IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 21, 2017, accepted February 2, 2018, date of publication February 12, 2018, date of current version March 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2804925

A Novel Framework for Estimating Viewer Interest by Unsupervised Multimodal Anomaly Detection

YUMA SASAKA, (Student Member, IEEE), TAKAHIRO OGAWA, (Member, IEEE), AND MIKI HASEYAMA, (Senior Member, IEEE)
Graduate School of Information Science and Technology, Hokkaido University, Hokkaido 060-0814, Japan
Corresponding author: Yuma Sasaka (sasaka@lmd.ist.hokudai.ac.jp)

This work was supported by JSPS KAKENHI under Grant JP17H01744 and Grant JP15K12023.

ABSTRACT A reliable method to estimate viewer interest is highly sought after for human-centered video information retrieval. A method that estimates viewer interest while users are watching Web videos is presented in this paper. The method uses a framework for anomaly detection based on collaborative use of facial expression and biological signals such as electroencephalogram (EEG) signals. To the best of our knowledge, there have been no studies that have taken into account two actual mechanisms of the behavior of users while they are watching Web videos. First, whereas most Web videos garner very little attention, a small number attract millions of views. Therefore, a framework for anomaly detection is newly applied to facial expression and EEG in order to model the imbalanced distribution of popularity. Second, since the number of Web videos that are labeled by users as interesting/not interesting is generally too small to estimate viewer interest by a supervised approach, the proposed method utilizes parametric techniques for anomaly detection, which estimates viewer interest in an unsupervised way. Unlike some related studies for estimating viewer interest, our method takes into account actual mechanisms of the behavior of users while they are watching Web videos by utilizing parametric techniques for anomaly detection. Then viewer interest can be estimated on the basis of an anomaly score calculated from our proposed method. Consequently, successful estimation of viewer interest based on a framework for anomaly detection, via collaborative use of facial expression and biological signals, becomes feasible.

INDEX TERMS Viewer interest, unsupervised anomaly detection, facial expression, biological signals.

I. INTRODUCTION

It has become possible to access videos on the Web (hereafter Web videos) owing to the development of services such as YouTube[1] and NETFLIX[2] that recommend Web videos. The number of Web videos has been constantly increasing because of user uploads and diversified services that recommend Web videos [1]. Many recommendation methods have therefore been studied in order to provide Web videos that match the desire of a user [2]–[11]. Recommendation methods that are based on individual Web video preference such as interesting/not interesting have been studied in recent years [9]–[11]. These methods model individual Web video preference by extracting image and audio features from Web videos (hereafter Web video features), and effective Web video recommendation is realized through Web video classification based on the model. However, these Web video features, which are generally not related to individual Web video preference, deteriorate the performance of Web video classification. Furthermore, it is difficult to find a set of Web video features related to individual Web video preference since the particular element of a Web video that stimulates a user is different for each user. Therefore, it is necessary to introduce a new idea that solves the problem by estimating individual Web video preference based on alternative features.

Recently, with the development of various types of sensor technology [12]–[19], some studies have focused on psychophysiological data, such as facial expression and electroencephalogram (EEG) data, to estimate individual Web

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

video preference [20]–[23]. Facial expression is a major cue in measuring viewer interest while users are watching Web videos [20], and biological signals such as EEG signals are also used to estimate viewer interest [21], [22]. The use of both facial expression and biological signals collaboratively is effective for enhancing performance of viewer interest estimation [23]. Observation of biological signals has become easier, and the quality of observed signals has been improved due to the development and miniaturization of biological sensors [17]–[19]. Results of those studies indicate that biological signals are also effective cues for estimating viewer interest. Thus, both facial expression and biological signals are widely used.

Motivated by the aforementioned discussion, several methods for estimating viewer interest have been proposed [20]–[23]. However, to the best of our knowledge, most approaches including the studies for estimating viewer interest do not take into account "actual mechanisms of users while they are watching Web videos", which are reported in [6] and [24]–[27]. Specifically, there are the following two mechanisms.

1) Since users tend to watch Web videos that they would prefer to by obtaining information about the Web videos through social networking services (SNSs) and word-of-mouth [24], most Web videos garner very little attention, whereas a small number of Web videos attract millions of views [25], [26].
2) Due to the ever increasing number of available Web videos, the number of Web videos that are labeled by users as interesting/not interesting is too small [6], [27] to estimate viewer interest by a supervised approach.

Related studies needed a large amount of training data to estimate viewer interest based on the assumption that users watch a sufficiently large number of both interesting and not interesting Web videos to construct classifiers. However, considering the first mechanism, the number of uninteresting videos watched by users is small. Consequently, the first mechanism was not considered in related studies, and the feasibility of methods proposed for viewer interest estimation is low. Since a supervised approach was used in related studies for estimating viewer interest using class labels that are assigned by users, the second mechanism was also not considered in related studies. Thus, the feasibility of methods proposed for viewer interest estimation is clearly low. Therefore, to take into account the actual mechanisms, we newly define the notion of "estimating viewer interest" used in the related studies [20]–[23] as that of "estimating high viewer interest for particularly interesting Web videos" as shown in Fig. 1. Based on the new notion, an unsupervised method, which estimates viewer interest for Web videos that users selectively watched, is desirable to realize a feasible framework for viewer interest estimation. However, in the context of estimating viewer interest, such a method that is suitable for using facial expression and biological signals has not been proposed as far as we know.

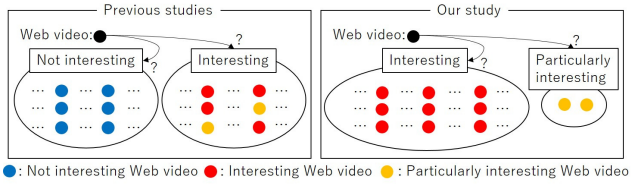

FIGURE 1. Difference between the notion of estimating viewer interest in previous studies and that in our study.

In this paper, we propose a novel method for estimating viewer interest, which can be regarded as a psychological state evoked by multimedia [20]–[23], via collaborative use of facial expression and biological signals for Web video recommendation. Unlike previous studies for estimating viewer interest, the proposed method takes into account the two mechanisms of users while they are watching Web videos. In order to do so, associated with the concept of anomaly detection [28], i.e., finding patterns in data that do not conform to expected behavior, we newly derive an anomaly detection approach that is suitable for using facial expression and biological signals. By using the framework, our method can estimate viewer interest even if the number of uninteresting videos watched by users is small. Moreover, by utilizing parametric techniques [29] for anomaly detection, our method estimates viewer interest by an unsupervised approach even if the number of Web videos that are labeled by users as interesting/not interesting is too small to estimate viewer interest.

In this paper, considering the two mechanisms of users, we only focus on the validation of using the framework for anomaly detection as the first method towards realizing feasible estimation of viewer interest. This is the placement of our method. In summary, the contributions of our study are threefold:

- In the context of estimating viewer interest, we newly derive an anomaly detection approach that is suitable for using facial expression and biological signals.
- By using a framework for anomaly detection, our method estimates viewer interest even if the number of uninteresting Web videos watched by users is small.
- By utilizing parametric techniques for anomaly detection, our method estimates viewer interest by an unsupervised approach.

The remainder of this paper is organized as follows. In section II, a model of "Interest" for multimedia application is presented. In section III, a brief review of related studies is given. In section IV, extraction of two different features, features of facial expression and those of biological signals, used in our method is presented. In section V, we explain the proposed method. In section VI, the results of experiments to verify the effectiveness of the proposed method are presented. Finally, the conclusion of this paper is given in section VII.

II. A MODEL OF INTEREST FOR MULTIMEDIA APPLICATION

In this section, we define the concept of viewer interest based on several studies. We firstly explain factors of ''Interest'' in II-A. We then derive the model of viewer interest in our study in II-B.

A. FACTORS OF "INTEREST"

''Interest'' as a psychological entity has been explained by Berlyne [30]. According to Berlyne, ''interest'' as an emotional state fosters curiosity and the drive to explore the object or situation at hand. Basically, the key measures are increased arousal and sensation seeking, i.e., objects inspire curiosity via novelty and emotional conflict. Silva [31], [32] incorporated a cognitive perspective to the measures, by which ''interest'' is driven by stimulus complexity.

Emotional resonances can also be considered as a key measure of ''interest'' for cultural heritage experiences [33]. In the context of using multimedia applications, emotional resonances that accompany novelty or complexity also represent an important component of ''interest''. For example, when a viewer watches a moving story of an animal, the viewer may feel sorry for the animal. On the other hand, when a viewer watches funny comedians and hears their jokes, the viewer may feel happy. The experience of ''interest'' is followed by a sense of positive emotion derived from intellectual engagement [34]. Therefore, emotions and cognitions are important factors of the conceptualization of ''interest'' for multimedia applications.

As discussed above, based on the works of Berlyne [30] and Silva [31], we developed a model that consists of five parts: three cognitive in nature and cognitive factors (former three) and two emotional in nature (latter two), as shown below.

- *Comprehension*: whether the representation/function of the multimedia content is clearly understandable
- *Complexity*: whether the perceptual complexity of the multimedia content is high or low
- *Novelty*: how familiar or unusual the object is
- *Valence*: whether viewing the multimedia content makes the person feel happy or sad
- *Attractiveness*: how attractive the object is

According to an online survey reported in [33] in which the number of subjects was over 1000, *Novelty* is closely correlated to *Comprehension* and *Complexity* in a cultural heritage context. This is also true in the context of using multimedia applications. As an illustration, novel Web videos are complex and consequently difficult to understand. It was also shown in the survey [33] that *Attraction* is positively related to *Valence* in a cultural heritage context. When using multimedia content, we generally feel happy if the objects are attractive. In summary, ''interest'' representing user's experiences within the context of using multimedia applications needs to include cognitive factors and emotional factors. We therefore define the factors of interest as consisting of a cognitive factor (representing a high/low cognitive response to the complexity/ease of comprehension associated with the stimulus) and an emotional factor, valence (positive/negative affect), as shown in Fig. 2.

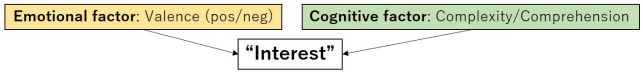

FIGURE 2. A cognitive factor (representing a high/low cognitive response to the complexity/ease of comprehension associated with the stimulus) and an emotional factor, valence (positive/negative affect), creating factors of "interest".

B. MODEL OF VIEWER INTEREST IN OUR STUDY

As mentioned in the previous subsection, we constructed a method for viewer interest estimation based on two factors. To focus on the validation of using the framework for anomaly detection, as the first method towards realizing feasible estimation of viewer interest, we model the cognitive component and valence. Specifically, the cognitive component of interest is inferred from activation of the rostral prefrontal cortex; this variable can be captured using spontaneous EEG measures of electrocortical activation. Valence (positive/negative) can be captured by facial expression [20], [23]. Again, we use facial expression and EEG to estimate viewer interest since these two features are widely used and we are able to capture them without distracting viewers while they are watching Web videos.

III. RELATED WORK

Various methods for estimating viewer interest have been proposed. In this section, we give a brief review of the sources of information they use: low level features calculated from multimedia content and features captured from humans. In addition, we describe the validity of the features of facial expression and those of biological signals used in our method.

A. METHODS BASED ON FEATURES CALCULATED FROM MULTIMEDIA CONTENT

Estimation of viewer interest is a big challenge for Web content recommendation and retrieval. There have been several efforts to estimate viewer interest utilizing only Web video features [9]–[11], [35]. Grabner *et al.* [35] analyzed image sequences recorded by a static video camera to predict the parts in an image sequence that are considered interesting by many viewers. In [9]–[11], low level features were analyzed to model viewers' video preference. Despite a number of studies, there is a common agreement that visual features extracted from videos are mainly used. In addition to visual features, motion and shot features were used in [9] to reflect a field or objects changing rate in images of videos and to find key frames of videos, while textual features that can be represented by a sparse histogram over vocabularies were extracted in [10] Furthermore, audio features were extracted from video clips in [11]

As mentioned in section I, Web video features are not related to individual Web video preference since they are

different from person to person. Therefore, it is difficult to create a general model for Web video preference estimation using usual pattern recognition approaches. Moreover, because of the second mechanism mentioned in section I, creating personalized models is also difficult.

B. METHODS BASED ON FEATURES CAPTURED FROM HUMANS

With the development of various types of sensor technology, utilization of features captured from humans has been investigated to estimate ''interest'' [20]–[23], [36]–[39].

The concept of ''interest'' as human attention has mostly been studied through understanding which visual stimuli can attract human attention [40]. In several studies, gaze patterns of subjects watching images or videos were recorded and the data were analyzed [36]–[38] on the basis of the concept. However, we should point out that the concept is not necessarily equivalent to viewer interest. If a person watches an image or a Web video in order to understand what is happening, this does not mean that he or she really considers the observations as interesting.

Unlike the studies mentioned above, ''interest'' was considered as an affective state in some studies [20]–[23], [39]. Yeasin *et al.* [20] presented a spatio-temporal approach for recognizing six universal facial expressions and using the video sequences of facial expressions to compute levels of viewer interest. They finally showed the results of estimating viewer interest level in a positive/negative manner. Le and Vea [39] focused on both positive and negative emotional states of customers' facial expressions towards products. They reported that facial expressions are major cues to model product purchase intentions. For obtaining facial expression data, they used Kinect v2, which has a great ability to see and respond to a multitude of interactions and plays a key role in the development of consumer depth sensors in various types of markets [12]–[16]. On the other hand, EEG was used in some psychological and neurological studies for estimating viewer interest [21]–[23]. Liu *et al.* [22] proposed a platform to simultaneously record EEG and eye movement for subjects with video stimuli for an automatic movie trailer evaluation system. Similarly, Fairclough *et al.* [21] classified psychophysiological data such as EEG data to estimate viewer interest in a binary high/low manner. The EEG signals were captured from subjects while they were watching movie trailers.

Given the studies mentioned above, it is assumed that facial expression and EEG signals are important measures to estimate viewer interest. In our previous study [23], we captured both facial expression and EEG data from subjects and analyzed them by machine learning methods. However, as mentioned in section I, most approaches including our previous study for estimating viewer interest have not taken into account ''actual mechanisms of users while they are watching Web videos''

TABLE 1. AUs used for calculating features of facial expression in the proposed method.

Jaw Open	Left cheek Puff
Lip Pucker	Right cheek Puff
Jaw Slide Right	Left eye Closed
Lip Stretcher Right	Right eye Closed
Lip Stretcher Left	Right eye brow Lowerer
Lip Corner Puller Left	Left eye brow Lowerer
Lip Corner Puller Right	Lower lip Depressor Left
Lip Corner Depressor Left	Lower lip Depressor Right
Lip Corner Depressor Right	

IV. FEATURE EXTRACTION

In this section, we explain extraction of the two different features, features of facial expression and those of biological signals, used in our method.

A. FEATURES OF FACIAL EXPRESSION

In this subsection, we explain the extraction of features of facial expression. We obtain facial expression by using Kinect for Windows v2,[3] which is playing a key role in the development of consumer depth sensors in various types of markets [12], [13]. First, we detect 17 animation units (AUs)[4] on the face, with the AUs being automatically selected by Kinect. In our method, we measure the degree of change from a numeric weight varying between 0 and 1. More detailed information is not available to the public. However, the features have been used in several related studies [23], [39]. The names of AUs are shown in Table 1. We then capture the movements 15 times per second. In order to extract autonomic features, we calculate the mean from 10-sec data windows, with a sliding window of 5 sec. In this way, we capture the movements of each part of the face (17 dimensions) as features of facial expression $\boldsymbol{x}_n^F \in \mathbb{R}^{17}$ ($n = 1, 2, \cdots, N$; N being the number of samples).

B. FEATURES OF BIOLOGICAL SIGNALS

In this subsection, we explain the extraction of features of biological signals. Since the use of neurophysiological measures is relatively new in computer science, especially as they pertain to watching a Web video, we extract features of biological signals from EEG, based on prior studies using EEG in multimedia applications [21]–[23], [33], [41], [42]. Biological signals are obtained by using alphatec IV-s. The sampling rate for EEG signals is 1024 Hz. Alphatec IV-s is used for obtaining EEG signals from Fp1, and EEG signals are closely related to the degree of interest of users [33]. Segmentation of EEG signals is performed at a fixed interval with an overlapped Hamming window. Note that the length of the window is also 10 seconds with a sliding window of 5 sec. We then apply short-time Fourier transform (STFT) to each

[3]https://www.microsoft.com/en-us/kinectforwindows/

[4]https://msdn.microsoft.com/en-us/library/microsoft.kinect.face.faceshapeanimations.aspx

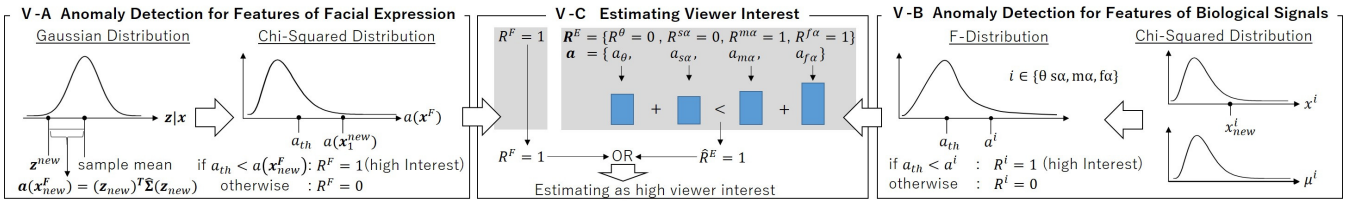

FIGURE 3. Overview of the proposed method.

TABLE 2. Details of features calculated from signals obtained by alphatec IV-s.

INDEX	DESCRIPTION	RANGE
x_θ	Power spectrum of θ wave	4-7Hz
$x_{s\alpha}$	Power spectrum of slow-α wave	7-9Hz
$x_{m\alpha}$	Power spectrum of mid-α wave	9-11Hz
$x_{f\alpha}$	Power spectrum of fast-α wave	11-13Hz

segment. Consequently, we obtain the power spectrum of δ wave (1-4 Hz), θ wave (4-7 Hz), slow-α wave (7-9 Hz), mid-α wave (9-11 Hz), fast-α wave (11-13 Hz) and β wave (13 Hz-) from each segment.

In studies using EEG [43]–[45], the main focus is on the well-known α wave and θ wave, occurring within the 7-13 Hz and 4-7 Hz frequency bands, respectively, with α wave being the most prominent signal in EEG [43]. Alpha wave is seen as being complementary to θ wave; when the amplitude of one decreases with performance, the other is generally expected to increase and vice versa [44], [45]. Specifically, α band power increases in the absence of a task, and α band power decreases as more and more groups of neurons are activated to meet some task demand. Furthermore, studies using EEG in multimedia applications [21], [23], [33], [41] showed close correlations of α wave and θ wave with viewer attention and interest in multimedia content.

We did not use signals of δ wave and β wave captured from EEG since they are not of interest for the purpose of our study. The δ wave is normally present during deep sleep [46] and is obscured by excessive noise from muscle movements. Furthermore, β wave is widely considered to reflect motor activity [47]. We therefore conclude that using them is not appropriate for the purpose of our study.

From the above discussion, in our study, we used the power spectrum of θ wave (4-7 Hz), slow-α wave (7-9 Hz), mid-α wave (9-11 Hz), and fast-α wave (11-13 Hz). The measurements, which were obtained by using alphatec IV-s, are shown in Table 2. In this way, we obtain biological signal features $\boldsymbol{x}_n^E = \{x_\theta, x_{s\alpha}, x_{m\alpha}, x_{f\alpha}\}$ $(n = 1, 2, \cdots, N)$.

V. PROPOSED METHOD

In this section, we explain the method for estimating viewer interest while users are watching videos, using a framework for anomaly detection, based on collaborative use of the features of facial expression and the features of biological signals. We explain the anomaly detection for the features of facial expression and biological signals in **V-A** and **V-B**, respectively. In **V-C**, we describe in detail the estimation of viewer interest using the results of anomaly detection for the features of facial expression and biological signals. The term "anomaly" used in this article means that "the degree of viewer interest is particularly high". An overview of the proposed method is shown in Fig. 3.

A. ANOMALY DETECTION FOR FEATURES OF FACIAL EXPRESSION

In this subsection, we explain our method of anomaly detection for features of facial expression. In order to take into account the actual mechanisms of behavior of users, we use parametric techniques for anomaly detection. Generally, when using parametric techniques, estimating the effective dimension of features for correctly detecting an anomaly is needed. In our method, we apply Variational Bayesian Principal Component Analysis (VBPCA) [48] to the features of facial expression in order to estimate the effective dimension and adjust the statical degrees of freedom since not all of the 17 features are correlated to viewer interest [39], [49]. Unlike some related studies, we use VBPCA only for obtaining the probability distribution of the features of facial expression, for which the statical degrees of freedom are adjusted.

In general, VBPCA assumes that $\boldsymbol{x}^F = \{\boldsymbol{x}_1^F, \boldsymbol{x}_2^F, \cdots, \boldsymbol{x}_N^F\}$ are obtained by adding noise to the linear transformation of principal components as follows:

$$\boldsymbol{x}_n^F = \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu} + \boldsymbol{\epsilon}_n, \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{D\times d}$ represents the projection matrix of principal features of facial expression $\boldsymbol{z}_n \in \mathbb{R}^d$ $(n = 1, 2, \cdots, N)$ to $\boldsymbol{x}_n^F$ $(n = 1, 2, \cdots, N; N)$. Note that $\boldsymbol{\mu}$ denotes the data offset and $\boldsymbol{\epsilon}_n \in \mathbb{R}^D$ $(n = 1, 2, \cdots, N)$ denotes noise. From this assumption, the prior distribution and the conditional probability distribution of $\boldsymbol{x}^F$ given $z = \{z_1, z_2, \cdots, z_N\}$ are denoted as follows:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\boldsymbol{0}, \sigma^2\boldsymbol{I}_D), \tag{2}$$

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I}_d), \tag{3}$$

$$\boldsymbol{x}^F|\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{x}^F|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I}_D). \tag{4}$$

We calculate the optimal model parameters, $\hat{\boldsymbol{W}}$, $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}^2$, for Eqs. (2)-(4) by means of a variational approach based on [48]. From all of the above calculations, we can obtain the conditional probability distribution of $\boldsymbol{x}^F$

given z. Based on Bayes' theorem, the conditional probability distribution of z given $\boldsymbol{x}^F$ can be calculated from that of $\boldsymbol{x}^F$ given z. Therefore, the probability distribution of the features of facial expression, for which the statical degrees of freedom are adjusted, is calculated as follows:

$$z|\boldsymbol{x}^F \sim \mathcal{N}(\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{W}}^T(\hat{\sigma}^2\boldsymbol{I}_D)^{-1}(\boldsymbol{\mu}_{\boldsymbol{x}^F} - \hat{\boldsymbol{\mu}}), \hat{\boldsymbol{\Sigma}}), \tag{5}$$

where $\hat{\boldsymbol{\Sigma}} = (\boldsymbol{I}_d + \hat{\boldsymbol{W}}^T(\hat{\sigma}^2\boldsymbol{I}_D)^{-1}\hat{\boldsymbol{W}})^{-1}$, and $\boldsymbol{\mu}_{\boldsymbol{x}^F}$ is the sample mean of $\boldsymbol{x}^F$.

We define an anomaly score based on Hotelling's statistics. With a hypothesis that the features of facial expression follow a Gaussian distribution, the anomaly score of newly observed features of facial expression $\boldsymbol{x}^F_{new}$ can be considered as the Mahalanobis distance between $\boldsymbol{x}^F_{new}$ and $\boldsymbol{\mu}_{\boldsymbol{x}^F}$ in the latent space defined by Eq.(5).

$$a(\boldsymbol{x}^F_{new}) = (z_{new})^T\hat{\boldsymbol{\Sigma}}^{-1}(z_{new}), \tag{6}$$

where $z_{new} = \hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{W}}^T(\hat{\sigma}^2\boldsymbol{I}_D)^{-1}(\boldsymbol{x}^F_{new} - \boldsymbol{\mu}_{\boldsymbol{x}^F})$, and $a(\boldsymbol{x}^F_{new})$ denotes the anomaly score of $\boldsymbol{x}^F_{new}$. From Hotelling's statistics, the anomaly score, which is defined by Eq.(6), follows a chi-squared distribution with d degrees of freedom when $N \gg d$. Therefore, we calculate a threshold a_{th} for deciding whether the anomaly score is significantly high or not based on a chi-squared distribution as follows:

$$1 - \alpha = \int_0^{a_{th}} \chi^2(x|d)dx, \tag{7}$$

where α represents a parameter for calculating a confidence interval and $\chi^2(x|d)$ corresponds to the chi-squared distribution with d degrees of freedom.

In our method, we determine that $\boldsymbol{x}^F_{new}$ corresponds to high viewer interest when the anomaly score $a(\boldsymbol{x}^F_{new})$ is larger than a_{th}. In order to estimate viewer interest for a Web video, we estimate that the viewer thinks the video is interesting when the newly observed feature of facial expression $\boldsymbol{x}^F_{new}$, which is determined as corresponding to high viewer interest, continues for more than j^F ($j^F \geq 2$) samples. Finally, we obtain the result R^F that is binary.

B. ANOMALY DETECTION FOR FEATURES OF BIOLOGICAL SIGNALS

In this subsection, we explain our method of anomaly detection for features of biological signals. As we mentioned, the use of neurophysiological measures is relatively new in computer science, and we also define the anomaly score based on related studies [41], [50], [51] in which EEG was used. Those studies showed that the event-related power decrease (event-related desynchronization: ERD) of α wave and θ wave is a major cue for analyzing EEG data. An affective or motivational stimulus results in ERD of α wave and θ wave [50]. ERD indicates the degree to which neuron populations no longer oscillate in synchrony as they are activated to process a given task. In essence, ERD is defined as the percentage of band power decrease or increase between the resting period preceding a task and the task itself, as shown in the following equation:

$$\text{ERD} = \frac{(\text{bandpower})_{\text{rest}} - (\text{bandpower})_{\text{task}}}{(\text{bandpower})_{\text{rest}}}. \tag{8}$$

As an illustration, when more desynchronization is observed in $x^{f\alpha}$ from a user while the user is watching a video, it is indicative of higher viewer interest, which can be regarded as a psychological state evoked by multimedia [21], [22]. Motivated by the above discussion, we firstly define ERD for newly observed features of biological signals $\boldsymbol{x}^E_{new} = \{x^\theta_{new}, x^{s\alpha}_{new}, x^{m\alpha}_{new}, x^{f\alpha}_{new}\}$ in our study as follows:

$$\begin{aligned}\text{ERD}^i &= \frac{\mu^i - x^i_{new}}{\mu^i} \\ &= 1 - \frac{x^i_{new}}{\mu^i} \quad (i \in \{\theta, s\alpha, m\alpha, f\alpha\}),\end{aligned} \tag{9}$$

where $\boldsymbol{\mu}_{x^E} = \{\mu^\theta, \mu^{s\alpha}, \mu^{m\alpha}, \mu^{f\alpha}\}$ is the sample mean vector of $\boldsymbol{x}^E_n$. We use $\boldsymbol{\mu}_{x^E}$ as a reference owing to the difficulty of setting a reference state for each individual in a real situation.

As well as the anomaly detection for features of facial expression, we also use parametric techniques for anomaly detection to consider the actual mechanisms of behavior of users. In order to use parametric techniques for anomaly detection, we define the anomaly vector $\boldsymbol{a} = \{a^\theta, a^{s\alpha}, a^{m\alpha}, a^{f\alpha}\}$ and the probability distribution, which the anomaly vector follows, based on Eq.(9). However, deriving the probability distribution is difficult and complex since Eq.(9) has a fractional expression with substraction. To deal with this difficulty, we focus on the ratio of each element of $\boldsymbol{\mu}_{x^E}$ to that of $\boldsymbol{x}^E_{new}$, and we define the anomaly vector as follows:

$$a^i = \begin{cases} \dfrac{x^i_{new}}{\mu^i} & \text{if } i = \theta \\ \dfrac{\mu^i}{x^i_{new}} & \text{otherwise.} \end{cases} \tag{10}$$

In this way, we define the anomaly vector with consideration of the complementary nature between α wave and θ wave mentioned in section **IV-B**.

To use parametric techniques for anomaly detection, we then derive the probability distribution that the defined anomaly vector follows, based on blind source separation (BSS) and related studies [52]–[54]. In BSS, signals are considered as mixtures of several sources with the assumption that the signals are transmitted without any delay. This is also true for EEG signals since electric field change is transmitted instantly to the scalp [53], [54]. Using the hypothesis that the sources of signals are not moveable and no reactances exist in the conductive route between the sources and the scalp, the transfer function h_k between the k-th source ($k = 1, 2 \cdots, K$; K being the number of sources and it being the same as the number of neurons in the brain) and the scalp

is the same as that in the whole brain as follows:

$$X(t) = \sum_{k=1}^{K} h_k s_k(t), \tag{11}$$

where $X(t)$ is the observed EEG signals and $s_k(t)$ is the signal transmitted from the k-th source. In order to derive the probability distribution that the defined anomaly vector follows, we build up three hypotheses related to the complex amplitude $s_k(n,f)(k = 1, 2 \cdots, K)$ of $s_k(t)$ in timeslot (n,f), where f is the index for a frequency, as follows:

- Stationarity in arbitrary EEG frame n
- $E[s_{k_1}(n,f)s_{k_2}(n,f)^*] = 0$ for $k_1 \neq k_2$
- Following a Gaussian distribution

From these hypotheses, EEG signals can be considered as following a Gaussian distribution. As a result, the power spectrum follows a chi-square distribution. Therefore, the defined anomaly vector follows an F-distribution.

We calculate a threshold a_{th} for deciding whether the anomaly score is significantly high or not based on the chi-squared distribution as follows:

$$1 - \alpha = \int_0^{a_{th}} F(x|l,m)dx, \tag{12}$$

where α represents a parameter for calculating a confidence interval and $F(x|l,m)$ corresponds to the F-distribution with parameters l and m related to degree of freedom.

In our method, we determine that $\boldsymbol{x}^E_{new}$ corresponds to high viewer interest when each element of the anomaly vector $\boldsymbol{a}$ is larger than a_{th}. In each element of the newly observed features of biological signals $\boldsymbol{x}^E_{new}$, we estimate viewer interest for a Web video. Specifically, we estimate that the viewer thinks the video is interesting when each element of the newly observed features of biological signals $\boldsymbol{x}^E_{new}$, which is determined as corresponding to high viewer interest, continues for more than j^E ($j^E \geq 2$) samples. Finally, we obtain the results $\boldsymbol{R}^E = \{R^\theta, R^{s\alpha}, R^{m\alpha}, R^{f\alpha}\}$, where each element of $\boldsymbol{R}^E$ is binary.

C. ESTIMATING VIEWER INTEREST

In this subsection, we describe in detail the estimation of viewer interest using the results of anomaly detection for the features of facial expression and biological signals. We integrate the results R^F and $\boldsymbol{R}^E$ from **V-A** and **V-B**, respectively. There are many state-of-the-art methods to integrate the results [55]–[58]. However, they need a large amount of training data to do that. This trait is inappropriate for the purpose of our study.

In this paper, we focus on simple voting methods for pattern recognition based on [59]. Firstly, we integrate each result of $\boldsymbol{R}^E$ into $\hat{R}^E$ by utilizing the ''Sum rule'' reported in [59]. In detail, all anomaly scores of the results that are estimated as high viewer interest are added. In addition, all anomaly scores of the other results are added. We estimate $\hat{R}^E$ as the class of the highest sum.

We then estimate the final viewer interest. In order to detect the state of ''particularly interesting'', we estimate viewer interest as high if either R^F or $\hat{R}^E$ represents high viewer interest.

VI. EXPERIMENTAL RESULTS

In this section, we show experimental results to verify the effectiveness of our method. Firstly, we explain the experimental conditions in **VI-A**. Secondly, in order to fully investigate the generalizability of the proposed method, we present the results of both within-subjects and between-subjects studies. For the within-subjects estimation in **VI-B**, we construct a user-dependent model for each subject and use 50-fold cross-validation for estimation, where 50 is the number of trailers used in this experiment. We determine each subject's thresholds for all methods based on viewer interest estimation performance. The final performance is the average across all folds. On the other hand, in **VI-C**, the between-subjects study uses leave-one-subject-out cross-validation, testing on each subject in turn.

A. EXPERIMENTAL CONDITIONS

In this subsection, we explain the experimental conditions. First of all, the keyword ''movie trailer'' was given as a query to YouTube. We then obtained 50 trailers of videos based on [21]–[23], and the 50 trailers consisted of five genres of film: science fiction, comedy, action, horror and romance. The subjects were 11 men aged 22 to 24 years, and they watched all of the 50 trailers. All of the subjects were volunteers, and instructions were given by an experimenter orally. The length of each trailer was about 90 seconds, and each genre contained 10 trailers. Each subject watched the trailers while sitting on a chair about 0.5 meters away from the display. The trailers were displayed on a monitor of 1920×1080 in resolution in full screen mode. Kinect was set behind the display to obtain features of the subject's facial expression. In addition, each subject wore alphatec IV-s on his head to obtain biological signals. In order to deal with artifacts (e.g., spike noise and muscle artifacts), we conducted the experiment in a laboratory room that is used only for obtaining EEG signals. We took care about the influence of other electronic devices, and we monitored the noise level of the EEG signals by using software for alphatec IV-s.

The subjects then evaluated each of the trailers after watching in four grades.[5] Facial expression features and biological signal features were calculated every second. As a result, datasets including four elements (trailer, rating, features of facial expression, and features of biological signal) could be obtained.

In this experiment, we estimated viewer interest with respect to two classes, i.e., ''High Interest'' and ''Neutral''. The class ''High Interest'' corresponds to the video evaluation rated 4 by subjects, and the class ''Neutral'' corresponds to the video evaluation rated 1, 2 or 3 by the subjects. Details about the prepared ratings are shown in Table 3.

[5] 1=Not at all interesting, 4=Extremely interesting

TABLE 3. Details of the datasets used in the experiment per subject. "HI" represents the number of videos classified into "High Interest", and "NT" represents the number of videos classified into "Neutral".

Subject	A	B	C	D	E	F	G	H	I	J	K
HI	14	16	3	15	16	11	4	7	9	9	15
NT	36	34	47	35	34	29	46	43	41	41	35

TABLE 4. Details of the comparative methods. "ADFace" represents anomaly detection of features of facial expression and "ADBio" represents anomaly detection of features of biological signals. "Est" corresponds to estimating viewer interest.

Name	ADFace	ADFBio	Est	Parameter
Ours	**V-A**	**V-B**	**V-C**	a_{th}
C1	**V-A**	-	-	
C2	-	**V-B**	-	
C3	Clustering-based [60]		**V-C**	Num. of clusters
C4	LOF [61]			Num. of nearest neighbors
C5	COF [62]			
C6	Histogram-based [63]			Num. of bins
C7	Information theoretic-based [64]			Expected num. of outliers

We compared the results of estimation by our method with results of seven comparative methods shown in Table 4. C1 and C2 were used to verify the effectiveness of collaborative use of facial expression and biological signals for estimating viewer interest. C3-C7 are simple unsupervised techniques for anomaly detection, which are suitable for the purpose of our study. Specifically, a clustering-based algorithm is applied to the dataset for anomaly detection in C3 [60]. C4 and C5 are known as Local Outlier Factor (LOF) [61] and Connectivity-based Outlier Factor (COF) [62], respectively. Both of them use k-nearest neighbors to find the radius of the smallest hyper-sphere centered at data instance. C6 uses histograms for anomaly detection based on [63], and it is the simplest non-parametric approach for anomaly detection. Information theoretic is applied to the dataset for anomaly detection in C7, assuming that anomalies in the dataset induce irregularities in the information content of the dataset. The aim of using C3-C7 is to show the effectiveness of using a parametric approach for anomaly detection, as the first method towards realizing feasible estimation of viewer interest.

We evaluated the performance by utilizing the F-measure, correctly classified rate (CCR) and area under the receiver operating characteristic curve (AUC) in a plot of Anomalous Video Accuracy (AVA) against False Alarm Rate (FAR), while changing the parameter a_{th} mentioned in **V-A** and **V-B**. FAR and AVA are defined as follows:

$$\text{FAR} = 1 - \frac{\text{Num. of correctly estimated NT}}{\text{Num. of NT}},$$
$$\text{AVA} = \frac{\text{Num. of correctly estimated HI}}{\text{Num. of HI}}.$$

''HI'' represents the number of videos classified into ''High Interest'', and ''NT'' represents the number of videos classified into ''Neutral''. The parameters of each method that are used for calculating the receiver operating characteristic curve are also shown in Table 4.

B. EXPERIMENTAL RESULTS FOR WITHIN-SUBJECTS

In this subsection, we show the experimental results for within-subjects. Table 5 and Fig. 4 summarize the performances of our method and comparative methods C1-C7 for user-dependent models. The results show that our method outperforms the comparative methods. The AUC of our method is much higher than those of C1 and C2, indicating the effectiveness of collaborative use of facial expression and biological signals based on our method. Furthermore, in most cases (10 out of 11 subjects) in Table 5, our method also succeeds in outperforming comparative methods C3-C7, which are well-known techniques for anomaly detection. As expected, this is because C3-C7 are purely data-driven. The results of analysis are shown in detail below.

First of all, C3 is effective only when the anomalies do not form significant clusters among themselves. However, the features that we used have particular patterns when viewer interest is high. Secondly, C4 and C5 usually require sufficient close neighbors. Therefore, if the dataset has anomalies that have sufficient close neighbors, the performances of C4 and C5 become low. In the experiment, some anomaly instances of the features of facial expression have enough close neigbors, since some subjects did not often change their facial expression while watching videos. This trait is not suitable for the features of facial expression, whereas our method detects anomaly instances of the features of facial expression, which have sufficient close neigbors, based on statistical models. In addition, the performance of C6 is obviously low since construction of the histogram is influenced by the small amount of training data, which is a general problem caused by the actual mechanism of users. Finally, the performance of our method is higher than that of C7. This is because our method can estimate viewer interest even if the number of particularly interesting videos is small. On the other hand, information theoretic measures can generally detect the presence of anomalies only when there are significantly large numbers of anomalies present in the data.

C. EXPERIMENTAL RESULTS FOR BETWEEN-SUBJECTS

In this subsection, we show the experimental results for between-subjects to fully investigate the generalizability of the proposed method.

Table 6 summarizes the F-measures of the user-independent and user-dependent models for estimation of viewer interest. It is very encouraging that the user-independent model (average F-measure $=$ 0.760) outperforms the user-dependent model (average F-measure $=$ 0.747) in our method, given sufficient features of facial expression and biological signals. Furthermore, F-measures of the comparative methods in the user-independent model also outperform those in the user-dependent model since they are purely data-driven as mentioned in the previous subsection. However, the F-measure of our method is significantly higher than those of the comparative methods, given sufficient features of facial expression and biological signals with $p < 0.01$ by Welch's t-test.

TABLE 5. Correctly classified rate and F-measure of estimating viewer interest.

	Correctly Classified Rate								F-measure							
Subject	Ours	C1	C2	C3	C4	C5	C6	C7	Ours	C1	C2	C3	C4	C5	C6	C7
A	**0.600**	0.580	0.540	0.560	0.580	0.540	0.360	0.540	**0.750**	0.600	0.633	0.600	0.671	0.612	0.423	0.405
B	**0.560**	0.460	**0.560**	0.460	0.540	0.540	0.420	0.500	**0.653**	0.579	0.636	0.608	0.588	0.588	0.457	0.452
C	**0.720**	0.560	0.560	0.560	0.540	0.440	0.380	0.360	0.776	0.494	**0.779**	0.491	0.491	0.435	0.456	0.445
D	**0.580**	0.560	0.540	0.660	0.580	0.520	0.480	0.440	**0.755**	0.392	0.711	0.646	0.677	0.613	0.356	0.228
E	**0.540**	0.580	0.560	0.520	0.640	0.640	0.340	0.360	**0.706**	0.431	0.586	0.595	0.571	0.595	0.337	0.367
F	**0.760**	0.580	0.740	0.720	0.600	0.540	0.320	0.460	**0.784**	0.506	0.571	0.504	0.651	0.653	0.367	0.475
G	**0.520**	0.440	0.480	0.440	**0.520**	0.480	0.460	0.420	**0.696**	0.652	0.681	0.645	0.651	0.645	0.458	0.426
H	**0.720**	0.620	0.700	0.640	0.460	0.540	0.480	0.500	**0.873**	0.428	0.769	0.576	0.422	0.694	0.497	0.424
I	0.720	**0.760**	0.520	**0.760**	0.640	0.620	0.420	0.380	0.640	**0.673**	0.640	0.669	0.584	0.582	0.476	0.437
J	**0.700**	0.600	0.660	0.620	0.580	0.460	0.380	0.420	**0.789**	0.581	0.721	0.747	0.673	0.569	0.466	0.334
K	**0.680**	**0.680**	0.660	0.560	0.580	0.500	0.400	0.520	**0.793**	0.453	0.755	0.706	0.677	0.662	0.345	0.467
Average	**0.645**	0.584	0.593	0.591	0.569	0.529	0.404	0.445	**0.747**	0.496	0.666	0.617	0.605	0.604	0.422	0.405

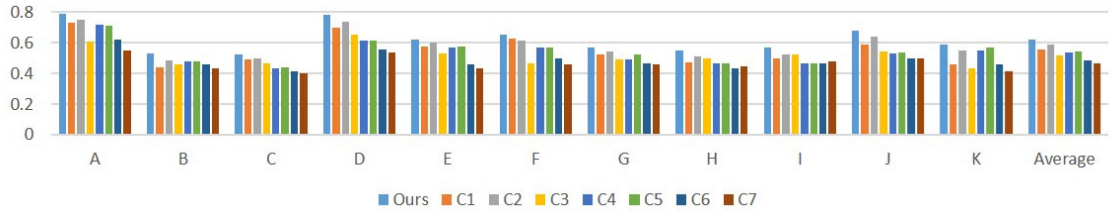

FIGURE 4. AUC of estimating viewer interest for all subjects.

TABLE 6. F-measure of estimating viewer interest by a user-dependent model (UDM) and a user-independent model (UIM).

	Ours		C1		C2		C3		C4		C5		C6		C7	
Subject	UDM	UIM	UDM	UIM	UDM	UIM	UDM	UIM	UDM	UIM	UDM	UIM	UDM	UIM	UDM	UIM
A	0.750	**0.773**	0.600	**0.621**	0.633	**0.643**	0.600	**0.633**	**0.671**	0.653	0.612	**0.635**	0.423	**0.525**	0.405	**0.446**
B	0.653	**0.686**	0.579	**0.587**	0.636	0.636	0.608	**0.612**	0.588	**0.607**	0.588	**0.654**	0.457	**0.488**	0.452	**0.543**
C	**0.776**	0.764	0.494	**0.524**	0.621	**0.635**	0.491	**0.533**	**0.491**	0.483	**0.435**	0.431	0.456	**0.532**	**0.445**	0.412
D	0.755	**0.786**	0.392	**0.431**	0.711	**0.738**	0.646	**0.657**	**0.677**	0.667	0.613	**0.648**	0.356	**0.378**	0.228	**0.332**
E	0.706	**0.732**	0.431	0.431	0.586	**0.631**	**0.595**	0.588	0.571	**0.647**	**0.595**	0.566	0.337	**0.383**	0.367	**0.383**
F	0.784	**0.762**	0.506	**0.557**	0.571	**0.611**	0.504	**0.558**	0.651	**0.663**	0.653	**0.677**	0.367	**0.426**	0.475	**0.544**
G	0.696	**0.727**	0.652	**0.681**	0.681	**0.721**	0.645	**0.678**	0.651	**0.652**	**0.645**	0.632	0.458	**0.532**	0.426	**0.456**
H	0.873	**0.876**	0.428	**0.472**	0.769	**0.785**	0.576	**0.587**	0.422	**0.523**	0.694	**0.722**	0.497	**0.544**	0.424	**0.433**
I	0.640	**0.668**	0.345	**0.345**	0.640	**0.647**	0.669	**0.702**	0.584	**0.592**	0.582	**0.583**	0.476	**0.511**	0.437	**0.439**
J	0.789	**0.795**	0.581	**0.596**	0.721	**0.735**	0.747	**0.755**	0.673	**0.682**	0.569	**0.635**	0.466	**0.468**	0.334	**0.355**
K	0.793	**0.795**	0.453	**0.458**	0.755	**0.768**	0.706	**0.709**	0.677	**0.683**	**0.662**	0.632	0.345	**0.439**	**0.467**	0.441
Average	0.747	**0.760**	0.496	**0.518**	0.666	**0.686**	0.617	**0.637**	0.605	**0.623**	0.604	**0.620**	0.422	**0.475**	0.405	**0.436**

To verify the robustness to the change in the amount of training data, we then estimated viewer interest when the amount of training data is small under the user-independent condition. Firstly, in each subject, we divided 50 groups (trailers) of facial expression and biological signals into 30 groups and 20 groups. We used the latter 20 groups for each subject as training data. Thus, we evaluated the robustness of our method when the number of groups for each subject as training data was changed from 10 to 20. For example, if we estimate the viewer interest for subject A, we use the groups for other subjects as training data, and the range of changes is from 10 groups $\times 10$ subjects to 20 groups $\times 10$ subjects.

As shown in Fig. 5, our method has robust performance despite changes in the amount of training data. Moreover, when the number of trailers used as training becomes 100, the performance of our method is obviously high. In addition to robustness, our method shows better performance as the number of trailers used as training decreases to that in the user-dependent mode (see Table 5 and Fig. 4). Therefore, our method realizes effective estimation of viewer interest even if the amount of training data is changed. This trait is desirable for creating a method towards realizing feasible estimation of viewer interest.

The size of the dataset might be one of limitations of our study for investigating the generalizability of the proposed method. It is difficult to know the exact size of the dataset needed to generalize the models since this is the first study on anomaly detection for viewer interest estimation using facial expression and biological signals. However, in our method, the sample means of the features of facial expression and biological signals are only keys to determine the estimation performance. Therefore, we can easily generalize the models

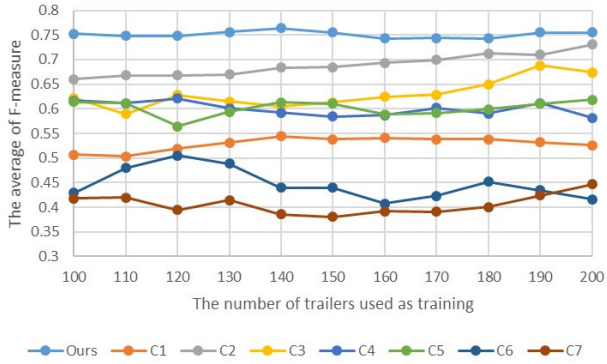

FIGURE 5. Average of F-measures for all subjects.

even if the number of samples seems to be small. Moreover, Fig. 5 shows that the performance of our method is stable when using 100 trailers or more. For all of these reasons, we believe that the results show the generalizability of our method.

VII. CONCLUSION

In this paper, we have proposed a novel method for estimating viewer interest while users are watching Web videos, via collaborative use of facial expression and biological signals. In the proposed method, we use a framework for anomaly detection to take into account two actual mechanisms of users while they are watching Web videos. By comparing the performance of our method with the performances of well-known methods, we showed the validity of using parametric techniques for anomaly detection. We also compared the results of a user-dependent model with those of a user-independent model. The results showed that the performance of our method is better if more training data are available. Consequently, we have realized a successful method for estimating viewer interest based on a framework for anomaly detection via collaborative use of facial expression and biological signals. We believe that our study is the first step towards realizing a feasible method for estimation of viewer interest.

REFERENCES

[1] X. Cheng, J. Liu, and C. Dale, "Understanding the characteristics of Internet short video sharing: A YouTube-based measurement study," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1184–1194, Aug. 2013.

[2] D. K. Krishnappa, M. Zink, C. Griwodz, and P. Halvorsen, "Cache-centric video recommendation: An approach to improve the efficiency of YouTube caches," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 4, 2015, Art. no. 48.

[3] Q. Huang, B. Chen, J. Wang, and T. Mei, "Personalized video recommendation through graph propagation," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 10, no. 4, 2014, Art. no. 32.

[4] P. Cui, Z. Wang, and Z. Su, "What videos are similar with you?: Learning a common attributed representation for video recommendation," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 597–606.

[5] J. Tarvainen, M. Sjöberg, S. Westman, J. Laaksonen, and P. Oittinen, "Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2085–2098, Dec. 2014.

[6] 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.

[7] X. Zhao, H. Luan, J. Cai, J. Yuan, X. Chen, and Z. Li, "Personalized video recommendation based on viewing history with the study on YouTube," in *Proc. 4th Int. Conf. Internet Multimedia Computing Service*, 2012, pp. 161–165.

[8] Z. Dai *et al.*, "A real-time video recommendation system for live programs," in *Proc. 4th IEEE Int. Conf. Netw. Infrast. Digit. Content (IC-NIDC)*, Sep. 2015, pp. 498–502.

[9] Y. Hou *et al.*, "Predicting movie trailer viewer's 'like/dislike' via learned shot editing patterns," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 29–44, Jan./Mar. 2016.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] J. Fu, D. Miao, W. Yu, S. Wang, Y. Lu, and S. Li, "Kinect-like depth data compression," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1340–1352, Oct. 2013.

[15] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using Kinect sensor," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1110–1120, Aug. 2013.

[16] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2094–2107, Nov. 2015.

[17] 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.

[18] 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)*, Nov. 2014, pp. 55–58.

[19] 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 Human Int. Conf.*, 2014, Art. no. 15.

[20] 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.

[21] 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.

[22] S. Liu *et al.*, "What makes a good movie trailer?: Interpretation from simultaneous EEG and eyetracker recording," in *Proc. ACM Multimedia Conf.*, 2016, pp. 82–86.

[23] Y. Sasaka, T. Ogawa, and M. Haseyama, "Multimodal interest level estimation via variational Bayesian mixture of robust CCA," in *Proc. ACM Multimedia Conf.*, 2016, pp. 387–391.

[24] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida, "On word-of-mouth based discovery of the Web," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf.*, 2011, pp. 381–396.

[25] 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.

[26] A. Brodersen, S. Scellato, and M. Wattenhofer, "YouTube around the world: Geographic popularity of videos," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 241–250.

[27] M. Soleymani and M. Pantic, "Human-centered implicit tagging: Overview and perspectives," in *Proc. IEEE Int. Conf. Syst., Man, (SMC)*, Oct. 2012, pp. 3304–3309.

[28] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, p. 15, 2009.

[29] F. J. Anscombe and I. Guttman, "Rejection of outliers," *Technometrics*, vol. 2, no. 2, pp. 123–147, 1960.

[30] D. E. Berlyne, *Conflict, Arousal, and Curiosity*. New York, NY, USA: 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] A. J. Karran and U. Kreplin, "The drive to explore: Physiological computing in a cultural heritage context," in *Advances in Physiological Computing*. Springer, 2014, pp. 169–195.
[34] S. Hidi and K. A. Renninger, "The four-phase model of interest development," *Edu. Psychol.*, vol. 41, no. 2, pp. 111–127, 2006.
[35] 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.
[36] 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.
[37] J. M. Henderson, "Human gaze control during real-world scene perception," *Trends Cognit. Sci.*, vol. 7, no. 11, pp. 498–504, 2003.
[38] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proc. IEEE CVPR*, vol. 1. Jun. 2005, pp. 631–637.
[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*. Amsterdam, The Netherlands: Elsevier, 2005.
[41] 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.
[42] 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.
[43] W. Klimesch, "Memory processes, brain oscillations and EEG synchronization," *Int. J. Psychophysiol.*, vol. 24, nos. 1–2, pp. 61–100, 1996.
[44] W. Klimesch, P. Sauseng, and S. Hanslmayr, "EEG alpha oscillations: The inhibition–timing hypothesis," *Brain Res. Rev.*, vol. 53, no. 1, pp. 63–88, 2007.
[45] T. A. Rihs, C. M. Michel, and G. Thut, "Mechanisms of selective inhibition in visual spatial attention are indexed by α-band EEG synchronization," *Eur. J. Neurosci.*, vol. 25, no. 2, pp. 603–610, 2007.
[46] J. A. Hobson and E. F. Pace-Schott, "The cognitive neuroscience of sleep: Neuronal systems, consciousness and learning," *Nature Rev. Neurosci.*, vol. 3, no. 9, pp. 679–693, 2002.
[47] S. N. Baker, "Oscillatory interactions between sensorimotor cortex and the periphery," *Current Opinion Neurobiol.*, vol. 17, no. 6, pp. 649–655, 2007.
[48] C. M. Bishop, "Variational principal components," in *Proc. 9th Int. Conf. Artif. Neural Netw. (ICANN)*, vol. 1. 1999, pp. 509–514.
[49] C.-H. Wu, W.-L. Wei, J.-C. Lin, and W.-Y. Lee, "Speaking effect removal on emotion recognition from facial expressions based on eigenface conversion," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1732–1744, Dec. 2013.
[50] K. Onoda *et al.*, "Anticipation of affective images and event-related desynchronization (ERD) of alpha activity: An MEG study," *Brain Res.*, vol. 1151, pp. 134–141, Jun. 2007.
[51] M. C. M. Bastiaansen, K. B. E. Böcker, and C. H. M. Brunia, "ERD as an index of anticipatory attention? Effects of stimulus degradation," *Psychophysiology*, vol. 39, no. 1, pp. 16–28, 2002.
[52] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.
[53] S. Romero, M. A. Mañanas, and M. J. Barbanoj, "A comparative study of automatic techniques for ocular artifact reduction in spontaneous EEG signals based on clinical target variables: A simulation case," *Comput. Biol. Med.*, vol. 38, no. 3, pp. 348–360, 2008.
[54] R. Vigario, J. Sarela, V. Jousmiki, M. Hamalainen, and E. Oja, "Independent component approach to the analysis of EEG and meg recordings," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 5, pp. 589–593, May 2000.
[55] V. C. Raykar *et al.*, "Supervised learning from multiple experts: Whom to trust when everyone lies a bit," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 889–896.
[56] V. C. Raykar *et al.*, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, Apr. 2010.
[57] V. C. Raykar and S. Yu, "Eliminating spammers and ranking annotators for crowdsourced labeling tasks," *J. Mach. Learn. Res.*, vol. 13, pp. 491–518, Feb. 2012.
[58] J. Goldberger, "Combining soft decisions of several unreliable experts," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2334–2338.
[59] M. van Erp, L. Vuurpijl, and L. Schomaker, "An overview and comparison of voting methods for pattern recognition," in *Proc. 8th Int. Workshop Frontiers Handwriting Recognit.*, Aug. 2002, pp. 195–200.
[60] M.-F. Jiang, S.-S. Tseng, and C.-M. Su, "Two-phase clustering process for outliers detection," *Pattern Recognit. Lett.*, vol. 22, nos. 6–7, pp. 691–700, 2001.
[61] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. Int. Conf. Manage. Data (SIGMOD)*, 2000, pp. 93–104.
[62] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, 2002, pp. 535–548.
[63] M. Goldstein and A. Dengel, "Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm," in *KI, Poster Demo Track*. Citeseer, 2012, pp. 59–63.
[64] W. Lee and D. Xiang, "Information-theoretic measures for anomaly detection," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2001, pp. 130–143.

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.

TAKAHIRO OGAWA (S'03–M'08) received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Japan, in 2003, 2005, and 2007, respectively. He joined the Graduate School of Information Science and Technology, Hokkaido University, in 2008, where he is currently an Associate Professor. His research interests are multimedia signal processing and its applications. He is a member of the ACM, the EURASIP, the IEICE, and the ITE. He has been an Associate Editor of the *ITE Transactions on Media Technology and Applications*.

MIKI HASEYAMA (S'88–M'91–SM'06) received the B.S., M.S., and Ph.D. degrees in electronics from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University, in 1994, as an Associate Professor. She was a Visiting Associate Professor with Washington University in St. Louis, 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, Institute of Image Information and Television Engineers, and Information Processing Society of Japan. She has been a Vice President of the Institute of Image Information and Television Engineers, Japan, an Editor-in-Chief of the *ITE Transactions on Media Technology and Applications*, a Director, International Coordination and Publicity of The Institute of Electronics, Information and Communication Engineers.