulamis, P. Kokkinos, and E. M. Varvarigos, "Event detection in Twitter microblogging," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2810–2824, Dec. 2016.

[13] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, and A. Jaimes, "Sensing trending topics in Twitter," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1268–1282, Oct. 2013.

[14] Z. Cheng, X. Chang, L. Zhu, R. C. Kanjirathinkal, and M. Kankanhalli, "MMALFM: Explainable recommendation by leveraging reviews and images," *ACM Trans. Inf. Syst.*, vol. 37, no. 2, pp. 1–28, Mar. 2019.

[15] L. Zhu, X. Lu, Z. Cheng, J. Li, and H. Zhang, "Flexible multi-modal hashing for scalable multimedia retrieval," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 2, pp. 1–20, Mar. 2020.

[16] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[17] K. Somandepalli, N. Kumar, R. Travadi, and S. Narayanan, "Multimodal representation learning using deep multiset canonical correlation," 2019, *arXiv:1904.01775*. [Online]. Available: http://arxiv.org/abs/1904.01775

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[19] A. D. Doulamis, N. D. Doulamis, and S. D. Kollias, "A fuzzy video content representation for video summarization and content-based retrieval," *Signal Process.*, vol. 80, no. 6, pp. 1049–1067, Jun. 2000.

[20] K. Hirasawa, K. Maeda, T. Ogawa, and M. Haseyama, "Mvgan maximizing time-lag aware canonical correlation for baseball highlight generation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2020, pp. 1–6.

[21] B. Dhingra, Z. Zhou, D. Fitzpatrick, M. Muehl, and W. W. Cohen, "Tweet2Vec: Character-based distributed representations for social media," 2016, *arXiv:1605.03481*. [Online]. Available: http://arxiv.org/abs/1605.03481

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[24] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*. [Online]. Available: http://arxiv.org/abs/1705.06950

[25] E. A. Hadhrami, M. A. Mufti, B. Taha, and N. Werghi, "Ground moving radar targets classification based on spectrogram images using convolutional neural networks," in *Proc. 19th Int. Radar Symp. (IRS)*, Jun. 2018, pp. 1–9.

[26] 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," in *Proc. Interspeech*, Aug. 2017, pp. 3512–3516.

[27] T. Haruyama, S. Takahashi, T. Ogawa, and M. Haseyama, "Estimation of important scenes in soccer videos based on collaborative use of audio-visual CNN features," in *Proc. IEEE 7th Global Conf. Consum. Electron. (GCCE)*, Oct. 2018, pp. 710–711.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[30] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.

[31] L. C. Parra, S. Haufe, and J. P. Dmochowski, "Correlated components analysis–extracting reliable dimensions in multivariate data," 2018, *arXiv:1801.08881*. [Online]. Available: http://arxiv.org/abs/1801.08881

[32] M. A. Hasan, "On multi-set canonical correlation analysis," in *Proc. Int. Joint Conf. Neural Netw.*, Jun. 2009, pp. 1128–1133.

[33] K. Maeda, Y. Ito, T. Ogawa, and M. Haseyama, "Supervised fractional-order embedding geometrical multi-view CCA (SFGMCCA) for multiple feature integration," *IEEE Access*, vol. 8, pp. 114340–114353, 2020.

[34] J. de Leeuw, "Derivatives of generalized eigen systems with applications," Dept. Statist., Univ. California, Los Angeles, CA, USA, Tech. Rep., 2007.

[35] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[36] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, "MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks," in *Proc. Int. Conf. Artif. Neural Netw.*, 2019, pp. 703–716.

[37] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *J. Roy. Stat. Soc. C, Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.

[38] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, Dec. 1936.

[39] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[40] Y. Chen, X. Sean Zhou, and T. S. Huang, "One-class SVM for learning in image retrieval," in *Proc. Int. Conf. Image Process.*, vol. 1, 2001, pp. 34–37.

[41] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, "Cloud-assisted multiview video summarization using CNN and bidirectional LSTM," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 77–86, Jan. 2020.

KAITO HIRASAWA (Student Member, IEEE) received the B.S. degree in electronics and information engineering from Hokkaido University, Japan, in 2020, where he is currently pursuing the M.S. degree with the Graduate School of Information Science and Technology. His research interest includes sports video analysis.

KEISUKE MAEDA (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Japan, in 2015, 2017, and 2019, respectively. He is currently a specially appointed Assistant Professor with the Office of Institutional Research, Hokkaido University. His research interests include multimodal signal processing and machine learning and its applications. He is a member of the IEICE.

TAKAHIRO OGAWA (Senior Member, IEEE) received the 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 the Graduate School of Information Science and Technology, Hokkaido University, in 2008. He is currently an Associate Professor with the Faculty of Information Science and Technology, Hokkaido University. His research interests include AI, the IoT, and big data analysis for multimedia signal processing and its applications. He is a member of ACM, IEICE, and ITE. He was the Special Session Chair of IEEE ISCE, in 2009, the Doctoral Symposium Chair of ACM ICMR, in 2018, the Organized Session Chair of IEEE GCCE, from 2017 to 2019, the TPC Vice Chair of IEEE GCCE, in 2018, the Conference Chair of IEEE GCCE, in 2019, and so on. He has been an Associate Editor of *ITE Transactions on Media Technology and Applications*.

MIKI HASEYAMA (Senior Member, IEEE) received the 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 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 Fellow of ITE and a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and ASJ. She has been the Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE) and the Director of the International Coordination and Publicity of IEICE. She has also been the Editor-in-Chief of *ITE Transactions on Media Technology and Applications*.