IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 26, 2019, accepted July 12, 2019, date of publication July 24, 2019, date of current version August 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2930713

Music Video Recommendation Based on Link Prediction Considering Local and Global Structures of a Network

YUI MATSUMOTO[1], (Student Member, IEEE), RYOSUKE HARAKAWA[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]Department of Electrical, Electronics and Information Engineering, Nagaoka University of Technology, Nagaoka 940-2188, Japan
[3]Faculty of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan

Corresponding author: Yui Matsumoto (matsumoto@lmd.ist.hokudai.ac.jp)

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

ABSTRACT A novel method for music video recommendation is presented in this paper. The contributions of this paper are two-fold. (i) The proposed method constructs a network, which not only represents relationships between music videos and users but also captures multi-modal features of music videos. This enables collaborative use of multi-modal features such as audio, visual, and textual features, and multiple social metadata that can represent relationships between music videos and users on video hosting services. (ii) A novel scheme for link prediction considering local and global structures of the network (LP-LGSN) is newly derived by fusing multiple link prediction scores based on both local and global structures. By using the LP-LGSN to predict the degrees to which users desire music videos, the proposed method can recommend users' desired music videos. The experimental results for a real-world dataset constructed from YouTube-8M show the effectiveness of the proposed method.

INDEX TERMS Music video, recommendation, link prediction, network analysis, social metadata.

I. INTRODUCTION

On video hosting services, music videos[1] have been watched actively by a large number of users [1], [2]. In fact, on YouTube,[2] 23% of all uploaded videos are music videos [3], [4], and 86% of queries are words related to music [5]. On existing video hosting services, users input queries and receive videos related to the queries. In other words, users must verbalize which music videos they desire. Therefore, these video hosting services are not enough to satisfy users who watch music videos without the clear desires. Since users' desires depend on multiple factors about the users (e.g., personality, emotions or social environment) and music videos (e.g., musical pieces, videos or lyrics), there is a case in which users cannot verbalize their own desires. In such situation, recommender systems, which can provide music videos corresponding to automatically predicted users' preference, are useful.

To the best of our knowledge, however, there have been few works that target music videos to solve the above problem. Methods for musical piece recommendation [6]–[34] and video recommendation [35]–[61] can be applied to realize music video recommendation. In these methods, approaches based on matrix factorization (MF) [62], probabilistic models [63] and deep learning [64] have been widely adopted. These approaches enable accurate analysis of users' operation histories (e.g., listening histories and bookmarks) and features obtained from musical pieces or videos. When we focus on musical pieces and videos on the Web, approaches based on network analysis [24]–[34], [56]–[61] have been adopted. In these methods, a network for which nodes include musical pieces/videos and users is constructed by using social metadata as well as features of musical pieces or videos. It has been reported that the network representation based on social metadata as well as such features is effective for

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

[1]We call videos with attached musical pieces "music videos." Music videos consist of musical pieces, moving pictures, and texts such as description.

[2]https://www.youtube.com/

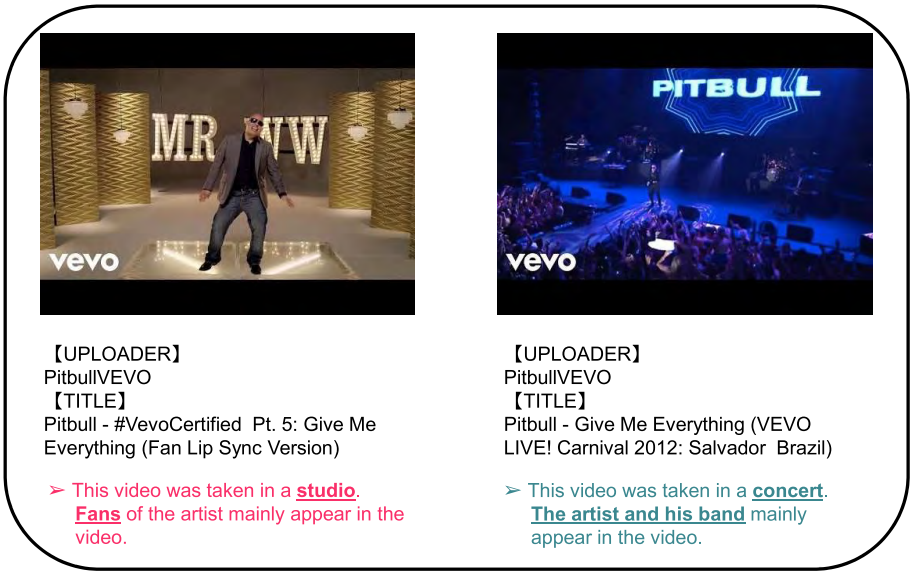


FIGURE 1. Example of two music videos that include the same musical piece but different moving pictures.

estimating local similarities between musical pieces/videos and users [65], [66]. Therefore, these methods realize accurate recommendation by analyzing the local similarities on the network. However, these methods [24]–[34], [56]–[61] still have the following two difficulties.

(Problem I)
Conventional methods, which target only musical pieces or videos, do not capture multi-modal features including audio, visual and textual features. Thus, these methods cannot accurately distinguish music videos by utilizing only a part of multi-modal features. For example, if two music videos include the same musical piece but different moving pictures, we cannot distinguish them by using only audio features (see Fig. 1). This is because these methods cannot realize direct comparison of music videos with users on the same space, which enables construction of a network.

(Problem II)
Conventional methods search musical pieces or videos on the network for recommendation based on only local similarities. By using local similarities, these methods can measure which music videos are neighbor on the network. However, these methods cannot recommend musical pieces or videos that are not neighbor on the network but correspond to users' preference. To reduce the ill influence of using local similarities, we should use not only local similarities but also global similarities. The global similarities can measure which music videos have the same role on the network. However, network analysis of the conventional methods cannot realize simultaneous utilization of the local and global similarities.

To solve the above problems, we propose a novel method for music video recommendation in this paper. The challenging point of this paper is solutions of both (Problem I) and (Problem II). To solve (Problem II), (Problem I) must be solved in advance. When analyzing a network for which nodes are music videos and users, relationships between nodes must be accurately defined. However, as mentioned in (Problem I), conventional methods cannot realize direct comparison of music videos and users. Therefore, conventional methods cannot define relationships between nodes and network analysis of conventional methods cannot work well. In the proposed method, by constructing a network that can solve (Problem I), we can perform accurate network analysis that can solve (Problem II). The contributions of this paper are two-fold.

(Contribution i)
We have constructed a network for which nodes are music videos and users and links represent their relationships. The proposed network is constructed by utilizing novel features that can represent latent relationships among multi-modal features. Thus, we can solve (Problem I).

(Contribution ii)
We have developed a novel scheme called link prediction considering local and global structures of a network (LP-LGSN). By using not only local similarities but also global similarities, we can recommend music videos that cannot be recommended by conventional methods. Thus, the proposed method can solve (Problem II).

In **(Contribution i)**, we apply sub-sampled canonical correlation analysis (CCA) [67] to audio, visual and textual features. This enables extraction of latent features that realize direct comparison of music videos with users on the

same latent space. By collaboratively using the obtained latent features and social metadata such as ''related videos,'' ''tags,'' ''uploaders,'' and ''keywords,'' we can construct a network that enables direct comparison of music videos with users. In **(Contribution ii)**, we fuse weighted Jaccard coefficients [68] and node2vec [69]-based similarities through empirical distribution functions [70]. By this fusion, LP-LGSN enables extraction of pairs of unlinked nodes for which degrees of local and global similarities are both high. Since LP-LGSN can estimate relevance between unlinked nodes accurately, we can predict which music videos users desire more accurately. As a result, successful music video recommendation becomes feasible.

The preliminary version of this work can be found in a conference paper [71]. We have improved our previous method [71] in the following two aspects.

- While our previous method extracts audio and textual features from music videos, the proposed method extracts visual features as well for more accurate recommendation.
- While our previous method performs link prediction that only utilizes node2vec-based similarities, the proposed method enables more accurate link prediction by using LP-LGSN that fuses multiple link prediction scores.

The rest of this paper is organized as follows. In Section II, related works of the proposed method are explained. In Section III, a method for construction of a network via sub-sampled CCA and music video recommendation via LP-LGSN are explained. In Section IV, results of experiments for real-world music videos obtained from a public dataset, YouTube-8M [72] are presented to verify the effectiveness of our method. Conclusions are given in Section V.

II. RELATED WORKS

The purpose of this work is to realize music video recommendation. To the best of our knowledge, there have been few works that target music video recommendation. Thus, we explain existing methods for musical piece recommendation and video recommendation that can be applied to music video recommendation. At the end of this section, we discuss the limitations of these methods for music video recommendation.

A. MUSICAL PIECE RECOMMENDATION AND VIDEO RECOMMENDATION

Many methods for recommending musical pieces [6]–[23] and videos [35]–[55] have been proposed.

Conventionally, MF that can estimate users' preference from their operation histories has been proposed [9]–[11], [36]–[43]. Specifically, singular value decomposition (SVD), for which the input is a user-item rating matrix and output is latent factors of users' ratings [11], [39], was utilized for musical piece recommendation. For video recommendation, various MF-based models were proposed [36]–[43]. Specifically, Liu *et al.* [40] realized accurate MF of a user-video rating matrix by grouping videos based on content similarities such as similarities of genres or description. Du et al. [41] constructed a general weighted MF-model, collaborative embedding regression, and recommended videos based on late fusion of regression results for multiple features. In the paper [42], implicit feedback-based MF was used to calculate users' latent features from users' access points and users' rating histories. Also, Wu *et al.* [43] proposed a method based on factorization machines (FM) that are constructed with users' listening histories, users' attributes (e.g., moods) and videos' attributes (e.g., popularity). By using the constructed FM, this method can predict which context information works well for recommendation. However, if we cannot obtain abundant users' ratings or other operation histories and a user-item matrix is sparse, MF cannot work well for accurate recommendation.

To solve this problem, probabilistic models have also been widely utilized [13]–[18], [44]–[49]. These models can estimate probabilities of users desiring each content [63]. Specifically, latent Dirichlet allocation (LDA) to capture factors of users' preference by using tags of musical pieces was adopted for musical piece recommendation [15]. Cheng et al. [17] developed LDA and constructed a novel topic model to capture both users' long-term preference and short-term needs from audio features, popularity and listening histories. Also, they proposed location-aware topic model (LTM) [18] to capture relevance between listening location and audio features. Meanwhile, Jin *et al.* [16] constructed a Markov embedding model that predicts degrees of users' desires for musical pieces by using information about playlists. For video recommendation, LDA was widely utilized for modeling semantic topics of visual features of videos and textual features of description on Web pages [46], [47]. Also, multi-site probabilistic factorization, which can capture users' preference from users' view counts of multiple video hosting services, was proposed [48].

When we can prepare a large amount of training data, methods based on deep learning [73] are useful for accurate recommendation [19]–[23], [50]–[55]. Specifically, deep convolutional neural networks (CNN) were adopted by training with tens or hundreds of thousands of audio features and labels [22], [23]. These methods can predict latent factors of acoustic characteristics. For video recommendation, fine-tuned neural networks that embed videos into the space where similar videos are located close to each other by using audio and visual features were adopted [52]. Also, deep belief networks for which the input is a user-item matrix and output is representation of users' interest were adopted [55]. Ma et al. [53] recommended videos by constructing a neural network whose inputs were CNN-based visual features, recurrent neural network (RNN)-based textual features and metadata. Dang et al. [54] performed data argumentation to multiple pre-trained CNN models to improve prediction of relationships of videos by using visual features of videos.

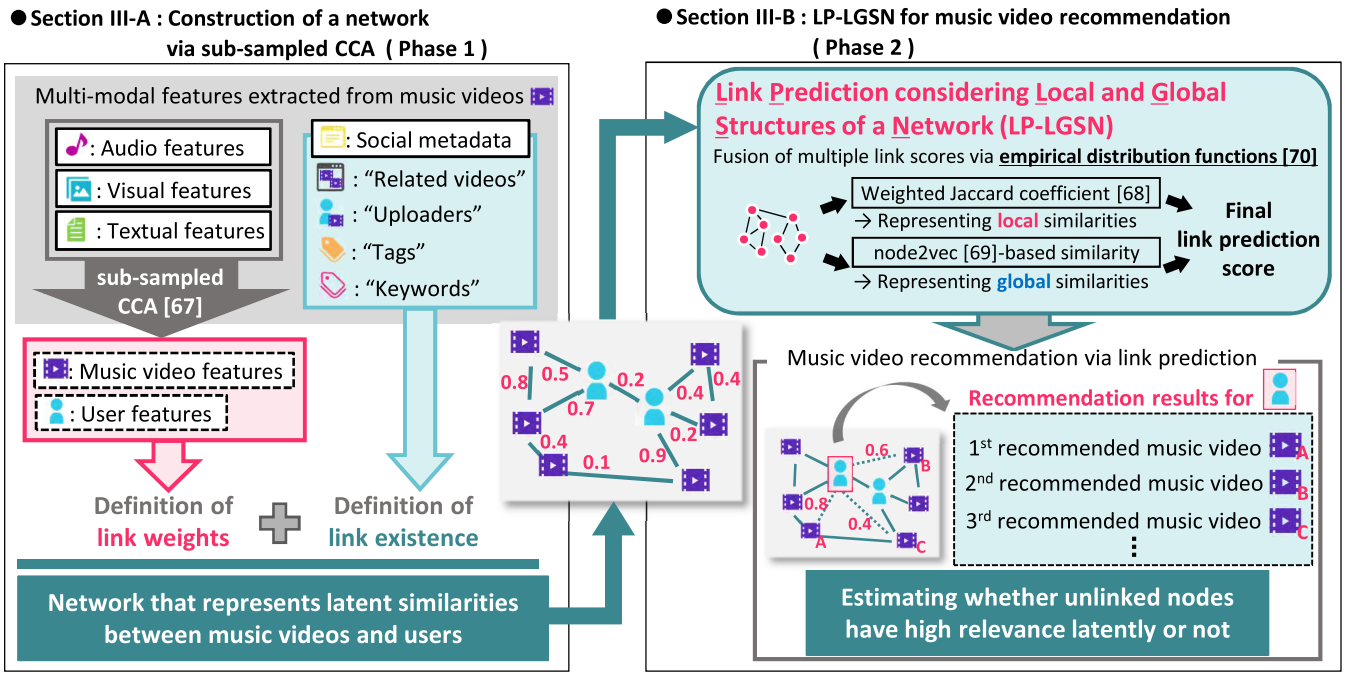


FIGURE 2. Overview of the proposed method for music video recommendation based on LP-LGSN.

B. MUSICAL PIECE RECOMMENDATION AND VIDEO RECOMMENDATION BASED ON NETWORK ANALYSIS

Approaches based on network analysis have attracted much attention for analysis of contents on the Web. These approaches construct a network for which nodes are the contents and links are defined by social metadata; thus, such network analysis can estimate latent similarities between contents [65], [66]. Network analysis has been widely adopted for musical piece recommendation [24]–[34] and video recommendation [56]–[61]. The most-widely-used approach based on network analysis is construction of a graph regularization model that can rank music videos accurately [27]–[32]. Specifically, a hypergraph that can capture interactions among musical pieces, users and tags was analyzed for musical piece recommendation [27], [28]. Also, methods in the papers [29], [30] construct heterogeneous networks for which nodes are musical pieces, users and other information (e.g., albums, genres or playing sequences) by using users' operation histories. The method in the paper [31] performs graph regularization with total variation for accurately ranking musical pieces.

Methods that use basic technologies of network analysis such as graph clustering [33], [57], [58], random walk [33], [59], community detection [60], [61] and graph embedding [34], [61] have been also proposed. Pang et al. [58] realize accurate video recommendation by graph clustering that uses visual features. Also, Shang et al. [57] constructed a graph based on Web page links and performed clustering based on Jaccard coefficients for video recommendation. Furthermore, the method in [59] realizes accurate recommendation by learning similarities of the network structure via CNN. The method in [33] is based on the combination of random walk and clustering via affinity propagation. By constructing a graph for which nodes are users and capturing users' social relationships via community detection, methods in [60], [61] realize accurate video recommendation. The method in the paper [61] also performs graph embedding after community detection and recommends videos based on improved k-nearest neighbor (k-NN) of embedded features.

C. PROBLEMS TO BE SOLVED IN THIS WORK

We have introduced existing methods for musical piece recommendation and video recommendation in the above sections. However, these methods have two difficulties. First, these methods do not have a framework to directly compare music videos with users, which needs to realize music video recommendation. To realize direct comparison, multi-modal features of music videos should be fully utilized for representing music videos accurately. Second, conventional methods, which only utilize local similarities, cannot recommend musical pieces or videos that are not neighbor on the network but correspond to users' preference. To reduce the ill influence of utilizing local similarities, we should realize simultaneous utilization of the local and global similarities. In this paper, we present an alternative method that can simultaneously solve the above problems.

III. MUSIC VIDEO RECOMMENDATION BASED ON LP-LGSN

As shown in Fig. 2, our method consists of two phases. A method for construction of a network via sub-sampled CCA (Phase 1) is explained in Section III-A and a method for music

TABLE 1. Metadata used for building links of a network that associates music videos and users.

Metadata	Description
Related videos	Metadata for associating music videos related to each other
Uploaders	Metadata identifying who uploads each music video
Tags	Metadata annotated by uploaders, which represent the summary of music video contents
Keywords	Metadata annotated by uploaders to show their own information (e.g., genres related to their uploaded videos)

video recommendation via LP-LGSN (Phase 2) is explained in Section III-B.

A. CONSTRUCTION OF A NETWORK VIA SUB-SAMPLED CCA (PHASE 1)

According to reports showing that network analysis is effective for estimating latent similarities between contents on social media [65], [66], [74], we construct a network for which nodes are music videos and users. Here, we denote music videos by $c^{(i)}$ ($i = 1, 2, \ldots, N_c$; N_c being the number of music videos) and we denote users by $u^{(j)}$ ($j = 1, 2, \ldots, N_u$; N_u being the number of users). We also define the network as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of music videos $c^{(i)}$ and users $u^{(j)}$, and $\mathcal{E}$ is the set of links.

First, we define links by social metadata shown in Table 1. We adopted them according to a report that "related videos," "uploaders" and "tags" are effective for associating contents similar to each other on video hosting services [66], [74]–[76]. Since contents of "related videos" are similar to each other, we can associate music videos whose contents are similar by associating music videos based on "related videos." In addition, it is expected that users' preference is reflected in music videos that have been already associated to users. Therefore, "uploaders" are effective for associating music videos with users. Moreover, it is expected that "tags" and "keywords" can represent which music videos or users have related contents, respectively. Thus, such metadata are effective for associating music videos or users. Specifically, we build links $f_{p^{(x)},q^{(y)}}$ ($p^{(x)}, q^{(y)} \in \{\mathcal{S}_c, \mathcal{S}_u\}$; $\mathcal{S}_c$ and $\mathcal{S}_u$ being the set of music videos and users, respectively) if the following requirements are satisfied.

Case where $(p^{(x)}, q^{(y)}) = (c^{(x)}, c^{(y)})$:
"Related videos" of $p^{(x)}$ include $q^{(y)}$ and vice versa or one or more "tags" of $p^{(x)}$ and $q^{(y)}$ are the same.

Case where $(p^{(x)}, q^{(y)}) = (c^{(x)}, u^{(y)})$:
$q^{(y)}$ is an "uploader" of $p^{(x)}$.

Case where $(p^{(x)}, q^{(y)}) = (u^{(x)}, u^{(y)})$:
One or more "keywords" of $p^{(x)}$ and $q^{(y)}$ are the same.

Note that we do not build self-loops.

Next, we define link weights based on similarities between music videos and users by using multi-modal features of music videos via sub-sampled CCA [67]. For accurate estimation of similarities, we need to project features obtained from music videos so that music videos and users can be directly compared to each other. In addition, we need to calculate such projection efficiently to deal with a large number of music videos on real-world video hosting services. For these reasons, we adopt sub-sampled CCA, which enables efficient projection of heterogeneous features into the same latent space. Concretely, we extract an audio feature $\boldsymbol{v}_a^{(i)}$, a visual feature $\boldsymbol{v}_v^{(i)}$, and a textual feature $\boldsymbol{v}_t^{(i)}$ from $c^{(i)}$. Then we select a small number of features that approximate distribution of the whole features from these features $\boldsymbol{v}_\zeta^{(i)}$ ($\zeta \in \{a, v, t\}$) via k-means clustering [77]. In order to improve scalability, we apply CCA to only the selected features. By the above computation, we can obtain matrices $\boldsymbol{U}_\zeta$ that enables projection of audio, visual and textual features into the feature space where these features can be directly compared. Thus, we can calculate latent features $\hat{\boldsymbol{v}}_\zeta^{(i)}$ as $\boldsymbol{U}_\zeta^{\mathrm{T}} \boldsymbol{v}_\zeta^{(i)}$. Second, we calculate a music video feature $\hat{\boldsymbol{v}}_c^{(i)}$ of $c^{(i)}$ and a user feature $\hat{\boldsymbol{v}}_u^{(j)}$ of $u^{(j)}$ with a latent feature $\hat{\boldsymbol{v}}_\zeta^{(i)}$ as follows:

$$\hat{\boldsymbol{v}}_c^{(i)} = [(\hat{\boldsymbol{v}}_v^{(i)})^{\mathrm{T}} (\hat{\boldsymbol{v}}_a^{(i)})^{\mathrm{T}} (\hat{\boldsymbol{v}}_t^{(i)})^{\mathrm{T}}]^{\mathrm{T}}, \tag{1}$$

$$\hat{\boldsymbol{v}}_u^{(j)} = \frac{1}{|C(u^{(j)})|} \sum_{c^{(i)} \in C(u^{(j)})} \hat{\boldsymbol{v}}_c^{(i)}, \tag{2}$$

where $C(u^{(j)})$ is the set of music videos uploaded by $u^{(j)}$. In Eq. (1), we can accurately represent semantic contents of music videos, which are defined by multiple factors such as musical pieces or videos, by concatenating the obtained features. Moreover, in Eq. (2), we can represent a user as a feature that can integrate features of multiple music videos that are related to the user. In other words, we can calculate the features of all users in the same computation even if the number of music videos are different depending on each user. By the above computation, music videos and users can be directly compared to each other. Finally, we define link weights $w(p^{(x)}, q^{(y)})$ of $f_{p^{(x)},q^{(y)}}$ based on similarities of the music video features and the user features. In the proposed method, we define similarities based on the following equation, which is effective for constructing a network for which nodes are contents on social media [78], [79].

$$w(p^{(x)}, q^{(y)}) = \left| \frac{\hat{\boldsymbol{v}}_p^{(x)\mathrm{T}} \hat{\boldsymbol{v}}_q^{(y)}}{\|\hat{\boldsymbol{v}}_p^{(x)}\| \, \|\hat{\boldsymbol{v}}_q^{(y)}\|} \right|. \tag{3}$$

By collaborative use of multi-modal features and multiple social metadata, the proposed method enables construction of a network that can represent latent similarities between music videos and users.

B. LP-LGSN FOR MUSIC VIDEO RECOMMENDATION (PHASE 2)

We newly derive LP-LGSN through fusion of multiple link prediction scores based on similarities of both local and global structures of the obtained network $\mathcal{G}$. We define $e(n^{(m)}, n^{(l)})$ as a link between nodes $n^{(m)}$ and $n^{(l)} \in \mathcal{V}$.

First, we calculate local similarities of $\mathcal{G}$. In the proposed method, we adopt weighted Jaccard coefficients [68] since the coefficients can extract similarities of nodes that are neighbor on a target node. Specifically, weighted Jaccard coefficients [68] $L_{\mathrm{jac}}[e(n^{(m)}, n^{(l)})]$ are defined as follows:

$$L_{\mathrm{jac}}[e(n^{(m)}, n^{(l)})] = \sum_{r \in \Gamma(n^{(m)}) \cap \Gamma(n^{(l)})} \frac{w(n^{(m)}, r) + w(n^{(l)}, r)}{\sum_{n'^{(m)} \in \Gamma(n^{(m)})} w(n'^{(m)}, n^{(m)}) + \sum_{n'^{(l)} \in \Gamma(n^{(l)})} w(n'^{(l)}, n^{(l)})}, \tag{4}$$

where $\Gamma(n^{(m)})$ is a set of neighbors of $n^{(m)}$. Since weighted Jaccard coefficients represent the proportion of common neighbors, the local structure of a network can be represented.

Second, we calculate global similarities of $\mathcal{G}$. In the proposed method, we apply node2vec [69] to $\mathcal{G}$ in order to obtain node features. Node2vec can extract similarities of nodes that are not only neighbor on a target node but also distant from it. Therefore, we can extract pairs of nodes that are distant from each other but similar to each other. Specifically, we calculate node2vec-based similarities $L_{\mathrm{n2v}}[e(n^{(m)}, n^{(l)})]$ as follows:

$$L_{\mathrm{n2v}}[e(n^{(m)}, n^{(l)})] = \left| \frac{\tilde{\boldsymbol{v}}_n^{(m)\mathrm{T}} \tilde{\boldsymbol{v}}_n^{(l)}}{\|\tilde{\boldsymbol{v}}_n^{(m)}\| \, \|\tilde{\boldsymbol{v}}_n^{(l)}\|} \right|, \tag{5}$$

where $\tilde{\boldsymbol{v}}_n^{(m)}$ is a node feature of $n^{(m)}$. Since node2vec enables distributed representation of nodes based on a random walk [80], the obtained similarities can represent the global structure of a network.

Furthermore, we construct empirical distribution functions of the two kinds of link prediction scores $F_{\mathrm{idx}}(z)$ (idx $\in$ {n2v, jac}) as follows [70]:

$$F_{\mathrm{idx}}(z) = \frac{1}{N_{\mathrm{link}}} \sum_{\xi=1}^{N_{\mathrm{link}}} \mathbb{I}(L_{\mathrm{idx}}[\xi] \leq z). \tag{6}$$

Note that we sort each element of the obtained link prediction scores $L_{\mathrm{idx}}[e(n^{(m)}, n^{(l)})]$ in ascending order and denote them by $L_{\mathrm{idx}}[\xi]$ ($\xi = 1, 2, \cdots, N_{\mathrm{link}}$; N_{link} being the number of all links). In addition, $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the condition is satisfied and 0 otherwise. By the above computation, we equalize the occurrence probability of link prediction scores in a constant interval of the score-axis by building each distribution. Thus, direct comparison of multiple link prediction scores becomes feasible [81].

Finally, we perform link prediction for each node of a user $n^{(m)}$. Here, we define a set of nodes of music videos unlinked to $n^{(m)}$ as $\mathcal{U}^{(m)}$. We fuse multiple link prediction scores of a link between $n^{(m)}$ and $n^{(u)}$ ($n^{(u)} \in \mathcal{U}^{(m)}$) to obtain a final link prediction score $L_{\mathrm{fin}}[e(n^{(m)}, n^{(u)})]$ as follows:

$$L_{\mathrm{fin}}[e(n^{(m)}, n^{(u)})] = \frac{1}{N_{\mathrm{idx}}} \sum_{\mathrm{idx} \in \{\mathrm{n2v,jac}\}} F_{\mathrm{idx}}(L_{\mathrm{idx}}[e(n^{(m)}, n^{(u)})]), \tag{7}$$

where N_{idx} is the number of fused scores ($= 2$). By Eq. (7), LP-LGSN enables extraction of pairs of unlinked nodes for which degrees of local and global similarities are both high. Therefore, the proposed method enables accurate estimation of which music videos and users have high relevance on $\mathcal{G}$ by acquiring node pairs with a high value of $L_{\mathrm{fin}}[e(n^{(m)}, n^{(u)})]$.

By the above computation, we can predict which music videos users desire. Thus, the proposed method can realize recommendation of users' desired music videos by presenting recommendation results in the descending order of $L_{\mathrm{fin}}[e(n^{(m)}, n^{(u)})]$.

IV. EXPERIMENTAL RESULTS

In this section, we present experimental results for real-world music videos by simulating link prediction to verify the effectiveness and limitation of our method for music video recommendation. Experimental settings are explained in Section IV-A. Recommendation performance is evaluated in Section IV-B and the effectiveness of LP-LGSN and performance limitation are described in Section IV-C and IV-D, respectively.

A. SETTINGS

In this experiment, we used YouTube-8M [72] for data collection. YouTube is the best-known video hosting service. In fact, about 90% of the time that users watch music videos on the Web is spent on YouTube [1]. YouTube-8M is the largest dataset of public video datasets [72] and contains a large number of music videos. To the best of our knowledge, YouTube provides the largest amount of information about videos and users among video hosting services. Thus, we collected as much information about real-world music videos and users as possible via YouTube-8M. Specifically, we first selected users who have 10 or more videos whose "entities"[3] are "Music video" and collected them. Thus, a dataset containing 10331 music videos and 473 users was constructed. Then we used audio features provided by YouTube-8M [82]. These audio features were calculated based on CNNs that can recognize musical instrument and vocals in videos. Also, we used CNN-based features that can capture objects in videos as visual features, which are provided by YouTube-8M [72]. In addition, we obtained textual features by applying doc2vec [83] to titles and description attached to music videos. Before applying doc2vec, we lemmatized each word and removed stop words by using natural language toolkit (NLTK) [84]. In addition, we obtained metadata shown in Table 1 from YouTube API v3[4] in order to build links. Here, tags and keywords where document frequencies were less than five and more than 90 percentiles were removed in order to remove noisy tags and keywords.

For evaluation, we randomly removed a certain percentage of links between nodes of music videos and users as "test links" so that the percentage of the remaining links, called "training links", became {10, 20, 30, 40, 50, 60,

[3] In YouTube-8M, semantic topics called entities are attached to each video for tasks such as classification and clustering.

[4] https://developers.google.com/youtube/v3

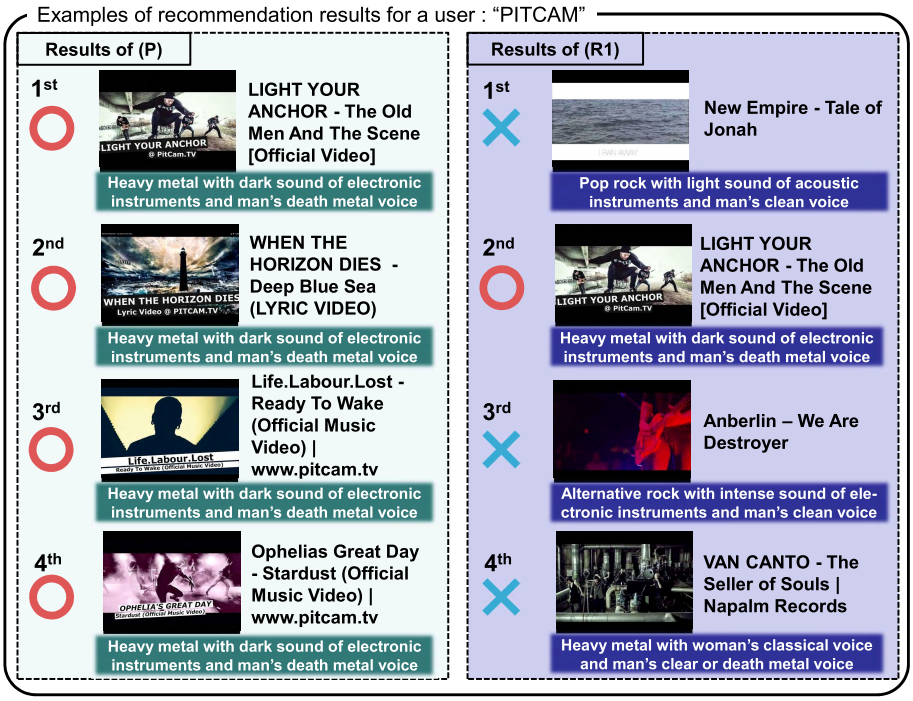


FIGURE 3. Examples of recommendation results for (P) and (R1). Thumbnails of music videos, titles in black letters, and characteristics of music videos in white letters are shown.

70, 80, 90}%. When we calculated user features, we did not utilize features obtained from music videos for which nodes are unlinked to users. We assumed that users prefer their uploaded music videos according to a previous report [85], and we defined their uploaded music videos as the ground truth. Specifically, recommendation results are correct if test links between nodes of music videos and users were predicted. Here, we denote the proposed method by (P). We compared (P) with the following reference methods, (R1), (R2), (R3) and (R4), which are network-based methods that target music videos as in the proposed method.

- **(R1):** This is a method that does not use audio, visual and textual features. This method defines link existence based on multiple social metadata shown in the same manner as (P) and constructs an unweighted network.
- **(R2):** This is a method based on a recently published paper [71]. This method uses audio and textual features and calculates music video features as $[(\hat{\boldsymbol{v}}_a^{(i)})^{\mathrm{T}}(\hat{\boldsymbol{v}}_t^{(i)})^{\mathrm{T}}]^{\mathrm{T}}$. Furthermore, this method performs link prediction based on only node2vec [69]-based similarities. Specifically, this method calculates final link prediction scores by Eq. (5).
- **(R3):** This is a method that performs link prediction using only weighted Jaccard coefficients [68]. Specifically, this method constructs a network in the same manner as (P) and calculates final link prediction scores by Eq. (4).
- **(R4):** This is a method that performs link prediction using node2vec [69]-based similarities. Specifically, this method constructs a network in the same manner as (P) and calculates final link prediction scores by Eq. (5).

B. EVALUATION OF RECOMMENDATION PERFORMANCE

In this section, we evaluate the effectiveness of our method for music video recommendation. First, examples of results for (P) and (R1) where the percentages of test links are 50% are shown in Fig. 3 to qualitatively verify the effectiveness of our method. Note that (R1) does not utilize audio, visual and textual features, but (P) utilizes those features. From the results, we confirmed that (P) highly ranked music videos for which contents, e.g., musical pieces and moving pictures, are similar to each other acoustically and visually. Thus, it can be seen that there are cases where (P) can realize more accurate music video recommendation than (R1) can by using multimodal features.

Next, to quantitatively evaluate each method, we utilize mean average precision (MAP@k) and mean prediction accuracy. MAP@k is the mean of AP@k of all users, and AP@k is defined as follows:

$$\mathrm{AP@}k = \frac{1}{N_k}\sum_{i=1}^{k}\alpha_i\frac{N_i}{i},$$

where N_i is the number of correctly recommended musical pieces within i music videos of the recommendation results, and α_i is 1 if the i-th recommendation result is correct and 0 otherwise. In addition, we define mean prediction accuracy as the mean of prediction accuracy of all users. Specifically,

TABLE 2. MAP@4 of the recommendation results.

	Percentages of training links (%)									
	10	20	30	40	50	60	70	80	90	Mean
(P)	**0.709**	**0.843**	**0.893**	**0.916**	**0.924**	**0.921**	**0.909**	**0.874**	**0.807**	**0.866**
(R1)	0.597	0.721	0.799	0.846	0.870	0.873	0.863	0.824	0.755	0.794
(R2)	0.688	0.804	0.851	0.873	0.887	0.881	0.873	0.846	0.778	0.831
(R3)	0.670	0.813	0.867	0.891	0.894	0.884	0.854	0.809	0.717	0.822
(R4)	**0.709**	0.827	0.870	0.892	0.901	0.895	0.893	0.863	0.803	0.850

TABLE 3. SD of AP@4 of the recommendation results.

	Percentages of training links (%)									
	10	20	30	40	50	60	70	80	90	Mean
(P)	**0.394**	**0.302**	**0.250**	**0.219**	**0.208**	**0.215**	**0.234**	**0.279**	**0.351**	**0.272**
(R1)	0.440	0.395	0.343	0.306	0.280	0.279	0.289	0.324	0.382	0.338
(R2)	0.407	0.336	0.297	0.272	0.251	0.262	0.270	0.305	0.373	0.308
(R3)	0.409	0.326	0.279	0.255	0.255	0.263	0.294	0.336	0.404	0.313
(R4)	0.399	0.318	0.277	0.245	0.234	0.243	0.249	0.291	0.357	0.290

TABLE 4. Mean prediction accuracy of the recommendation results.

	Percentages of training links (%)									
	10	20	30	40	50	60	70	80	90	Mean
(P)	0.556	0.703	**0.769**	**0.799**	**0.801**	**0.776**	**0.715**	**0.586**	**0.396**	**0.678**
(R1)	0.476	0.588	0.661	0.702	0.721	0.703	0.651	0.537	0.362	0.600
(R2)	0.558	0.676	0.729	0.751	0.752	0.725	0.675	0.554	0.371	0.643
(R3)	0.468	0.642	0.718	0.742	0.735	0.701	0.631	0.504	0.330	0.608
(R4)	**0.576**	**0.705**	0.754	0.780	0.778	0.753	0.699	0.572	0.384	0.667

TABLE 5. SD of prediction accuracy of the recommendation results.

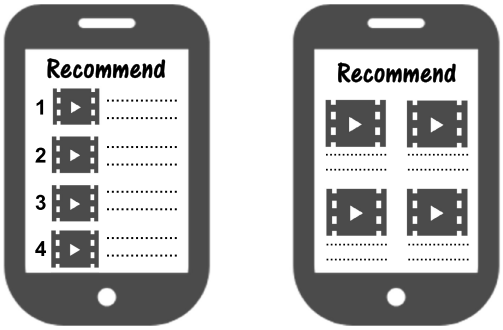
	Percentages of training links (%)									
	10	20	30	40	50	60	70	80	90	Mean
(P)	0.378	**0.335**	**0.302**	**0.283**	**0.277**	**0.283**	**0.289**	**0.282**	0.244	**0.297**
(R1)	0.402	0.386	0.359	0.336	0.321	0.319	0.312	0.296	0.246	0.331
(R2)	0.391	0.353	0.329	0.312	0.302	0.307	0.302	0.289	0.244	0.314
(R3)	**0.358**	0.344	0.323	0.311	0.311	0.313	0.312	0.288	**0.232**	0.310
(R4)	0.388	0.344	0.318	0.296	0.288	0.295	0.293	0.286	0.242	0.306

prediction accuracy is defined as follows:

$$\text{Prediction accuracy} = \frac{N_k}{k}.$$

In this experiment, we set k to 4 in reference to a report that people can only remember about four chunks in short-term memory tasks [86]. In order to evaluate variation of the above evaluation metrics, we also calculated standard deviation (SD) of AP@4 and prediction accuracy. For accurate evaluation, random removal of links and link prediction were repeated 10 times. Then the average of all evaluation metrics for all 10 times was calculated.

Tables 2-5 show the results of music video recommendation. From Table 2, which shows results for MAP@4, we can see that (P) outperforms the other reference methods at all percentages of training links. Here, MAP@4 can consider how many ground truth music videos are ranked in higher ranks. In other words, we can evaluate the effectiveness of adopting the recommendation methods for applications that present recommendation results in a ranking format (see Fig. 4(a)). Also, from Table 4, which shows results of mean prediction accuracy, we can see that (P) outperforms the other reference methods at most percentages. Here, mean prediction accuracy can evaluate the effectiveness of adopting the recommendation methods for applications that present recommendation results in a listing format (see Fig. 4(b)). Therefore, we can confirm that (P) is effective for any application. Furthermore, from Tables 3 and 5, we can see that SD of both metrics in (P) is mostly the smallest of those of all reference methods. Thus, we can confirm that (P) can most accurately recommend music videos for all users without variation.



(a) Ranking format (b) Listing format

FIGURE 4. Examples of applications that present recommendation results.

Next, we evaluated differences between (P) and the other reference methods in detail. When we compared (P) with (R1), the effectiveness of utilization of audio, visual and textual features was verified. From the results obtained by (P) and (R2), we can see that our method outperforms our previous method [71]. Thus, we confirmed the effectiveness of introducing visual features into construction of a network and performing LP-LGSN. When we compared (P) with (R3), we can see that (R3) did not realize more successful link prediction than (P). In real-world video hosting services, there may be a lack of essential social metadata. Thus, there is a case in which nodes have actually similar characteristics on the network but are not common neighbors. In this case, it can be considered that the performance of the link prediction that utilizes only weight Jaccard coefficients can be lowered. Thus, we can confirm that LP-LGSN can estimate relevance between unlinked nodes more accurately than can link prediction of (R3). Moreover, we can see that the performance of LP-LGSN is improved when compared to that of (R4). Note that LP-LGSN extracts pairs of unlinked nodes for which degrees of local and global similarities are both high. It can be considered that LP-LGSN can obtain reliable link prediction scores even if there are unreliable links in the network. Therefore, we can confirm that (P) can realize more accurate link prediction than (R4) can. As a result, we have confirmed the effectiveness of the proposed method.

C. EVALUATION OF THE EFFECTIVENESS OF LP-LGSN

In this section, we evaluate whether LP-LGSN actually contributes to improvement of recommendation performance or not. To verify whether link prediction scores can represent the degrees to which users desire music videos, we calculated Pearson's correlation coefficients between link prediction scores and the ground truth. The evaluation scheme is motivated by the papers [87], [88], in which the discriminant power of features was evaluated by using Pearson's correlation coefficients between the features and labels. Specifically, we calculated the correlation coefficients between link prediction scores of all music videos and labels that indicate 1 if each music video is the ground truth and 0 otherwise.

TABLE 6. Pearson's correlation coefficient of each method.

(P)	(R1)	(R2)	(R3)	(R4)
0.549	0.534	0.110	0.456	0.112

In the same manner as that for the evaluation in Section IV-B, we repeated the link prediction 10 times, and we calculated the average of Pearson's correlation coefficients. Note that the link prediction scheme adopted in each method is (P) and (R1): LP-LGSN, (R2) and (R4): node2vec-based similarities, and (R3): weighted Jaccard coefficients.

Table 6 shows the correlation coefficients of (P) and (R1)-(R4) in which the percentage of training links is 50%. From this table, we can see that the correlation coefficients of (R2) and (R4) are lower than that of (R3). Here, weighted Jaccard coefficients can clarify the difference between recommended music videos and other music videos since weighted Jaccard coefficients are zero if pairs of nodes do not have common neighbors. Thus, link prediction scores of (R3) have high correlation with the labels. However, recommendation performance of (R2) and (R4) is higher than that of (R3) since node2vec-based similarities are accurately calculated in all ranks with consideration of the global structure. Therefore, it can be considered that weighted Jaccard coefficients and node2vec-based similarities can realize accurate recommendation in higher ranks and in all ranks, respectively.

Finally, we can see that the (P) has the highest correlation in all methods. As a result, it can also be considered that this is because LP-LGSN has merits of both weighted Jaccard coefficients and node2vec-based similarities. Above all, we can confirm that link prediction scores obtained by LP-LGSN actually contribute to improvement of recommendation performance.

D. EVALUATION OF PERFORMANCE LIMITATION

To evaluate the effectiveness and limitation of (P) and (R1)-(R4), we calculated precision@k, recall@k and F-measure@k, which are defined as follows.

$$\begin{aligned}\text{Precision@}k &= \frac{N_k}{k},\\ \text{Recall@}k &= \frac{N_k}{N^{(\text{GT})}},\\ \text{F}-\text{measure@}k &= \frac{2\times\text{Precision@}k\times\text{Recall@}k}{\text{Precision@}k+\text{Recall@}k},\end{aligned}$$

where $N^{(\text{GT})}$ is the number of ground truth music videos of each user. In this evaluation, we calculated the averages of these evaluation metrics for all users. We also set k to 4 in the same manner as MAP@k and mean prediction accuracy. In addition, we set k to 1 in order to evaluate the results in the top rank, for which users mostly paid attention to [47], [49], [53]. Tables 7-15 show the results for (P) and (R1)-(R4) when the percentages of training links are 10%-90%, respectively. The results are discussed below.

TABLE 7. Precision, recall and F-measure when the percentage of training links is 10(%).

	Precision@1	Recall@1	F-measure@1	Precision@4	Recall@4	F-measure@4
(P)	0.660	0.044	0.082	0.556	0.148	0.234
(R1)	0.546	0.036	0.068	0.476	0.128	0.201
(R2)	0.641	0.044	0.082	0.558	0.153	0.240
(R3)	0.611	0.042	0.078	0.468	0.128	0.201
(R4)	**0.661**	**0.045**	**0.085**	**0.576**	**0.159**	**0.249**

TABLE 8. Precision, recall and F-measure when the percentage of training links is 20(%).

	Precision@1	Recall@1	F-measure@1	Precision@4	Recall@4	F-measure@4
(P)	**0.808**	**0.061**	**0.113**	0.703	0.211	0.324
(R1)	0.682	0.051	0.095	0.588	0.175	0.270
(R2)	0.769	0.059	0.110	0.676	0.207	0.317
(R3)	0.774	0.059	0.110	0.642	0.196	0.301
(R4)	0.792	**0.061**	**0.113**	**0.705**	**0.217**	**0.332**

TABLE 9. Precision, recall and F-measure when the percentage of training links is 30(%).

	Precision@1	Recall@1	F-measure@1	Precision@4	Recall@4	F-measure@4
(P)	**0.874**	**0.075**	**0.137**	**0.769**	**0.262**	**0.391**
(R1)	0.765	0.065	0.119	0.661	0.223	0.333
(R2)	0.824	0.072	0.132	0.729	0.252	0.375
(R3)	0.840	0.073	0.134	0.718	0.248	0.368
(R4)	0.841	0.073	0.134	0.754	0.261	0.388

TABLE 10. Precision, recall and F-measure when the percentage of training links is 40(%).

	Precision@1	Recall@1	F-measure@1	Precision@4	Recall@4	F-measure@4
(P)	**0.903**	**0.090**	**0.164**	**0.799**	**0.315**	**0.452**
(R1)	0.823	0.081	0.147	0.702	0.274	0.395
(R2)	0.845	0.085	0.155	0.751	0.301	0.430
(R3)	0.870	0.087	0.158	0.742	0.296	0.424
(R4)	0.867	0.087	0.159	0.780	0.313	0.446

TABLE 11. Precision, recall and F-measure when the percentage of training links is 50(%).

	Precision@1	Recall@1	F-measure@1	Precision@4	Recall@4	F-measure@4
(P)	**0.912**	**0.111**	**0.198**	**0.801**	**0.384**	**0.519**
(R1)	0.850	0.102	0.182	0.721	0.342	0.464
(R2)	0.865	0.106	0.188	0.752	0.366	0.493
(R3)	0.880	0.107	0.191	0.735	0.355	0.479
(R4)	0.877	0.107	0.191	0.778	0.379	0.510

TABLE 12. Precision, recall and F-measure when the percentage of training links is 60(%).

	Precision@1	Recall@1	F-measure@1	Precision@4	Recall@4	F-measure@4
(P)	**0.913**	**0.135**	**0.235**	**0.776**	**0.451**	**0.570**
(R1)	0.856	0.126	0.219	0.703	0.404	0.513
(R2)	0.859	0.128	0.223	0.725	0.434	0.543
(R3)	0.867	0.129	0.225	0.701	0.417	0.523
(R4)	0.872	0.131	0.227	0.753	0.450	0.564

Effectiveness

Tables 9-15 show that (P) mostly outperforms all of the reference methods when k is 1. Thus, we can confirm that the proposed method can present the top results most accurately. Also, Tables 9-12 show that precision@4, recall@4 and F-measure@4 of (P) are the highest values of all reference methods when percentages of training links are medium percentages. We can thus confirm that the proposed method is the most effective method for music video recommendation in real-world applications.

Limitation

From Tables 7 and 8, we can see that precision, recall and F-measure of (P) are mostly lower than (R4). This tendency can be also found in Table 4. When percentages of train-

TABLE 13. Precision, recall and F-measure when the percentage of training links is 70(%).

	Precision@1	Recall@1	F-measure@1	Precision@4	Recall@4	F-measure@4
(P)	**0.899**	**0.173**	**0.290**	**0.715**	0.526	0.606
(R1)	0.846	0.161	0.270	0.651	0.476	0.550
(R2)	0.852	0.165	0.276	0.675	0.519	0.587
(R3)	0.837	0.162	0.271	0.631	0.485	0.549
(R4)	0.877	0.170	0.285	0.699	**0.538**	**0.608**

TABLE 14. Precision, recall and F-measure when the percentage of training links is 80(%).

	Precision@1	Recall@1	F-measure@1	Precision@4	Recall@4	F-measure@4
(P)	**0.863**	**0.240**	**0.375**	**0.586**	0.600	0.593
(R1)	0.802	0.221	0.346	0.537	0.550	0.544
(R2)	0.827	0.232	0.362	0.554	0.616	0.583
(R3)	0.784	0.219	0.342	0.504	0.558	0.530
(R4)	0.845	0.237	0.370	0.572	**0.636**	**0.602**

TABLE 15. Precision, recall and F-measure when the percentage of training links is 90(%).

	Precision@1	Recall@1	F-measure@1	Precision@4	Recall@4	F-measure@4
(P)	**0.785**	0.380	**0.512**	**0.396**	0.666	0.497
(R1)	0.723	0.347	0.469	0.362	0.613	0.455
(R2)	0.751	0.370	0.495	0.371	0.717	0.489
(R3)	0.682	0.337	0.451	0.330	0.646	0.437
(R4)	0.778	**0.382**	**0.512**	0.384	**0.742**	**0.506**

ing links are low, weighted Jaccard coefficients cannot be accurately calculated since there are many nodes that do not have common neighbors. Furthermore, recall@4 and F-measure@4 of (P) are lower than those of (R4) as can be seen in Tables 13-15. When percentages of training links are high, weighted Jaccard coefficients cannot be accurately calculated since they are easily affected by noisy links. From the above results, we can consider that weighted Jaccard coefficients are sensitive to the amount and reliability of metadata. In such cases, single utilization of node2vec-based similarities is more effective for performance improvement. On the other hand, utilization of both link prediction scores is effective when the number of links is well-balanced. Therefore, in our future work, we will introduce a framework to automatically control the effects of weighted Jaccard coefficients and node2vec-based similarities into LP-LGSN.

V. CONCLUSION

In this paper, we have presented a novel method based on LP-LGSN for music video recommendation. The contributions of this paper are (i) construction of a network by collaborative use of multi-modal features and social metadata and (ii) derivation of LP-LGSN. From the results of our experiment in which real-world music videos were used, we confirmed the effectiveness of our method. Notably, we have shown the following results. (I) Our method can work well in real-world applications since evaluation metrics that can evaluate the effectiveness in various applications indicate high values in the higher ranks. (II) The effectiveness of fusing multiple link prediction scores via our LP-LGSN can be confirmed by evaluating link prediction scores. As described in Section IV-D, in our future work, we will introduce a framework to fuse link prediction scores that can automatically control the effects of local and global structure-based similarities into LP-LGSN.

REFERENCES

[1] IFPI, (2018). *Music Consumer Insight Report*. [Online]. Available: http://www.ifpi.org/downloads/Music-Consumer-Insight-Report-2018.pdf

[2] IFPI, (2018). *Global Music Report: State of the Industry*. [Online]. Available: http://www.ifpi.org/downloads/GMR2018.pdf

[3] X. Cheng, C. Dale, and J. Liu, "Understanding the characteristics of Internet short video sharing: YouTube as a case study," Jul. 2007, *arXiv:0707.3670*. [Online]. Available: https://arxiv.org/abs/0707.3670

[4] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: A view from the Edge," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas.*, Oct. 2007, pp. 15–28.

[5] L. A. Liikkanen and A. Salovaara, "Music on YouTube: User engagement with traditional, user-appropriated and derivative videos," *Comput. Hum. Behav.*, vol. 50, pp. 108–124, Sep. 2015.

[6] Y. Song, S. Dixon, and M. Pearce, "A survey of music recommendation systems and future perspectives," in *Proc. 9th Int. Symp. Comput. Modeling Retr.*, Jun. 2012, pp. 395–410.

[7] O. Celma, *Music Recommendation Discovery*. New York, NY, USA: Springer, 2010.

[8] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proc. IEEE*, vol. 96, no. 4, pp. 668–696, Apr. 2008.

[9] E. Zheng, G. Y. Kondo, S. Zilora, and Q. Yu, "Tag-aware dynamic music recommendation," *Expert Syst. Appl.*, vol. 106, pp. 244–251, Sep. 2018.

[10] J.-H. Su, W.-Y. Chang, and V. S. Tseng, "Integrated mining of social and collaborative information for music recommendation," *Data Sci. Pattern Recognit.*, vol. 1, no. 1, pp. 13–30, 2017.

[11] M. S. Reddy and T. Adilakshmi, "Music recommendation system based on matrix factorization technique-SVD," in *Proc. Int. Conf. Comput. Commun. Inform.*, Jan. 2014, pp. 1–6.

[12] D. Wang, S. Deng, S. Liu, and G. Xu, "Improving music recommendation using distributed representation," in *Proc. 25th Int. Conf. Companion World Wide Web*, Apr. 2016, pp. 125–126.

[13] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences," in *Proc. ISMIR*, Oct. 2006, pp. 1–6.

[14] D. Sánchez-Moreno, A. B. G. González, M. D. M. Vicente, V. F. L. Batista, and M. N. M. García, "A collaborative filtering method for music recommendation using playing coefficients for artists and users," *Expert Syst. Appl.*, vol. 66, pp. 234–244, Dec. 2016.

[15] N. Hariri, B. Mobasher, and R. Burke, "Context-aware music recommendation based on latenttopic sequential patterns," in *Proc. 6th ACM Conf. Recommender Syst.*, Sep. 2012, pp. 131–138.

[16] L. Jin, D. Yuan, and H. Zhang, "Music recommendation based on embedding model with user preference and context," in *Proc. IEEE 2nd Int. Conf. Big Data Anal.*, Mar. 2017, pp. 688–692.
[17] Z. Cheng and J. Shen, "Just-for-me: An adaptive personalization system for location-aware social music recommendation," in *Proc. Int. Conf. Multimedia Retr.*, Apr. 2014, p. 185.
[18] Z. Cheng and J. Shen, "On effective location-aware music recommendation," *J. ACM Trans. Inf. Syst.*, vol. 34, no. 2, p. 13, Apr. 2016.
[19] S. Oramas, O. Nieto, M. Sordo, and X. Serra, "A deep multimodal approach for cold-start music recommendation," Jun. 2017, *arXiv:1706.09739*. [Online]. Available: https://arxiv.org/abs/1706.09739
[20] S.-Y. Chou, L.-C. Yang, Y.-H. Yang, and J.-S. R. Jang, "Conditional preference nets for user and item cold start problems in music recommendation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1147–1152.
[21] X. Wang and Y. Wang, "Improving content-based and hybrid music recommendation using deep learning," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 627–636.
[22] A. van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2643–2651.
[23] D. Liang, M. Zhan, and D. P. Ellis, "Content-aware collaborative music recommendation using pre-trained neural networks," in *Proc. ISMIR*, Oct. 2015, pp. 295–301.
[24] M. Li, W. Jiang, and K. Li, "When and what music will you listen to? Fine-grained time-aware music recommendation," in *Proc. IEEE Int. Symp. Parallel Distrib. Process. Appl. IEEE Int. Conf. Ubiquitous Comput. Commun.*, Dec. 2017, pp. 1091–1098.
[25] A. Flexer and J. Stevens, "Mutual proximity graphs for improved reachability in music recommendation," *J. Music Res.*, vol. 47, no. 1, pp. 17–28, Jan. 2018.
[26] S. Oramas, V. C. Ostuni, T. D. Noia, X. Serra, and E. D. Sciascio, "Sound and music recommendation with knowledge graphs," *ACM Trans. Intell. Syst. Technol.*, vol. 8, p. 21, Oct. 2016.
[27] S. Tan, J. Bu, C. Chen, B. Xu, C. Wang, and X. He, "Using rich social media information for music recommendation via hypergraph model," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 7, no. 1, p. 22, Oct. 2011.
[28] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, L. Zhang, and X. He, "Music recommendation by unified hypergraph: Combining social media information and music content," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 391–400.
[29] C. Guo and X. Liu, "Automatic feature generation on heterogeneous graph for music recommendation," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 807–810.
[30] C. Guo and X. Liu, "Dynamic feature generation and selection on heterogeneous graph for music recommendation," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 656–665.
[31] K. Benzi, V. Kalofolias, X. Bresson, and P. Vandergheynst, "Song recommendation with non-negative matrix factorization and graph total variation," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2439–2443.
[32] K. Mao, G. Chen, Y. Hu, and L. Zhang, "Music recommendation using graph based quality model," *Signal Process.*, vol. 120, pp. 806–813, Mar. 2016.
[33] D. Bugaychenko and A. Dzuba, "Musical recommendations and personalization in a social network," in *Proc. 7th ACM Conf. Recommender Syst.*, Oct. 2013, pp. 367–370.
[34] D. Wang, G. Xu, and S. Deng, "Music recommendation via heterogeneous information graph embedding," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 596–603.
[35] Y. Li, H. Wang, H. Liu, and B. Chen, "A study on content-based video recommendation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4581–4585.
[36] S. Roy and S. C. Guntuku, "Latent factor representations for cold-start video recommendation," in *Proc. 10th ACM Conf. Recommender Syst.*, Sep. 2016, pp. 99–106.
[37] M. Yan, J. Sang, C. Xu, and M. S. Hossain, "A unified video recommendation by cross-network user modeling," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 12, no. 4, p. 53, Aug. 2016.
[38] J. Niu, S. Wang, Y. Su, and S. Guo, "Temporal factor-aware video affective analysis and recommendation for cyber-based social media," *IEEE Trans. Emerg. Topics Comput.*, vol. 5, no. 3, pp. 412–424, Jul. 2017.
[39] M. Yan, W. Shang, and Z. Li, "Application of SVD technology in video recommendation system," in *Proc. IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci.*, Jun. 2016, pp. 1–5.
[40] Y. Liu, G. Zhang, X. Jin, and Y. Jia, "Rating matrix pre-padding for video recommendation," in *Proc. Conf. Technol. Appl. Artif. Intell. (TAAI)*, Dec. 2018, pp. 164–167.
[41] X. Du, H. Yin, L. Chen, Y. Wang, Y. Yang, and X. Zhou, "Personalized video recommendation using rich contents from videos," *IEEE Trans. Knowl. Data Eng.*, to be published.
[42] J. Zhang, Y. Zhou, D. Wu, and C. Yang, "Context-aware video recommendation by mining users' View preferences based on access points," in *Proc. 27th Workshop Netw. Operating Syst. Support Digit. Audio Video*, Jun. 2017, pp. 37–42.
[43] G. Wu, V. Swaminathan, S. Mitra, and R. Kumar, "Context-aware video recommendation based on session progress prediction," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1428–1433.
[44] Y. Zhou, J. Wu, T. H. Chan, S. Ho, D.-M. Chiu, and D. Wu, "Interpreting video recommendation mechanisms by mining view count traces," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2153–2165, Aug. 2018.
[45] X. Jia, A. Wang, X. Li, G. Xun, W. Xu, and A. Zhang, "Multi-modal learning for video recommendation based on mobile application usage," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Nov. 2015, pp. 837–842.
[46] Q. Zhu, M.-L. Shyu, and H. Wang, "VideoTopic: Content-based video recommendation using a topic model," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2013, pp. 219–222.
[47] S. Basu, Y. Yu, V. K. Singh, and R. Zimmermann, "Videopedia: Lecture video recommendation for educational Blogs using topic modeling," in *Proc. Int. Conf. Multimedia Modeling*, Jan. 2016, pp. 238–250.
[48] C. Yang, H. Yan, D. Yu, Y. Li, and D. M. Chiu, "Multi-site user behavior modeling and its application in video recommendation," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 175–184.
[49] L. Sun, X. Wang, Z. Wang, H. Zhao, and W. Zhu, "Social-aware video recommendation for online social groups," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 609–618, Mar. 2017.
[50] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youTube recommendations," in *Proc. 10th ACM Conf. Recommender Syst.*, Sep. 2016, pp. 191–198.
[51] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, and D. Sampath, "The YouTube video recommendation system," in *Proc. 4th ACM Conf. Recommender Syst.*, Sep. 2010, pp. 293–296.
[52] J. Lee and S. Abu-El-Haija, "Large-scale content-only video recommendation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 987–995.
[53] J. Ma, G. Li, M. Zhong, X. Zhao, L. Zhu, and X. Li, "LGA: Latent genre aware micro-video recommendation on social media," *Multimedia Tools Appl.*, vol. 77, no. 3, pp. 2991–3008, Feb. 2018.
[54] J. Dong, X. Li, C. Xu, G. Yang, and X. Wang, "Feature re-learning with data augmentation for content-based video recommendation," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2018, pp. 2058–2062.
[55] C. Hongliang and Q. Xiaona, "The video recommendation system based on DBN," in *Proc. IEEE Int. Conf. Comput. Inf. Technol., Ubiquitous Comput. Commun., Dependable, Autonomic Secure Comput., Pervasive Intell. Comput.*, Oct. 2015, pp. 1016–1021.
[56] L. Cui, L. Dong, X. Fu, Z. Wen, N. Lu, and G. Zhang, "A video recommendation algorithm based on the combination of video content and social network," *Concurrency Comput., Pract. Exper.*, vol. 29, no. 14, 2017, Art. no. e3900.
[57] W. Shang, S. Shang, S. Feng, and M. Shi, "An improved video recommendations based on the hyperlink-graph model," in *Proc. 4th Intl Conf. Appl. Comput. Inf. Technol./3rd Intl Conf. Comput. Sci./Intell. Appl. Inform./1st Intl Conf. Big Data, Cloud Comput., Data Sci. Eng.*, Dec. 2016, pp. 379–383.
[58] S. Pang, W. Wang, and H. Zhu, "A personalized video recommendation algorithm based on complex network," in *Proc. 2nd Int. Conf. Artif. Intell., Technol. Appl.*, Mar. 2018, pp. 66–69.
[59] Z. Zhao, Q. Yang, H. Lu, T. Weninger, D. Cai, X. He, and Y. Zhuang, "Social-aware movie recommendation via multimodal network learning," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 430–440, Feb. 2018.
[60] X. Zhou, L. Chen, Y. Zhang, L. Cao, G. Huang, and C. Wang, "Online video recommendation in sharing community," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2015, pp. 1645–1656.
[61] X. Zhou, L. Chen, Y. Zhang, D. Qin, L. Cao, G. Huang, and C. Wang, "Enhancing online video recommendation using social user interactions," *Int. J. Very Large Data Bases*, vol. 26, no. 5, pp. 637–656, 2017.

[62] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Comput.*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[63] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[64] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," Jul. 2017, *arXiv:1707.07435*. [Online]. Available: https://arxiv.org/abs/1707.07435

[65] I. Pitas, *Graph-Based Social Media Analysis*. Boca Raton, FL, USA: CRC Press, 2016.

[66] J. Cao, Y. Zhang, R. Ji, F. Xie, and Y. Su, "Web video topics discovery and structuralization with social network," *Neurocomputing*, vol. 172, pp. 53–63, Jan. 2015.

[67] R. Harakawa, T. Ogawa, and M. Haseyama, "Accurate and efficient extraction of hierarchical structure of Web communities for Web video retrieval," *ITE Trans. Media Technol. Appl.*, vol. 4, no. 1, pp. 49–59, 2016.

[68] F. Mitzlaff and G. Stumme, "Ranking given names: Algorithms and evaluation paradigms," in *Proc. Int. Conf. Social Inform.*, Dec. 2012, pp. 185–191.

[69] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 855–864.

[70] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge, U.K.: Univ. Press, 1998.

[71] Y. Matsumoto, R. Harakawa, T. Ogawa, and M. Haseyama, "Construction of network using heterogeneous social metadata for music video recommendation," in *Proc. IEEE 6th Global Conf. Consum. Electron. (GCCE)*, Oct. 2017, pp. 1–2.

[72] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8m: A large-scale video classification benchmark," Sep. 2016, *arXiv:1609.08675*. [Online]. Available: https://arxiv.org/abs/1609.08675

[73] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[74] R. Harakawa, T. Ogawa, and M. Haseyama, "Extracting hierarchical structure of Web video groups based on sentiment-aware signed network analysis," *IEEE Access*, vol. 16, pp. 16963–16973, 2017.

[75] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of youTube video," in *Proc. 16th Interntional Workshop Qual. Service*, Jun. 2008, pp. 229–238.

[76] R. Hong, J. Tang, H.-K. Tan, C.-W. Ngo, S. Yan, and T.-S. Chua, "Beyond search: Event-driven summarization for Web videos," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 7, no. 4, p. 35, Nov. 2011.

[77] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, Jun. 1967, pp. 281–297.

[78] D. Takehara, R. Harakawa, T. Ogawa, and M. Haseyama, "Extracting hierarchical structure of content groups from different social media platforms using multiple social metadata," *Multimedia Tools Appl.*, vol. 76, no. 19, pp. 20249–20272, Oct. 2017.

[79] A. Narayanan, E. Shi, and B. I. P. Rubinstein, "Link prediction by de-anonymization: How we won the Kaggle social network challenge," in *Proc. Int. Joint Conf. Neural Netw.*, Aug. 2011, pp. 1825–1834.

[80] L. Backstrom and J. Leskovec, "Supervised random walks: Predicting and recommending links in social networks," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, Feb. 2011, pp. 635–644.

[81] R. Harakawa, T. Ogawa, and M. Haseyama, "Extraction of hierarchical structure of Web communities including salient keyword estimation for Web video retrieval," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 1021–1025.

[82] A. Jansen, J. F. Gemmeke, D. P. W. Ellis, X. Liu, W. Lawrence, and D. Freedman, "Large-scale audio event discovery in one million YouTube videos," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 786–790.

[83] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2014, pp. 1188–1196.

[84] E. B. L. Steven and E. Klein, *Natural Language Processing With Python*. Sebastopol, CA, USA: O'Reilly Media Inc., 2009

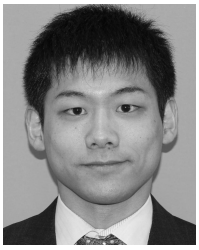
[85] Z. Deng, M. Yan, J. Sang, and C. Xu, "Twitter is faster: Personalized time-aware video recommendation from Twitter to youTube," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 2, p. 31, Dec. 2014.

[86] N. Cowan, "The magical number 4 in short-term memory: A reconsideration of mental storage capacity," *Behav. Brain Sci.*, vol. 24, no. 1, pp. 87–114, 2001.

[87] K. Maeda, S. Takahashi, T. Ogawa, and M. Haseyama, "Estimation of deterioration levels of transmission towers via deep learning maximizing canonical correlation between heterogeneous features," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 4, pp. 633–644, Aug. 2018.

[88] S.-B. Chen, Y. Zhang, C. H. Q. Ding, Z.-L. Zhou, and B. Luo, "A discriminative multi-class feature selection method via weighted $l_{2,1}$-norm and Extended Elastic Net," *Neurocomputing*, vol. 275, pp. 1140–1149, Jan. 2018.

YUI MATSUMOTO (S'17) received the B.S. degree in electronics and information engineering from Hokkaido University, Japan, in 2018, where she is currently pursuing the M.S. degree with the Graduate School of Information Science and Technology. Her research interests include music information retrieval and web mining. She is a Student Member of the IEICE.



RYOSUKE HARAKAWA (S'13–M'16) received the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan, in 2013, 2015, and 2016, respectively, all in electronics and information engineering. He is currently an Assistant Professor with the Department of Electrical, Electronics and Information Engineering, Nagaoka University of Technology. His research interests include multimedia information retrieval and web mining. He is a member of the IEICE and the Institute of Image Information and Television Engineers (ITE).

TAKAHIRO OGAWA (S'03–M'08–SM'18) 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 ACM, EURASIP, IEICE, and the Institute of Image Information and Television Engineers (ITE).

MIKI HASEYAMA (S'88–M'91–SM'06) 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, St. Louis, MO, 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 IEICE and the Information Processing Society of Japan (IPSJ), and a Fellow of the Institute of Image Information and Television Engineers (ITE). She has been the Vice-President of the ITE, the Editor-in-Chief of the *ITE Transactions on Media Technology and Applications*, and the Director of the International Coordination and Publicity of the IEICE.