IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 11, 2018, accepted October 2, 2018, date of publication October 18, 2018, date of current version November 9, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2876710

Favorite Video Classification Based on Multimodal Bidirectional LSTM

TAKAHIRO OGAWA, (Senior Member, IEEE), YUMA SASAKA, (Student Member, IEEE), KEISUKE MAEDA, (Student Member, IEEE), AND MIKI HASEYAMA, (Senior Member, IEEE)

Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan

Corresponding author: Takahiro Ogawa (ogawa@lmd.ist.hokudai.ac.jp)

This work was partly supported by the MIC/SCOPE #181601001 and JSPS KAKENHI Grant Number JP17H01744.

ABSTRACT Video classification based on the user's preference (information of what a user likes: WUL) is important for realizing human-centered video retrieval. A better understanding of the rationale of WUL would greatly contribute to the support for successful video retrieval. However, a few studies have shown the relationship between information of what a user watches and WUL. A new method that classifies videos on the basis of WUL using video features and electroencephalogram (EEG) signals collaboratively with a multimodal bidirectional Long Short-Term Memory (Bi-LSTM) network is presented in this paper. To the best of our knowledge, there has been no study on WUL-based video classification using video features and EEG signals collaboratively with LSTM. First, we newly apply transfer learning to the WUL-based video classification since the number of labels (liked or not liked) attached to videos by users is small, and it is difficult to classify videos based on WUL. Furthermore, we conduct a user study for showing that the representation of psychophysiological signals calculated from Bi-LSTM is effective for the WUL-based video classification. Experimental results showed that our deep neural network feature representations can distinguish WUL for each subject.

INDEX TERMS Multimodal fusion, video classification, LSTM, EEG.

I. INTRODUCTION

In recent years, studies on multimedia content analysis and retrieval have attracted much attention. Due to the development of many techniques for analyzing multimedia contents, automatic recognition of video contents has been successfully realized [1], [2].

However, many studies have focused almost exclusively on the information of what a user watches (WUW) and there have been few studies on estimation of individual video preference (information of what a user likes: WUL). The relationship between WUW and WUL is shown in Fig. 1. WUL is also key information for realizing various applications. Specifically, WUL enables users to find videos and notice a new pattern of videos they actually like. Furthermore, WUL enables newly registered users to better understand the video content since they are generally treated as the average users with recommendation of the most liked content [3]. Thus, both WUW and WUL should be exploited in a variety of practical applications, but few studies have focused on WUL.

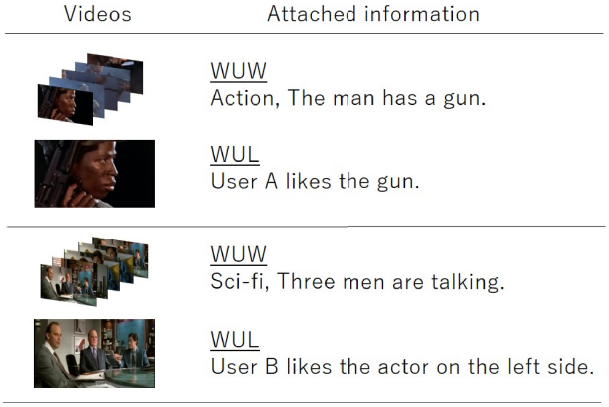


FIGURE 1. Relationship between WUW and WUL.

A. THE POTENTIAL OF WUL

Several video classification methods based on WUL have recently been proposed [4]–[7]. They model WUL by extracting WUW based on audio-visual features to achieve video classification. However, the audio-visual features decrease the performance of WUL-based classification of videos when WUL differs from person to person. Therefore, by using only these features, WUL-based modeling is difficult. Moreover, creating a personalized model is also difficult by using only these video features due to the limitation of

training data caused by the sparsity of user-item matrices on the Web [3]. Regarding these problems, we argue that using only WUW is not sufficient for WUL-based video classification.

Since various sensing techniques have been developed in recent years [8]–[12], several studies have focused only on psychophysiological data, such as electroencephalogram (EEG) data, in the context of multimedia content analysis. EEG signals are major cues in WUL estimation while users are watching videos [13]–[16]. Furthermore, since it has become easier to observe biological signals due to the recent developments of biological sensors, the quality of these signals has also become better [10]–[12]. Specifically, EEG signals are effective for estimating WUL, and they are widely used.

We make the assumption that WUL has visual patterns and has effects on psychophysiological signals such as EEG signals. Past studies have shown that the former can be estimated on the basis of classifiers using visual features and that the latter can be captured by automatic classifiers of sensing data. The reason for our assumption of a connection between the psychophysiological signals and video contents is on the basis of ''Implicit Human-Centered Tagging'' (IHCT). IHCT was proposed in [17] and explained in [18] as a method based on nonverbal behaviors. When interacting with multimedia content, the nonverbal behaviors provide effective information for improving the quality of tags associated with the multimedia content. This definition of the concept implies that WUL is observable in psychophysiological signals while users are watching videos and also that WUL cannot be estimated from only visible information within video contents.

It should be noted that WUL is a different concept from popularity. A popular video generally refers to a video that attracts millions of views and follows a long tail distribution [19], [20]. Therefore, popular videos might be liked by some users but not by other users.

B. CHALLENGES AND CONTRIBUTIONS

In this paper, a novel method for WUL-based video classification is presented. The method is realized by using both WUW and WUL collaboratively with a focus on video classification algorithms based on multimodal bidirectional long short-term memory (Bi-LSTM) [21], [22]. Bi-LSTM can be regarded as an extended version of Recurrent Neural Networks (RNNs), and it has been reported that Bi-LSTM can successfully perform multimedia recognition tasks [23]. For realizing WUL-based video classification, we address the following three important points.

The first point is realization of WUL-based video classification with the collaborative use of WUW and WUL [2]. Past studies have focused on either WUW or WUL. However, as mentioned above, both WUW and WUL should be exploited in a variety of practical applications. We therefore assume that WUL has visual patterns and has effects on psychophysiological signals and that collaborative use of both WUW and WUL enables realization of WUL-based video classification.

As the second point, we try to solve the difficulty of learning good video representation for WUL-based video classification due to the limitation of training data caused by the sparsity of user item matrices on the Web. While the number of videos on the Web is increasing, the number of label videos, i.e., labeled as liked or not liked, for classifying videos based on WUL is small [3], [18]. A large number of training videos has been required for WUL-based video classification in related studies. We therefore newly apply transfer learning to WUL-based video classification. Specifically, we use a framework of inductive transfer learning [24] that realizes accurate classification based on a transfer learning approach. In this framework, only a small amount of training data is necessary for constructing the prediction function. The framework is suitable for WUL-based video classification mentioned above.

The third point is the use of psychophysiological signals for our multimedia application, and this is a new approach in our target research field. Our method introduces feature vectors from EEG signals with Bi-LSTM, which reads all of the signals and produces the representation. There have been few studies using EEG representation with LSTM [25] for WUL-based video classification except for [26]. However, merely utilizing the vanilla LSTM is not sufficient to model WUL since the behavior of EEG signals representing personal liking can also be observed in the reverse order. The representation of EEG signals calculated from Bi-LSTM is effective for realizing feasible WUL-based video classification.

This paper is organized as follows. In section **II**, some reviews of related studies are shown. In section **III**, we explain our favorite video classification method. The proposed method extracts two kinds of features, video features and EEG-based features, and realizes classification based on multimodal Bi-LSTM. In section **IV**, results of experiments are shown to verify the effectiveness of our favorite video classification method. Finally, we conclude our paper and show our possible future work in section **V**.

II. RELATED WORKS

Various methods for affective video classification including WUL-based video classification have been proposed. In this section, we first show some models of affective states in related works. Next, we show brief reviews of features used for realizing the above tasks such as multimedia features and biological features. Furthermore, we explain the effectiveness of the use of EEG features in our method.

A. MODELS OF AFFECTIVE STATES

There are several conceptual models of affective states. Emotion is a psychophysiological process that is evoked by target objects or situations. Although it is easy to assign labels such as joy or fear for representing emotions, this approach has some drawbacks. Specifically, matching between

different languages representing emotions is quite difficult. As an illustration, a word representing ''disgust'' does not exist in Polish [27]. Thus, various studies tend to characterize emotions only in 2- or 3-dimensional space such as valence-arousal-dominance [26], [28], [29].

Next, we focus on interest, which is a psychological element. Berlyne explained interest as an emotional state in terms of curiosity evoked by target objects and situations [30]. Silva [31], [32] focused on a cognitive perspective and concerned relationship between ''interest'' and stimulus complexity. Fairclough *et al.* [14] divided interest into three types of process, arousal, valence and cognition, in their work.

Personal liking (*How much do you like the video?*) is one of the key measures as affective states and was used in [26], [28]. Note that the measure is different from valence scale since users may prefer videos even though the users feel sad or angry.

As mentioned above, different models of affective states were used in the existing work. In this paper, we define WUL as information of personal liking and use the model for WUL-based classification.

B. WUW-BASED AFFECTIVE VIDEO CLASSIFICATION

Affective video classification is an important cue for realizing effective retrieval and recommendation of Web contents. There have been several efforts to classify videos based on users' affective states (e.g., interested/not interested, like/dislike) utilizing only video features [4]–[7]. Grabner *et al.* studied a prediction method for selecting parts that interest many viewers from target videos [4]. In [5] and [6], modeling of users' preference for videos was conducted on the basis of low-level visual features in the videos. Such visual features obtained from target videos have been used in a number of studies. On the other hand, other kinds of features such as textual and audio features have also been introduced [5]–[7]. Deep learning frameworks were used to classify videos based on users' affective states in some studies [33], [34]. For example, a multimodal deep Boltzmann machine (MMDBM) was used for learning a joint density model targeting visual, auditory, and textual features to realize emotion tagging [33]. Instead of using MMDBM, Gan *et al.* [34] proposed a multimodal deep regression Bayesian network (MMDRBN) for constructing higher-level joint representation of visual and auditory features to realize emotion tagging. Although they reported advantages of MMDRBN compared with several multimodal methods, MMDRBN needs a large number of training samples with emotion tags for achieving high performance and avoiding overfitting. That trait is not suitable for the setting of WUL-based video classification since the numbers of both liked and not liked videos tagged by each user tend to be small.

We argue that using only WUW is not sufficient for WUL-based video classification. Therefore, we need to solve the problems by classifying videos based on WUL.

C. PSYCHOPHYSIOLOGICAL SIGNAL-BASED AFFECTIVE VIDEO CLASSIFICATION

Various kinds of sensing devices have been developed, and many studies using features obtained from such devices have been carried out to classify videos based on users' affective states while they are watching videos [13]–[16], [35]–[39].

For analyzing users' interest, i.e., users' attention, most studies have focused on the relationship between visual stimuli and users' attention [40]. Some studies have focused on gaze information obtained while users look at images and watch videos [35]–[37]. Unfortunately, the information used in the above studies is not always matched to WUL. For example, even though users watch videos to find what happens in the videos, this situation is not the same as users liking the videos.

On the other hand, ''interest'' was regarded as an affective state in several studies [13]–[16]. EEG signals have therefore been used to estimate users' interest in some studies. For example, in [15], both users' EEG signals and gaze information were recorded and used for evaluating videos automatically. Furthermore, in [14], EEG signals were recorded while users watched videos and the EEG signals were classified into two classes (high/low interest).

Videos were classified on the basis of WUL in [26], [28], [41], and [42]. EEG data were also used in those studies, and the effectiveness of using EEG for WUL-based video classification was shown. However, to the best of our knowledge, there has been little work using EEG representation with LSTM [25] for WUL-based video classification except for [26].

As shown in the above studies, EEG signals can be used as important modalities to classify videos based on WUL. However, related studies using EEG have focused only on WUL. Both WUW and WUL should be exploited in a variety of practical applications.

III. WUL-BASED CLASSIFICATION BASED ON MULTIMODAL BI-LSTM

A WUL-based video classification method using both WUW and WUL is presented in this section. First, we explain the representation of two kinds of signals, videos and EEG signals. We then describe in detail favorite video classification with multimodal Bi-LSTM. Figure 2 shows an overview of our favorite video classification method based on multimodal bi-directional LSTM.

A. REPRESENTATION OF VIDEOS

Many video analysis methods represent target videos by extracting features from a single frame or multiple consecutive frames and integrate the features over the frames. To recognize actions and events in videos, approaches based on deep Convolutional Neural Networks (CNNs) [43], [44] and RNNs [25], [45] have achieved excellent results. The availability of datasets such as Sports-1M [44] and ActivityNet [46] has encouraged research on video

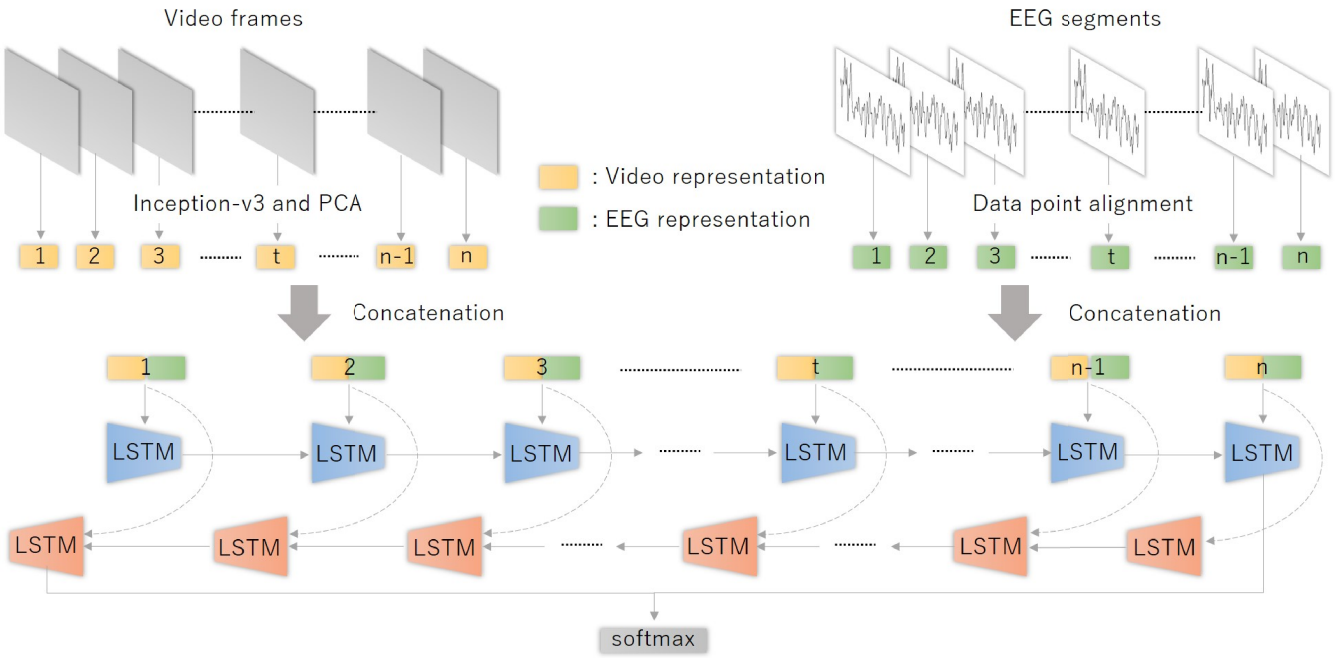


FIGURE 2. Overview of the proposed favorite video classification based on multimodal bi-directional LSTM.

classification of sports and human activities. However, due to the limitation of publicly available datasets, approaches for video analysis have been restricted to small-scale data, while large-scale video understanding remains an underaddressed problem [47]. Google released a multi-modal YouTube-8M dataset [48] that contains about 8M videos and 4716 unique tags to overcome this problem. The dataset is widely used in video understanding challenges and competitions. We therefore used the same feature extraction architecture as that in YouTube-8M to obtain video representation in the proposed method.

In YouTube-8M, visual features are pre-calculated at every second for each video. These visual features consist of ReLU activations of the last fully-connected layer from Inception-v3[1] trained on ImageNet [49]. Then principal component analysis (PCA) and whitening are used for reducing their dimension to 1024. Given a video V consisting of frames $F = (f_1, f_2, \cdots, f_t, \cdots, f_n; n$ being the number of frames), we extract a video representation $\boldsymbol{X}^v = (\boldsymbol{x}_1^v, \boldsymbol{x}_2^v, \cdots, \boldsymbol{x}_t^v, \cdots, \boldsymbol{x}_n^v)$ for each frame f_t with the trained network and the calculated PCA model. Thus, we can calculate $\boldsymbol{x}_t^v \in \mathbb{R}^{1024}$ from each frame of a video V.

B. REPRESENTATION OF EEG SIGNALS

We extract neurophysiological features of biological signals from EEG signals on the basis of results of prior studies in which EEG signals were used in some multimedia applications [14]–[16], [26], [28], [50], [51]. However, it should be noted that there has been little work using EEG representation with LSTM for WUL-based video classification except for [26]. Similar to [26], we used raw EEG signals as input to Bi-LSTM.

In the proposed method, EEG signals are recorded through alphatec IV-s, which provides EEG signals from FP1 with a sampling rate of 1024 Hz. The EEG signals obtained from FP1 have relationships with affective states such as liked/not liked [26], [28]. We segment target EEG signals into several parts at a fixed interval using a Hamming window. We extract an EEG representation $\boldsymbol{X}^e = (\boldsymbol{x}_1^e, \boldsymbol{x}_2^e, \cdots, \boldsymbol{x}_t^e, \cdots, \boldsymbol{x}_n^e)$ for each frame f_t, where the dimension of $\boldsymbol{x}_t^e$ is 1024, which corresponds to the number of data points.

C. FAVORITE VIDEO CLASSIFICATION

For each video frame f_t, we have feature sequences $\boldsymbol{X}^v$ and $\boldsymbol{X}^e$. In our method, it is important to realize effective multimodal fusion of different kinds of features. Early fusion, which concatenates different feature vectors, and late fusion, which performs pooling of multiple outputs, are well known as conventional fusion methods. To show the effectiveness of using both $\boldsymbol{X}^v$ and $\boldsymbol{X}^e$ collaboratively for WUL-based classification, we utilize the simplest way of concatenating feature vectors. Therefore, we denote $\boldsymbol{X}^{fw} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_n)$, where $\boldsymbol{x}_t \in \mathbb{R}^{2048}$ is calculated by concatenating $\boldsymbol{x}_t^v$ and $\boldsymbol{x}_t^e$. Then we calculate the sequence-to-one fixed length vector mapping to a forward LSTM model,

$$y^n = LSTM(\boldsymbol{X}^{fw}). \tag{1}$$

In the above equation, n represents the length of the target feature sequence, which is input to the LSTM model.

[1] https://www.tensorflow.org/tutorials/image_recognition

Furthermore, y^n is the final prediction value in the prediction sequence.

LSTM is an attractive sequential model for which the basic ideas are based on RNN. Since this model is equipped with input gates, forget gates and output gates in the memory block, it can adapt new data more rapidly than an RNN can. Therefore, we adopt LSTM for realizing sequence-to-one fixed length vector mapping. The classification proceeds on the basis of the last hidden vector. Note that in the LSTM model, the hidden vector $\boldsymbol{h}_t$ in time step t is modeled as follows:

$$\boldsymbol{h}_t = LSTM(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t). \tag{2}$$

The LSTM function is computed as follows:

$$LSTM(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t) = \boldsymbol{o}_t tanh(\boldsymbol{c}_t), \tag{3}$$

where

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o), \tag{4}$$
$$\boldsymbol{c}_t = \boldsymbol{f}_t\boldsymbol{c}_{t-1} + \boldsymbol{i}_t tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c), \tag{5}$$
$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f), \tag{6}$$
$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{b}_i). \tag{7}$$

Note that $\sigma()$ is a function that outputs element-wise sigmoid values. Furthermore, $\boldsymbol{i}_t, \boldsymbol{f}_t$ and $\boldsymbol{o}_t$ are the input gate, the forget gate, and the output gate, respectively. Then $\boldsymbol{c}_t$ is a cell activation vector obtained for the t-th input vector. The matrix $\boldsymbol{W}_{kk}$ ($\boldsymbol{k} \in \{x, h, i, f, o, c\}$) and the vector $\boldsymbol{b}_k$ represent the weight and bias, respectively.

Given feature sequences $\boldsymbol{X}^v$ and $\boldsymbol{X}^e$, it makes intuitive sense that the behavior of EEG signals representing personal liking can also be observed in the reverse order. For example, some people like a panda that is shown either after or before a horse in a video. Therefore, we build not only a forward LSTM (FW-LSTM) but also a backward LSTM (BW-LSTM).

$$y^n = LSTM(\boldsymbol{X}^{bw}), \tag{8}$$

where $\boldsymbol{X}^{bw} = (\boldsymbol{x}_n, \boldsymbol{x}_{n-1}, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_1)$. We then calculate the hidden vector $\boldsymbol{h}$ by concatenating $\boldsymbol{h}_n^{fw}$ and $\boldsymbol{h}_1^{bw}$. The classification of a video is based on the hidden vector $\boldsymbol{h}$ as follows:

$$y^n = softmax(\boldsymbol{W}_h\boldsymbol{h} + \boldsymbol{b}_h), \tag{9}$$

where $\boldsymbol{W}_h$ and $\boldsymbol{b}_h$ are the weight matrix and bias vector, respectively.

IV. EXPERIMENTS

Experimental results for verifying the effectiveness of the proposed method are shown in this section. We used a two-layer Bi-LSTM. The learning rate was set to 1.0×10^{-3} empirically. During the training time, LSTM was unrolled for 60 iterations. Therefore, the gradient horizon for LSTM was 60 seconds, which is the length of each video that we used in the experiment. The evaluation measure used in the experiment was F-measure. We compared the proposed method with baseline methods and some state-of-the-art methods for personal liking estimation using a dataset that we created. The dataset consists of 50 videos and EEG signals captured from 11 participants. In the experiment, we performed time patterns of evaluation of our method for verifying its generalizability. First, we focused on within-subject study, and a user-dependent model (UDM) was constructed for each subject. Then leave-one-video-out cross-validation, which tested on each video in turn for each subject, was performed. On the other hand, we focused on between-subject study, and a user-independent model (UIM) was constructed. We performed leave-one-subject-out cross-validation, which tested on all of the videos that each subject watched.

A. DATASETS

In the experiment, we prepared 50 videos for evaluation. Specifically, we obtained 50 movie trailer videos based on [14]–[16] by inputting the query keyword "movie trailer" to YouTube.[2] These videos included the following five genres: science fiction, comedy, action, horror and romance. Each genre included the same number of videos, i.e., 10 videos. The length of each video was 60 secs. Eleven subjects aged from 22 to 24 years participated in the experiment. The subjects were instructed to watch all of the 50 videos. Each subject sat on a chair, and the distance from the display to the subject was about 0.5 meters. The resolution of the display was 1920 × 1080 pixels, and all of the videos were shown in a full screen mode. For obtaining EEG signals, we used alphatec IV-s as described above. The EEG features were obtained every second.

In the experiment, each subject performed four grade[3] evaluation of all of the videos after watching them. Then we prepared datasets including videos, evaluation scores and EEG signals.

In the experiment, we classified the videos into two classes, "Liked video" and "Not Liked video". The class "Liked video" includes videos rated 3 or 4, and the class "Not Liked video" includes videos rated 1 or 2. The number of videos included in each class for each subject is shown in Table 1.

TABLE 1. The numbers of videos included in classes "Liked video" and "Not Liked video" per subject. "Li" and "NLi" represent "Liked video" and "Not Liked video", respectively.

Subject	A	B	C	D	E	F	G	H	I	J	K
Li	30	20	30	31	10	29	24	23	25	26	32
NLi	20	30	20	19	40	21	26	27	25	24	18

B. COMPARED APPROACHES

In order to verify the effectiveness of the proposed method, we compared the proposed method with the following ten baseline methods (former ten) and three state-of-the-art methods (latter three) for personal liking estimation. All of the comparative methods were individually tuned to achieve the best performance for fair comparisons.

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

[3] 1=Not at all liked, 4=Extremely liked

(1) **Video + Bi-LSTM**: This method uses only representation of videos for WUL-based video classification. The video representation is generated by our proposed approach. The classification of a video is based on the Bi-LSTM network.

(2) **EEG + Bi-LSTM**: This method uses only representation of EEG for WUL-based video classification. The EEG representation is generated by our proposed approach. The classification of a video is based on the Bi-LSTM network.

(3) **Video and EEG + FW-LSTM**: This method uses representation of videos and that of EEG for WUL-based video classification. The video and EEG representations are generated by our proposed approach. The classification of a video is based on the FW-LSTM network.

(4) **Video + FW-LSTM**: This method uses only representation of videos for WUL-based video classification. The video representation is generated by our proposed approach. The classification of a video is based on the FW-LSTM network.

(5) **EEG + FW-LSTM** [26]: This method uses only representation of EEG for WUL-based video classification. The EEG representation is generated by our proposed approach. The classification of a video is based on the FW-LSTM network.

(6) **Video and EEG + BW-LSTM**: This method uses representation of videos and that of EEG for WUL-based video classification. The video and EEG representations are generated by our proposed approach. The classification of a video is based on the BW-LSTM network.

(7) **Video + BW-LSTM**: This method uses only representation of videos for WUL-based video classification. The video representation is generated by our proposed approach. The classification of a video is based on the BW-LSTM network.

(8) **EEG + BW-LSTM**: This method uses only representation of EEG for WUL-based video classification. The video representation is generated by our proposed approach. The classification of a video is based on the BW-LSTM network.

(9) **Video + Average pooling**: Representation of a video is the average for all frames. The final classification result is obtained by the softmax function.

(10) **EEG + Average pooling**: Representation of EEG is the average for all frames. The final classification result is obtained by the softmax function.

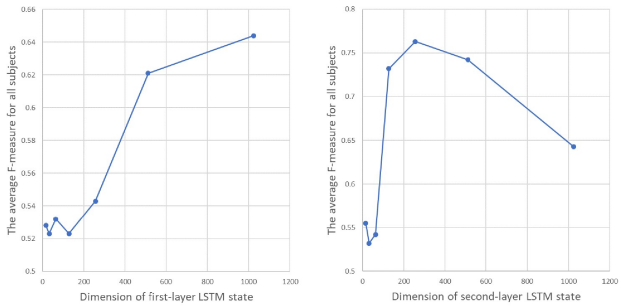
(11) **Koelstra *et al.*** [28]: In this method, EEG features including theta (4–8 Hz), slow alpha (8–10 Hz), alpha (8–12 Hz), beta (12–30 Hz) and gamma (30– Hz) spectral power are extracted. Fisher's linear discriminant is applied to the features, and Gaussian naive Bayes is used for low/high liking classification.

(12) **Yoon and Chung** [41]: Fast Fourier Transform analysis and Pearson's correlation coefficient-based feature selection are applied to EEG to extract effective features for emotion classification. Emotion classification is realized by a classifier on the basis of Bayes theorem and supervised learning using a perceptron convergence algorithm.

(13) **Naser and Saha** [42]: Feature extraction is performed by dual-tree complex wavelet packet transform. Furthermore, redundant feature elimination is performed based on singular value decomposition, QR factorization with column pivoting and F-ratio. Emotion classification is performed by a support vector machine [52].

C. EXPERIMENTAL RESULTS IN THE UDM SETTING

Experimental results in the UDM setting, i.e., for within-subjects, are shown in this subsection. The proposed method includes two important parameters: dimensions for the hidden states in the first-layer and second-layer LSTM networks. Thus, we confirmed the relationship between their dimensions and the average F-measure in a manner similar to that in [53]. Note that the dimension of the first or second layer was changed from 16 to 1024 with the dimension of the other layer being fixed to 512. The obtained relationship is shown in Fig. 3. The performance of our method is best when the dimensions of the first-layer and second-layer LSTM hidden states are 1024 and 256, respectively.



(a) (b)

FIGURE 3. Relationship between the dimension of the hidden states and performance in the UDM setting. The horizontal and vertical axes are the dimension and the average F-measure of the proposed method, respectively. (a) First-layer. (b) Second-layer.

1) COMPARISON WITH BASELINE METHODS

Table 2 shows the results obtained by using our method compared with results obtained by using the baseline methods. The results presented in the table indicate the following. 1) The proposed method almost outperforms the baseline methods. Specifically, the F-measure of the proposed method is much higher than the F-measures of **Video + Bi-LSTM** and **EEG + Bi-LSTM**, indicating the effectiveness of collaborative use of video and EEG representations with Bi-LSTM. 2) The methods using FW-LSTM outperform those using BW-LSTM. We can assume that the last few video frames in most videos correspond to salient scenes, and it is harder for BW-LSTM to obtain effective representations for WUL-based classification. 3) In the results for Subject H, the methods using only video representation outperformed other methods. This is because Subject H tended to like action and comedy videos. In the experiment, the genre of the most liked videos rated by Subject H (16 out of 23) was

TABLE 2. Comparison of the performances of our method and the baseline methods. The F-measure was calculated to verify the performance of favorite video classification for all subjects (A-K).

	A	B	C	D	E	F	G	H	I	J	K	Average
Ours	**0.763**	**0.695**	**0.721**	0.691	**0.752**	**0.764**	**0.631**	0.654	**0.595**	**0.596**	**0.667**	**0.684**
Video + Bi-LSTM	0.678	0.643	0.652	0.648	0.621	0.645	0.595	0.545	0.458	0.556	0.633	0.607
EEG + Bi-LSTM	0.662	0.644	0.663	0.671	0.701	0.674	0.608	0.638	0.584	0.595	0.642	0.643
Video and EEG + FW-LSTM	0.709	0.640	0.665	0.682	**0.752**	0.709	0.591	0.580	0.589	0.584	**0.667**	0.651
Video + FW-LSTM	0.677	0.635	0.646	0.646	0.615	0.638	0.581	0.505	0.427	0.551	0.632	0.596
EEG + FW-LSTM [26]	0.613	0.613	0.623	0.701	0.694	0.621	0.591	0.567	0.543	0.591	0.644	0.618
Video and EEG + BW-LSTM	0.688	0.657	0.645	**0.712**	0.743	0.712	0.588	0.563	0.589	0.584	0.644	0.648
Video + BW-LSTM	0.665	0.645	0.655	0.621	0.556	0.621	0.591	0.446	0.543	0.532	0.453	0.575
EEG + BW-LSTM	0.623	0.601	0.567	**0.712**	0.654	0.643	0.589	0.577	0.531	0.588	0.621	0.610
Video + Average pooling	0.667	0.571	0.577	0.353	0.500	0.519	0.560	**0.800**	0.566	0.558	0.553	0.566
EEG + Average pooling	0.643	0.556	0.512	0.522	0.542	0.523	0.542	0.487	0.453	0.324	0.534	0.513

action or comedy. 4) The F-measure of **EEG + Bi-LSTM** is higher than the F-measures of **EEG + FW-LSTM** [26] and **EEG + Average pooling**. Therefore, methods using EEG representation with LSTM networks such as Bi-LSTM are effective for WUL-based classification. We also mention this point in the comparison with state-of-the-art methods.

2) COMPARISON WITH STATE-OF-THE-ART METHODS

Table 3 summarizes the performances of the proposed method and state-of-the-art methods. As expected, our method outperformed previous methods that only use conventional EEG representations, and the effectiveness of the proposed Bi-LSTM-based EEG representation for favorite video classification was confirmed. This is because LSTM networks have a powerful ability to learn representations from raw EEG signals directly.

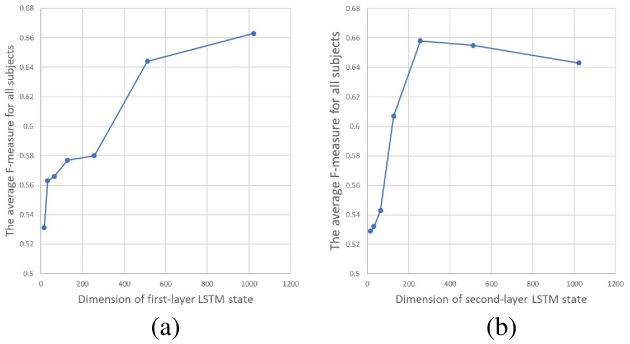
TABLE 3. Results in the UDM setting. Comparison with state-of-the-art methods. Metrics is the average F-measure for all subjects.

Method	The average F-measure
Ours	**0.684**
Koelstra *et al.* [28]	0.510
Yoon and Chung [41]	0.483
Naser and Saha [42]	0.536

D. EXPERIMENTAL RESULTS IN THE UIM SETTING

Experimental results in the UIM setting, i.e., for between-subjects, are shown in this subsection. As shown in the results for within-subjects, our method outperformed the comparative methods in the UDM setting. On the other hand, creating a UIM should also be discussed to analyze the user's tendency in multimedia applications and to solve the cold start problem in multimedia applications. We therefore constructed UIMs using only the proposed EEG representation.

As in the study for within-subjects, we first investigated the relationships between the dimensions and performance of the method of **EEG + Bi-LSTM**, which uses the proposed EEG representation with Bi-LSTM, in the same manner as that in the previous subsection. The obtained results are shown in Fig. 4. The performance of the method of **EEG + Bi-LSTM** was best when the first and second LSTM hidden states were 1024 and 256, respectively.



(a) (b)

FIGURE 4. Relationship between the dimension of hidden states and performance in the UIM setting. The horizontal and vertical axes are the dimension and the average F-measure of the proposed method, respectively. (a) First-layer. (b) Second-layer.

Figure 5 shows the performances of UIM and UDM for each subject using the method of **EEG + Bi-LSTM**. Interestingly, UIM (average F-measure = 0.676) outperforms UDM (average F-measure = 0.643). Furthermore, as we can see from the results for Subject H, the F-measure in the UIM setting is much higher than that in the UDM setting. We can see the same improvement from the results for other subjects whose liked videos cover several genres (all subjects except for H). This is because the behavior of EEG signals captured while subjects are feeling "like" is generally the same across all subjects, and the trait of the signals is effective for constructing a general model for WUL-based video classification.

Table 4 further shows the performances of the user-independent model using **EEG + Bi-LSTM** and state-of-the-art methods. The same improvement as that for the results in the UDM setting can be seen in Table 4. The proposed EEG representation with Bi-LSTM outperforms previous methods that only use conventional EEG representations, which again shows the effectiveness of the proposed Bi-LSTM-based EEG representation for favorite video classification.

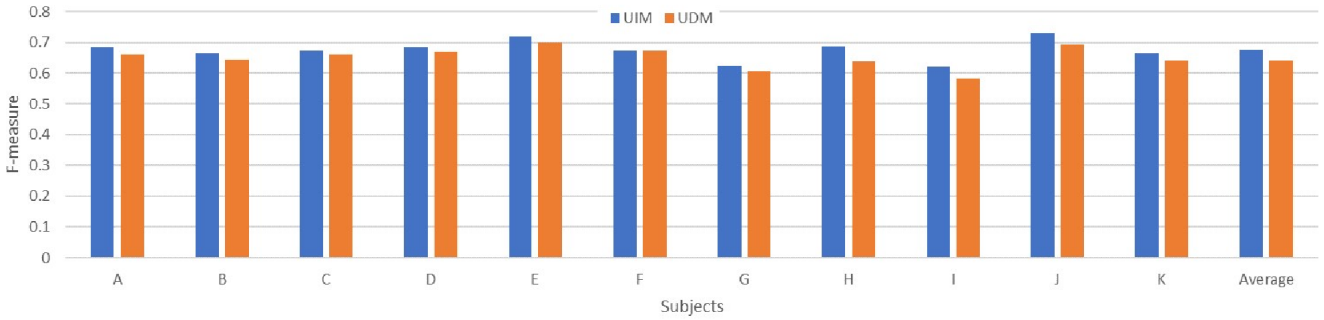


FIGURE 5. F-measures of EEG + Bi-LSTM in the UIM setting and those in the UDM setting for all subjects.

TABLE 4. Results in the UIM setting. This table also shows a comparison between our method and state-of-the-art methods. Metrics is the average F-measure for all subjects.

Method	The average F-measure
EEG + Bi-LSTM	**0.663**
EEG + FW-LSTM [26]	0.653
EEG + BW-LSTM	0.637
EEG + Average pooling	0.552
Koelstra *et al.* [28]	0.566
Yoon and Chung [41]	0.544
Naser and Saha [42]	0.612

V. CONCLUSIONS

A novel method for WUL-based video classification with multimodal Bi-LSTM is presented in this paper. The results of multiuser experiments demonstrated the effectiveness of WUL-based video classification. We also showed the effectiveness of the proposed EEG and video representations for WUL-based video classification. In both UDM and UIM settings, the proposed EEG representation with Bi-LSTM outperformed other conventional methods. Consequently, we can realize an accurate method for favorite video classification via collaborative use of WUW and WUL. Note that our study is a trial that realizes a feasible method for WUL-based video classification with Bi-LSTM. Since audio cues play an important role in affective state recognition, modeling collaborative use of audio and EEG signal representation is one of our future research directions.

REFERENCES

[1] J. Gu *et al.* (Dec. 2015). "Recent advances in convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1512.07108
[2] B. Wang and M. Larson, "Beyond concept detection: The potential of user intent for image retrieval," in *Proc. ACM Workshop Multimodal Understand. Social, Affective Subjective Attributes*, 2017, pp. 11–19.
[3] M. Yan, J. Sang, and C. Xu, "Unified YouTube video recommendation via cross-network collaboration," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 19–26.
[4] H. Grabner, F. Nater, M. Druey, and L. Van Gool, "Visual interestingness in image sequences," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 1017–1026.
[5] Y. Hou *et al.*, "Predicting movie trailer viewer's 'like/dislike' via learned shot editing patterns," *IEEE Trans. Affective Comput.*, vol. 7, no. 1, pp. 29–44, Jan./Mar. 2016.
[6] Q. Zhu, M.-L. Shyu, and H. Wang, "VideoTopic: Content-based video recommendation using a topic model," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2013, pp. 219–222.
[7] R. M. A. Teixeira, T. Yamasaki, and K. Aizawa, "Determination of emotional content of video clips by low-level audiovisual features," *Multimedia Tools Appl.*, vol. 61, no. 1, pp. 21–49, 2012.
[8] S. Zennaro *et al.*, "Performance evaluation of the 1st and 2nd generation Kinect for multimedia applications," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun./Jul. 2015, pp. 1–6.
[9] R.-D. Vatavu, "Audience silhouettes: Peripheral awareness of synchronous audience kinesics for social television," in *Proc. ACM Int. Conf. Interact. Exper. TV Online Video*, 2015, pp. 13–22.
[10] H. Ye, M. Malu, U. Oh, and L. Findlater, "Current and future mobile and wearable device use by people with visual impairments," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2014, pp. 3123–3132.
[11] J. Hernandez, Y. Li, J. M. Rehg, and R. W. Picard, "BioGlass: Physiological parameter estimation using a head-mounted wearable device," in *Proc. EAI 4th Int. Conf. Wireless Mobile Commun. Healthcare (Mobihealth)*, 2014, pp. 55–58.
[12] S. Ishimaru *et al.*, "In the blink of an eye: Combining head motion and eye blink frequency for activity recognition with Google Glass," in *Proc. 5th Augmented Hum. Int. Conf.*, 2014, Art. no. 15.
[13] Y. Sasaka, T. Ogawa, and M. Haseyama, "A novel framework for estimating viewer interest by unsupervised multimodal anomaly detection," *IEEE Access*, vol. 6, pp. 8340–8350, 2018.
[14] S. H. Fairclough, A. J. Karran, and K. Gilleade, "Classification accuracy from the perspective of the user: Real-time interaction with physiological computing," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, 2015, pp. 3029–3038.
[15] S. Liu *et al.*, "What makes a good movie trailer?: Interpretation from simultaneous eeg and eyetracker recording," in *Proc. ACM Conf. Multimedia*, 2016, pp. 82–86.
[16] Y. Sasaka, T. Ogawa, and M. Haseyama, "Multimodal interest level estimation via variational Bayesian mixture of robust CCA," in *Proc. ACM Conf. Multimedia*, 2016, pp. 387–391.
[17] M. Pantic and A. Vinciarelli, "Implicit human-centered tagging [Social Sciences]," *IEEE Signal Process. Mag.*, vol. 26, no. 6, pp. 173–180, Nov. 2009.
[18] M. Soleymani and M. Pantic, "Human-centered implicit tagging: Overview and perspectives," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2012, pp. 3304–3309.
[19] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of YouTube videos," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 365–374.
[20] A. Brodersen, S. Scellato, and M. Wattenhofer, "YouTube around the world: Geographic popularity of videos," in *Proc. ACM 21st Int. Conf. World Wide Web*, 2012, pp. 241–250.
[21] A. Graves *et al.*, *Supervised Sequence Labelling With Recurrent Neural Networks*, vol. 385. Berlin, Germany: Springer, 2012.
[22] Y. Bin, Y. Yang, Z. Huang, F. Shen, X. Xu, and H. T. Shen, "Bidirectional long-short term memory for video description," in *Proc. ACM Multimedia Conf.*, 2016, pp. 436–440.
[23] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, "Learning fashion compatibility with bidirectional LSTMs," in *Proc. ACM Multimedia Conf.*, 2017, pp. 1078–1086.

[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[26] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion recognition based on EEG using LSTM recurrent neural network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 10, pp. 355–358, 2017.
[27] J. A. Russell, "Culture and the categorization of emotions," *Psychol. Bull.*, vol. 110, no. 3, pp. 426–450, 1991.
[28] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, Oct./Mar. 2012.
[29] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.
[30] D. E. Berlyne, *Conflict, Arousal, and Curiosity*. New York, NY, US: McGraw-Hill, 1960.
[31] P. J. Silvia, "Interest—The curious emotion," *Current Directions Psychol. Sci.*, vol. 17, no. 1, pp. 57–60, 2008.
[32] P. J. Silvia, "Confusion and interest: The role of knowledge emotions in aesthetic experience," *Psychol. Aesthetics, Creativity, Arts*, vol. 4, no. 2, pp. 75–80, 2010.
[33] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, Nov. 2015.
[34] Q. Gan, S. Wang, L. Hao, and Q. Ji, "A multimodal deep regression Bayesian network for affective video content analyses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2017, pp. 5113–5122.
[35] M. Takahashi, S. Clippingdale, M. Naemura, and M. Shibata, "Estimation of viewers' ratings of TV programs based on behaviors in home environments," *Multimedia Tools Appl.*, vol. 74, no. 19, pp. 8669–8684, 2014.
[36] J. M. Henderson, "Human gaze control during real-world scene perception," *Trends Cognit. Sci.*, vol. 7, no. 11, pp. 498–504, 2003.
[37] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 631–637.
[38] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 500–508, Jun. 2006.
[39] H. T. Le and L. A. Vea, "A customer emotion recognition through facial expression using Kinect sensors v1 and v2: A comparative analysis," in *Proc. 10th Int. Conf. Ubiquitous Inf. Manage. Commun.*, 2016, Art. no. 80.
[40] J. K. Tsotsos, L. Itti, and G. Rees, "A brief and selective history of attention," in *Neurobiology of Attention*. New York, NY, USA: Academic, 2005.
[41] H. J. Yoon and S. Y. Chung, "EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm," *Comput. Biol. Med.*, vol. 43, no. 12, pp. 2230–2237, 2013.
[42] D. S. Naser and G. Saha, "Recognition of emotions induced by music videos using DT-CWPT," in *Proc. IEEE Indian Conf. Med. Inform. Telemedicine (ICMIT)*, Mar. 2013, pp. 53–57.
[43] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "DevNet: A deep event network for multimedia event detection and evidence recounting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2568–2577.
[44] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
[45] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 843–852.
[46] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 961–970.
[47] F. Li *et al.* (Jul. 2017). "Temporal modeling approaches for large-scale Youtube-8m video understanding." [Online]. Available: https://arxiv.org/abs/1707.04555
[48] S. Abu-El-Haija *et al.* (Sep. 2016). "Youtube-8m: A large-scale video classification benchmark." [Online]. Available: https://arxiv.org/abs/1609.08675
[49] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.
[50] I. Crk, T. Kluthe, and A. Stefik, "Understanding programming expertise: An empirical study of phasic brain wave changes," *ACM Trans. Comput.-Hum. Interact.*, vol. 23, no. 1, 2016, Art. no. 2.
[51] F. Cong *et al.*, "Linking brain responses to naturalistic music through analysis of ongoing EEG and stimulus features," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1060–1069, Aug. 2013.
[52] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[53] Y. Ye, Z. Zhao, Y. Li, L. Chen, J. Xiao, and Y. Zhuang, "Video question answering via attribute-augmented attention network learning," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2017, pp. 829–832.

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 Graduate School of Information Science and Technology, Hokkaido University. His research interests are multimedia signal processing and its applications. He has been an Associate Editor of the *ITE Transactions on Media Technology and Applications*. He is a member of the EURASIP, IEICE, and the Institute of Image Information and Television Engineers.

YUMA SASAKA (S'15) received the B.S. degree in electronics and information engineering from Hokkaido University, Japan, in 2016, where he is currently pursuing the M.S. degree with the Graduate School of Information Science and Technology. His research interests include biosignal processing and video information retrieval. He is a Student Member of the ACM.

KEISUKE MAEDA (S'14) received the B.S. and M.S. degrees in electronics and information engineering from Hokkaido University, Japan, in 2015 and 2017, respectively, where he is currently pursuing the Ph.D. degree with the Graduate School of Information Science and Technology. His research interests are multimodal processing and its applications. He is currently a Research Fellow of the Japan Society for the Promotion of Science. He is a Student Member of the IEICE.

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 Graduate School 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, ITE, and the Information Processing Society of Japan IPSJ. She has been the Vice-President of the Institute of Image Information and Television Engineers, Japan (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 Institute of Electronics, Information and Communication Engineers (IEICE).