IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 24, 2020, accepted October 30, 2020, date of publication November 9, 2020,
date of current version November 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3036908

Human-Centric Emotion Estimation Based on Correlation Maximization Considering Changes With Time in Visual Attention and Brain Activity

YUYA MOROTO[1], (Graduate Student Member, IEEE), KEISUKE MAEDA[2], (Member, IEEE), TAKAHIRO OGAWA[3], (Senior Member, IEEE), AND MIKI HASEYAMA[3], (Senior Member, IEEE)

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

Corresponding author: Yuya Moroto (moroto@lmd.ist.hokudai.ac.jp)

This work was supported in part by the MIC/SCOPE under Grant #181601001.

ABSTRACT A human-centric emotion estimation method based on correlation maximization with consideration of changes with time in visual attention and brain activity when viewing images is proposed in this paper. Owing to the recent developments of many kinds of biological sensors, many researchers have focused on multimodal emotion estimation using both eye gaze data and brain activity data for improving the quality of emotion estimation. In this paper, a novel method that focuses on the following two points is introduced. First, in order to reduce the burden on users, we obtain brain activity data from users only in the training phase by using a projection matrix calculated by canonical correlation analysis (CCA) between gaze-based visual features and brain activity-based features. Next, for considering the changes with time in both visual attention and brain activity, we obtain novel features based on CCA-based projection in each time unit. In order to include these two points, the proposed method analyzes a fourth-order gaze and image tensor for which modes are pixel location, color channel and the changes with time in visual attention. Moreover, in each time unit, the proposed method performs CCA between gaze-based visual features and brain activity-based features to realize human-centric emotion estimation with a high level of accuracy. Experimental results show that accurate human emotion estimation is achieved by using our new human-centric image representation.

INDEX TERMS Multimodal approach, CCA, changes with time, tensor analysis, eye gaze data, fNIRS.

I. INTRODUCTION

In the field of affective computing, it has been reported that since the construction of a genuine intelligent system such as a system for multimedia recommendation or retrieval corresponding to the user's semantics is difficult without considering the emotional mechanism, it is necessary to analyze human emotions [1]. It should be noted that emotions can be defined as the semantics describing the type and intensity of ''affections'', ''sensibility'', ''feelings'' or ''moods'' evoked by humans [2]. Actually, much research has been conducted for the realization of image retrieval systems based on these human emotions (Emotional Semantic Image Retrieval; ESIR) [2]. However, the realization of ESIR based only on content features such as visual features and text features is a difficult task. Owing to the recent development of many kinds of biological sensors [3]–[7], many researchers have studied human emotion estimation based on human bio-signals for making computers recognize human emotions [8]–[10]. In those studies, human emotions were estimated by applying machine learning techniques to bio-signals obtained from humans. Therefore, in this study, we focuse on the estimation of emotions based on bio-signals evoked by viewing images.

Researchers who have majored in psychology or neuroscience have claimed that human emotions are affected by human gaze focusing on objects [11], [12], and eye gaze data can be easily obtained due to recent sensor developments [7]. Thus, emotion estimation based on eye gaze data has attracted much attention [13], [14]. Specifically, it has been reported that the stimulus obtained from the first object gazed at is

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

most closely analogous to the emotion [15]. However, since it is difficult to perfectly represent inner information such as feelings and opinions from only eye gaze data, there is a limitation to the improvement in estimation accuracy.

On the other hand, many researchers have studied estimation of human emotions or preferences by using brain activities obtained from electroencephalograms (EEG), functional magnetic resonance imaging (fMRI) and functional near-infrared spectroscopy (fNIRS) [8], [9], [16]. Since explicit and inner information, i.e., gaze and brain activity information, becomes available, many researchers have focused on multimodal emotion estimation using both of these data for improving the estimation quality [17]–[20]. Furthermore, the miniaturization of devices has enabled the two kinds of bio-signals to be easily obtained. Then methods using these two kinds of bio-signals enable estimation of human emotions with higher performance than that of emotion estimation methods using only one bio-signal.

In the use of bio-signals, we should consider the burden on users in acquiring those signals. Eye gaze data can be obtained by using small sensors such as sensors in glasses, but obtaining brain activity data still imposes a heavy burden on users. Moreover, in the use of eye gaze data, it has been reported that the changes in visual attention with time and gazed objects are related to human emotion [15]. Although some works, which estimate human emotions based on eye gaze and brain activity data, were conducted [17]–[20], these works did not focus on the changes in bio-signals with time. Therefore, in order to realize emotion estimation with higher performance, we also need to collaboratively use changes in both eye gaze data and brain activity data with time.

According to the above discussion, we tackle the following two problems in this paper.

1) In order to reduce the burden on users, we obtain brain activity data from users only in the training phase. In other words, a method that realizes the collaborative use of eye gaze and brain activity data but does not need the acquisition of brain activity data in the test phase is desirable.
2) We should consider the changes with time in eye gaze data and brain activity data for improvement in emotion estimation accuracy. Thus, brain activity data that have higher temporal resolution are suitable for considering the changes with time. Moreover, by analyzing the relationship between eye gaze and brain activity data at each time, improvement in emotion estimation accuracy is expected.

A human-centric emotion estimation method based on correlation maximization between visual attention and brain activity with consideration of the changes with time is proposed in this paper. Then since the proposed method can be trained for each user and focuses on the extraction of each user's implicit state, we use "human-centric". The proposed method solves the above problems by using canonical correlation analysis (CCA) [21] and considering the changes with time. Concretely, these problems can be solved as follows.

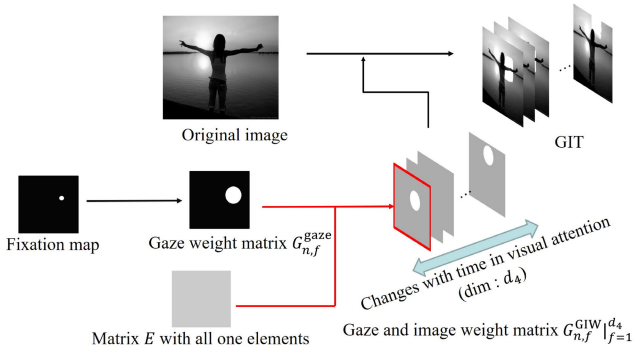


FIGURE 1. Construction of the GIT. We construct the fourth-order GIT for analyzing images with the changes over time in the visual attention, simultaneously. In this figure, a gray-scale image is used for simplifying the explanation.

1) A projection matrix is calculated by CCA from two kinds of features. Once this projection matrix has been calculated, we do not have to perform its recalculation. Thus, since we need to prepare brain activity data for only the training phase, the user burden is expected to be reduced.
2) We construct a gaze and image tensor (GIT) [22] as shown in Fig. 1 for considering the changes of visual attention with time. The axis of changes with time is added to images and corresponds to frames. This is a novel image representation approach considering the changes of visual attention [22]. Then we can perform fusion of gaze-based visual features and brain activity-based features.

The effectiveness of CCA has been reported in several fields including computer vision and human-computer interaction [23]–[26]. We therefore integrate gaze-based visual features and brain activity-based features by using the CCA-based approach. First, for analyzing the images and the changes with time in visual attention, simultaneously, we use the fourth-order GIT [22]. The first and second modes of this tensor are pixel locations, and the third mode corresponds to the color channels. These modes correspond to the information of images. In addition to these modes, we use the fourth mode of this tensor for considering the changes with time, that is, frames. Furthermore, by inputting the obtained GIT to convolutional neural network (CNN) [27] models, the proposed method enables the derivation of new gaze-based visual features. Next, the proposed method projects these gaze-based visual features in order to obtain emotion-correlated features by maximizing canonical correlation based on CCA using brain activity-based features obtained from users viewing images. By using these feature extraction and projection approaches, the proposed method can derive human-centric features suitable for the emotion estimation. As another advantage, since the brain activity-based features are used only for deriving the feature projection, their acquisition is not necessary for estimating emotion from a newly obtained image, i.e., our method has high applicability. Finally, in the classification

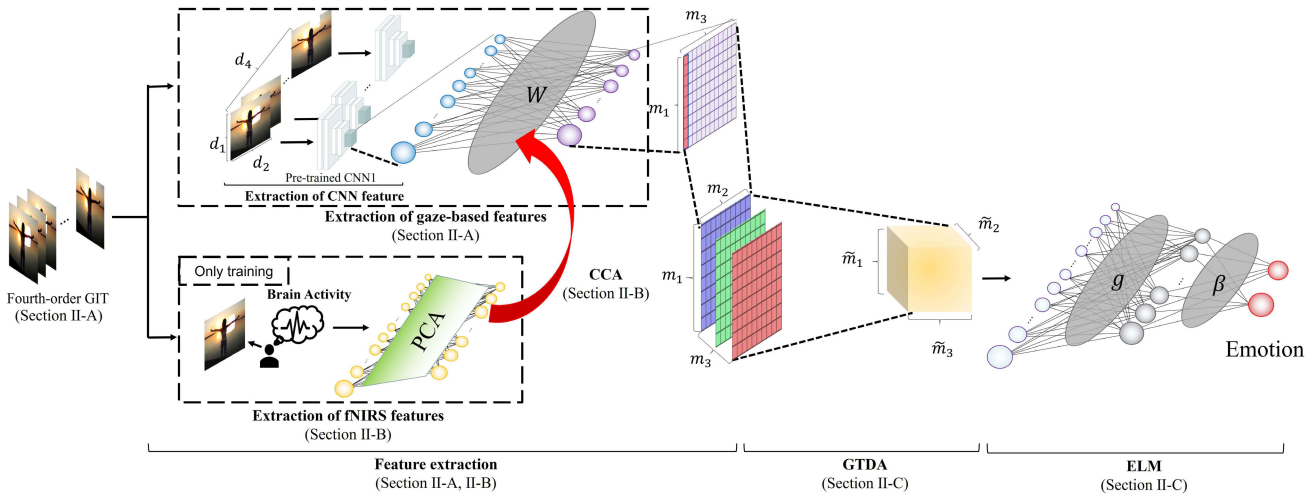


FIGURE 2. Overview of our method. First, we calculate the fourth-order GIT and extract gaze-based visual features. Next, we extract brain activity-based features and apply CCA to these features. By aligning the projected features, we generate the third-order GIT considering brain activity. Finally, we estimate emotions via tensor analysis (GTDA) and ELM.

step, the proposed method derives human-centric visual features from multiple CNN models and obtains a third-order tensor for which modes correspond to "the dimensions of the projected features", "the kinds of adopted CNN models" and "the time axis". Then, by applying generalized tensor discriminant analysis (GTDA) [28] to this third-order tensor and performing classification based on an extreme learning machine (ELM) [29] that enables training from a small number of training samples, the proposed method realizes the human-centric emotion estimation.

II. OUR ESTIMATION METHOD

In this section, we explain our emotion estimation method based on CCA between gaze-based visual features and brain activity-based features with consideration of the changes with time. The proposed method consists of three steps as shown in Fig. 2. In the first step (Sec. II-A), we calculate the gaze-based visual features. The fourth-order GIT, which is image representation for considering objects in images and the changes with time in visual attention, is used in our method. Then we calculate the pre-trained CNN-based visual features from the images corresponding to each frame of the fourth-order GIT as gaze-based visual features. In the second step (Sec. II-B), we calculate brain activity-based features [10] from each user. Furthermore, we perform CCA between the gaze-based visual features and the brain activity-based features at each frame in order to project the gaze-based visual features to novel features with consideration of the changes with time. In the final step (Sec. II-C), we align all projected features and construct the new third-order tensor. Then we estimate human emotions by using tensor-based machine learning.

A. CONSTRUCTION OF FOURTH-ORDER GIT AND EXTRACTION OF GAZE-BASED VISUAL FEATURES

The extraction of gaze-based visual features is shown in this subsection. For analyzing the image and the changes with time in visual attention, simultaneously, we construct the fourth-order GIT. The first and second modes d_1 and d_2 of the fourth-order GIT are pixel locations, the third mode d_3 corresponds to color channels, and the fourth mode d_4 means the number of samples in the time axis, that is, the number of frames. Given training images $\mathcal{X}_n^{\text{img}} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ ($n = 1, 2, \cdots, N$; N being the number of training images) and corresponding eye gaze data, we obtain gaze weight $\boldsymbol{G}_{n,f}^{\text{gaze}} \in \mathbb{R}^{d_1 \times d_2}$ of each frame f from the eye gaze data by applying a Gaussian filter to a fixation map obtained from the eye gaze data. Note that the fixation map, which is a gray-scale map, represents the human gaze area. In order to represent visual attention with the changes over time, we obtain a gaze and image weight (GIW) matrix $\boldsymbol{G}_{n,f}^{\text{GIW}} \in \mathbb{R}^{d_1 \times d_2}$ of each frame f by the following equation:

$$\boldsymbol{G}_{n,f}^{\text{GIW}} = d_4 \frac{\boldsymbol{G}_{n,f}^{\text{gaze}}}{\sum_{f=1}^{d_4} \boldsymbol{G}_{n,f}^{\text{gaze}}} + \boldsymbol{E}, \tag{1}$$

where all elements of a matrix $\boldsymbol{E} \in \mathbb{R}^{d_1 \times d_2}$ are one. We construct the fourth-order GIT $\mathcal{X}_n^{\text{4th}} \in \mathbb{R}^{d_1 \times d_2 \times d_3 \times d_4}$ as follows:

$$\boldsymbol{X}_{n,ch,f}^{\text{4th}} = \boldsymbol{X}_{n,ch}^{\text{img}} \circ \boldsymbol{G}_{n,f}^{\text{GIW}}, \tag{2}$$

where $\boldsymbol{X}_{n,ch,f}^{\text{4th}} \in \mathbb{R}^{d_1 \times d_2}$ and $\boldsymbol{X}_{n,ch}^{\text{img}} \in \mathbb{R}^{d_1 \times d_2}$ ($ch = 1, 2, \cdots, d_3$) are respectively a part of the fourth-order GIT $\mathcal{X}_n^{\text{4th}}$ and the image $\mathcal{X}_n^{\text{img}}$. Note that "$\circ$" represents the Hadamard product operator. The fourth-order GIT can reconstruct the original image as follows:

$$\mathcal{X}_n^{\text{img}} = \frac{1}{2d_4} \sum_{f=1}^{d_4} \mathcal{X}_{n,f}^{\text{4th}}, \tag{3}$$

where $\mathcal{X}_{n,f}^{\text{4th}} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is the part of $\mathcal{X}_n^{\text{4th}}$ in frame f. Note that we need to provide numeric values to the area which human does not gaze at for enabling the reconstruction based

on Eq. (3). Thus, we adopt the matrix $\boldsymbol{E}$ for which all elements are one.

In order to obtain more semantic gaze-based visual features, we extract three kinds of visual features from the image corresponding to each frame of the fourth-order GIT. In order to improve the representation ability for human emotions, we use several kinds of visual features. Since the visual features calculated from the GIT represent objects in images and the changes with time of visual attention, we regard these visual features as gaze-based visual features. Then, as visual features, we use the outputs of an intermediate layer of several kinds of CNN models since CNNs are well known in the field of object recognition [30]. Thus, by constructing the GIT based on CNN features, we can obtain gaze-based visual features that represent objects humans are viewing.

A large amount of training data is needed in order to train CNNs. However, the preparation of a large number of GIT is difficult since eye gaze data obtained from one user are limited. Thus, we perform transfer learning, which has been reported to be effective [31]. By using the ImageNet dataset [27], we pre-train the CNNs. In the proposed method, we use three kinds of state-of-the-art CNN models, Xception (X) [32], InceptionResnet-v2 (I) [33] and Densenet201 (D) [34]. We extract visual features $\boldsymbol{v}^p_{n,f} \in \mathbb{R}^{d_p}$ ($p \in \{\mathrm{X}, \mathrm{I}, \mathrm{D}\}$, d_p being the dimension of the outputs obtained from the last pooling layer of p) based on the pre-trained CNN from the image corresponding to each frame of the fourth-order GIT $\mathcal{X}^{4\mathrm{th}}_{n,f}$. Therefore, we extract gaze-based visual features with consideration of the objects in images and the changes with time of visual attention from the novel image representation, the fourth-order GIT.

B. EXTRACTION OF BRAIN ACTIVITY-BASED FEATURES AND CCA-BASED PROJECTION

In this subsection, we explain the extraction of brain activity-based features and the CCA-based projection with consideration of the changes with time. There are many kinds of brain activity data such as data obtained from EEG, fMRI and fNIRS. Data for brain activity obtained from EEG and from fNIRS are well known as having high temporal resolution. Specifically, since fNIRS measures blood oxygenation changes, fNIRS would be robust enough to avoid the effects of external activities such as eye blinks that occur while users view images [35]. Moreover, fNIRS equipment has few behavioral or physical restrictions on users [36]. Therefore, several studies focus on the relationship between human emotions and fNIRS signals [37]–[39], and we use fNIRS signals with eye gaze data in this study. There have been some studies in which both fNIRS and eye gaze data were used [40]–[42]. Specifically, in order to obtain fNIRS signals, we measure changes in deoxygenated and oxygenated hemoglobin levels from the head cortex by using near-infrared light. We calculate fNIRS features from fNIRS signals while each user is viewing images based on [10]. Concretely, we calculate the following 11-dimensional features from each channel in each frame as shown in Fig. 3.

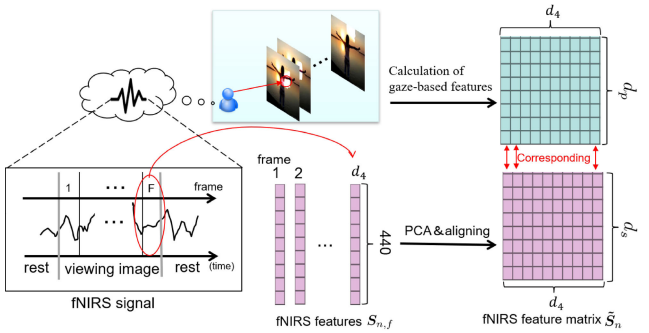


FIGURE 3. Calculation of brain activity-based features. In the proposed method, brain activity-based features correspond to gaze-based visual features in each frame.

- Statistical features (six dimensions)
  We calculate general statistics including average, variance, skewness, kurtosis, zero-crossing-rate and root-mean square of fNIRS signals in each time domain.
- Wavelet transform [43]-based features (five dimensions)
  By applying discrete wavelet transform to fNIRS signals, we can convert fNIRS signals into a frequency domain including high-frequency and low-frequency components. Then we calculate the ratio of energy to the total energy in each frequency component.

In the proposed method, we obtain fNIRS signals from ten channels in the front and back of the head, respectively. Note that we measure the changes in both oxygenated and deoxygenated hemoglobin levels, information that is biologically important for brain function. From the above, the dimension of fNIRS features is 440 calculated as 11-dimensional features × 20 channels × 2 (oxygenated/deoxygenated). In this way, we calculate fNIRS features $\boldsymbol{s}_{n,f} \in \mathbb{R}^{440}$ corresponding to a frame f of n-th image. Note that we apply dimension reduction to fNIRS features $\boldsymbol{s}_{n,f}$ since CCA tends to overfit the training data when the dimension of fNIRS features $\boldsymbol{s}_{n,f}$ is higher than the number of training images. As a dimension reduction method, we use principal component analysis (PCA) [44]. Consequently, we can newly obtain fNIRS features $\tilde{\boldsymbol{s}}_{n,f} \in \mathbb{R}^{d_s}$ (d_s being the dimension of fNIRS features after applying their dimension reduction).

We perform CCA between the above fNIRS features $\tilde{\boldsymbol{s}}_{n,f}$ and the gaze-based features $\boldsymbol{v}^p_{n,f}$ at each frame f as shown in Fig. 4. Concretely, we calculate the optimal projection pair $(\hat{\boldsymbol{w}}^p_{s,f}, \hat{\boldsymbol{w}}^p_{v,f}) \in \mathbb{R}^{d_s} \times \mathbb{R}^{d_p}$ by solving the following maximization problem:

$$\max_{(\boldsymbol{w}^p_{s,f}, \boldsymbol{w}^p_{v,f})} \frac{({\boldsymbol{w}^p_{s,f}}^\top \boldsymbol{C}^p_{sv,f} \boldsymbol{w}^p_{v,f})}{\sqrt{{\boldsymbol{w}^p_{s,f}}^\top \boldsymbol{C}_{ss,f} \boldsymbol{w}^p_{s,f}} \sqrt{{\boldsymbol{w}^p_{v,f}}^\top \boldsymbol{C}^p_{vv,f} \boldsymbol{w}^p_{v,f}}}, \tag{4}$$

where $^\top$ is a transposition operator. Specifically, the variances $\boldsymbol{C}_{ss,f}, \boldsymbol{C}^p_{vv,f}$ and the covariance $\boldsymbol{C}^p_{sv,f}$ at frame f are calculated as follows:

$$\boldsymbol{C}_{ss,f} = \frac{1}{N} \tilde{\boldsymbol{S}}_f \tilde{\boldsymbol{S}}_f^\top,$$

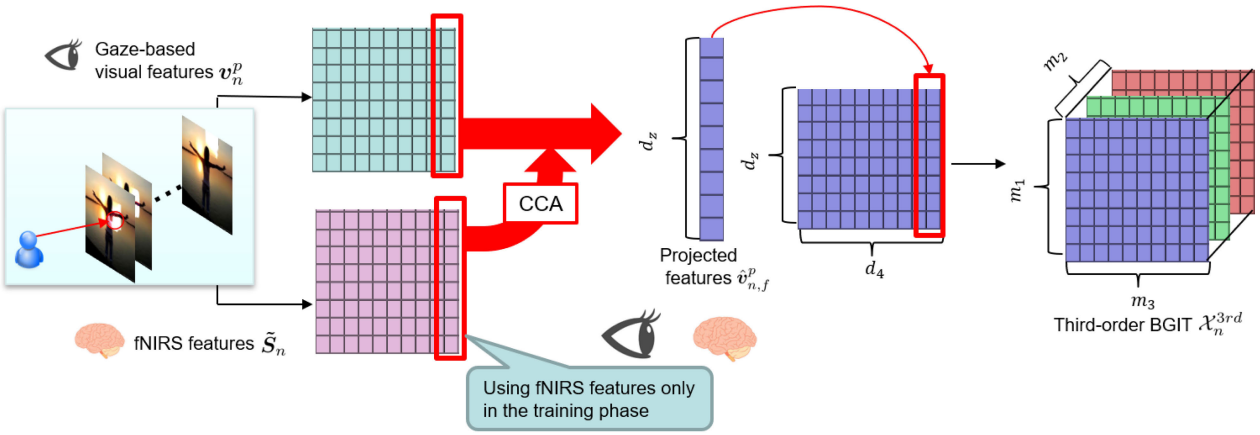


FIGURE 4. Projection on the basis of CCA between gaze-based visual features and fNIRS features at each frame. We use fNIRS features only in the training phase. The calculation at each frame enables consideration of the changes with time in visual attention and brain activity.

$$\boldsymbol{C}_{vv,f}^{p} = \frac{1}{N}\boldsymbol{V}_f^p \boldsymbol{V}_f^{p\top}, \tag{5}$$
$$\boldsymbol{C}_{sv,f}^{p} = \frac{1}{N}\tilde{\boldsymbol{S}}_f \boldsymbol{V}_f^{p\top},$$

where $\tilde{\boldsymbol{S}}_f = [\tilde{\boldsymbol{s}}_{1,f}, \tilde{\boldsymbol{s}}_{2,f}, \ldots, \tilde{\boldsymbol{s}}_{N,f}]$ and $\boldsymbol{V}_f^p = [\boldsymbol{v}_{1,f}^p, \boldsymbol{v}_{2,f}^p, \ldots, \boldsymbol{v}_{N,f}^p]$. Note that $\tilde{\boldsymbol{S}}_f$ and $\boldsymbol{V}_f^p$ are centered in each frame. Moreover, we can rewrite this maximization problem as follows:

$$(\hat{\boldsymbol{w}}_{s,f}^p, \hat{\boldsymbol{w}}_{v,f}^p) = \arg\max_{(\boldsymbol{w}_{s,f}^p, \boldsymbol{w}_{v,f}^p)} \boldsymbol{w}_{s,f}^{p\top} \boldsymbol{C}_{sv,f} \boldsymbol{w}_{v,f}^p$$
$$\text{s.t. } \boldsymbol{w}_{s,f}^{p\top} \boldsymbol{C}_{ss,f} \boldsymbol{w}_{s,f}^p = \boldsymbol{w}_{v,f}^{p\top} \boldsymbol{C}_{vv,f} \boldsymbol{w}_{v,f}^p = 1. \tag{6}$$

Furthermore, we obtain the following eigenvalue problem based on the Lagrange multiplier method and L1-regularization [45]:

$$\begin{bmatrix} \boldsymbol{O} & \boldsymbol{C}_{sv,f}^p \\ \boldsymbol{C}_{sv,f}^{p\top} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_{s,f}^p \\ \boldsymbol{w}_{v,f}^p \end{bmatrix}$$
$$= \lambda_f^p \begin{bmatrix} \boldsymbol{C}_{ss,f} + \zeta_s \boldsymbol{I}_s & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{C}_{vv,f}^p + \zeta_v \boldsymbol{I}_v^p \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_{s,f}^p \\ \boldsymbol{w}_{v,f}^p \end{bmatrix}, \tag{7}$$

where λ_f^p is the Lagrange multiplier, ζ_s and ζ_v are regularization parameters, and $\boldsymbol{I}_v^p$ and $\boldsymbol{I}_s$ are the identity matrices. Then we obtain the optimal projection pair $(\hat{\boldsymbol{w}}_{s,f}^p, \hat{\boldsymbol{w}}_{v,f}^p)$ by solving this eigenvalue problem. By using valid top d_z ($\leq \min(d_s, d_X, d_D, d_I)$) projection pairs in each frame, we obtain projection matrices $\hat{\boldsymbol{W}}_{s,f}^p \in \mathbb{R}^{d_z \times d_s}$ and $\hat{\boldsymbol{W}}_{v,f}^p \in \mathbb{R}^{d_z \times d_p}$.

Finally, we project the gaze-based visual features $\boldsymbol{v}_{n,f}^p$ by using the obtained projection matrices $\hat{\boldsymbol{W}}_{v,f}^p$ to calculate the projected features at each frame as follows:

$$\hat{\boldsymbol{v}}_{n,f}^p = \hat{\boldsymbol{W}}_{v,f}^p \boldsymbol{v}_{n,f}^p. \tag{8}$$

In this way, we can obtain the newly projected gaze-based visual features considering fNIRS features. Note that once we obtain the projection matrices, we can project new gaze-based visual features without acquiring fNIRS features. Then the above approach has the following two contributions. Since we need brain activity data only in the training phase, the burden on users is reduced. In addition, since the projection pair is calculated in each frame, we can consider the visual attention and the brain activity with the changes over time.

C. EMOTION ESTIMATION BASED ON TENSOR-BASED ANALYSIS

Emotion estimation on the basis of tensor-based analysis and simple machine learning is shown in this subsection. By using the projected features $\hat{\boldsymbol{v}}_{n,f}^p$, we newly construct the third-order GIT considering fNIRS (Brain activity-based Gaze and Image Tensor; BGIT) $\mathcal{X}_n^{3\text{rd}} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$ in order to analyze features with consideration of the changes over time. Note that $m_1 (= d_z)$ is the dimension of transformed features, $m_2 (= 3)$ is the kind of gaze-based visual features, and $m_3 (= d_4)$ is the number of frames of the fourth-order GIT. Concretely, we construct the third-order BGIT $\mathcal{X}_n^{3\text{rd}}$ as follows:

$$\mathcal{X}_{n,i,p,f}^{3\text{rd}} = \hat{\boldsymbol{v}}_{n,i,f}^p, \tag{9}$$

where $\mathcal{X}_{n,i,p,f}^{3\text{rd}}$ and $\hat{\boldsymbol{v}}_{n,i,f}^p$ are (i,p,f)-th element of the third-order BGIT $\mathcal{X}_n^{3\text{rd}}$ and i-th element of the projected gaze-based visual features $\hat{\boldsymbol{v}}_{n,f}^p$, respectively.

In the proposed method, we apply supervised feature transformation to the third-order BGIT $\mathcal{X}_n^{3\text{rd}}$ for improving discriminant ability. As a supervised feature transformation method, we use GTDA since it can be applied to tensors. GTDA is a tensor extension of the differential scatter discriminant criterion [46]. Concretely, in order to calculate the transformation set $\{\tilde{\boldsymbol{T}}_k \in \mathbb{R}^{m_k \times \tilde{m}_k}\}_{k=1}^3$ ($\tilde{m}_k < m_k$), we solve the following optimization problem:

$$\tilde{\boldsymbol{T}}|_{k=1}^3 = \arg\max_{\boldsymbol{T}|_{k=1}^3} \text{tr}\left(\boldsymbol{T}_k^\top \left(\boldsymbol{B}_k^b - \zeta_k \boldsymbol{B}_k^w\right) \boldsymbol{T}_k\right), \tag{10}$$

where ζ_k is the maximum eigenvalue of $\left(\boldsymbol{B}_k^w\right)^{-1} \boldsymbol{B}_k^b$. Note that $\boldsymbol{B}_k^w$ and $\boldsymbol{B}_k^b$ are respectively calculated by solving the

following equations:

$$\boldsymbol{B}_k^b = \sum_{j=1}^{c}\Bigg[n_j \mathrm{mat}_k\left(\left(\mathcal{M}_j - \mathcal{M}\right)\times_{\bar{k}} \boldsymbol{T}_k^\top\right) \mathrm{mat}_k^\top\left(\left(\mathcal{M}_j - \mathcal{M}\right)\times_{\bar{k}} \boldsymbol{T}_k^\top\right)\Bigg], \tag{11}$$

$$\boldsymbol{B}_k^w = \sum_{j=1}^{c}\sum_{l=1}^{n_j}\Bigg[\mathrm{mat}_k\left(\left(\mathcal{X}_l^{3\mathrm{rd}(j)} - \mathcal{M}_j\right)\times_{\bar{k}} \boldsymbol{T}_k^\top\right) \mathrm{mat}_k^\top\left(\left(\mathcal{X}_l^{3\mathrm{rd}(j)} - \mathcal{M}_j\right)\times_{\bar{k}} \boldsymbol{T}_k^\top\right)\Bigg], \tag{12}$$

where c is the number of classes, and $\mathcal{X}_l^{3\mathrm{rd}(j)}$ is l-th $\mathcal{X}_n^{3\mathrm{rd}}$ belonging to class j. Furthermore,

$$\mathcal{M}_j = \frac{1}{n_j}\sum_{l=1}^{n_j}\mathcal{X}_l^{3\mathrm{rd}(j)}, \tag{13}$$

$$\mathcal{M} = \frac{1}{N}\sum_{j=1}^{c} n_j \mathcal{M}_j, \tag{14}$$

where $\mathcal{M}_j$ is the mean tensor of j-th class, and $\mathcal{M}$ is the mean tensor of all tensors. Note that $\mathcal{X}_l^{3\mathrm{rd}(j)}\,|_{1\le j\le c}^{1\le l\le n_j}$, $\mathcal{M}_j\,|_{j=1}^{c}$, and $\mathcal{M}$ are all third-order tensors with sizes of $m_1 \times m_2 \times m_3$. Then we use tensor representation based on [28]. Specifically, a matrix $\mathrm{mat}_l(\mathcal{X}_l^{3\mathrm{rd}(j)}) \in \mathbb{R}^{m_1 \times \prod_{k\neq 1} m_k}$ represents the mode-l matricizing of the third-order BGIT $\mathcal{X}_l^{3\mathrm{rd}(j)}$. In addition, we denote the mode-k product of the third-order BGIT $\mathcal{X}_l^{3rd}$ and a matrix $\boldsymbol{T}_1$ as $\mathcal{X}_l^{3rd} \times_k \boldsymbol{T}_1$. Notably, $\mathcal{X}_l^{3\mathrm{rd}} \times_1 \boldsymbol{T}_1 \times_3 \boldsymbol{T}_3 = \mathcal{X}_l^{3\mathrm{rd}} \times_{\bar{2}} \boldsymbol{T}_2$ for example. We obtain a transformed tensor $\tilde{\mathcal{X}}_n^{3\mathrm{rd}}$ as follows:

$$\tilde{\mathcal{X}}_n^{3\mathrm{rd(j)}} = \mathcal{X}_n^{3\mathrm{rd}} \prod_{k=1}^{3} \times_k \tilde{\boldsymbol{T}}_k. \tag{15}$$

Therefore, we can obtain the transformed third-order BGIT $\tilde{\mathcal{X}}_n^{3\mathrm{rd}}$ by using GTDA.

Finally, the proposed method estimates human emotions by training the ELM-based classifier for which inputs are the transformed third-order BGIT $\tilde{\mathcal{X}}_n^{3\mathrm{rd}}$. ELM, which is a single-hidden layer feedforward network, can learn the emotion even if the amount of training samples is small by using random values as parameters of only one hidden layer. Note that ELM is almost the same concept as the randomized neural network [47] and the random vector functional link [48]. ELM calculates weights $\boldsymbol{\beta}$ between the hidden layer and the output layer as follows:

$$\boldsymbol{\beta} = \boldsymbol{H}^\dagger \boldsymbol{Y}, \tag{16}$$

where $\boldsymbol{Y}$ is a class matrix calculated as follows:

$$\boldsymbol{Y} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,c} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,c} \\ \vdots & & \ddots & \vdots \\ y_{N,1} & y_{N,2} & \cdots & y_{N,c} \end{bmatrix}, \tag{17}$$

where $y_{n,j}$ is a binary value. If n-th image belongs to class j, $y_{n,j} = 1$, otherwise $y_{n,j} = 0$. Moreover, $\boldsymbol{H}^\dagger$ is Moore-Penrose generalized inverse [49] of the hidden layer's outputs calculated as follows:

$$\boldsymbol{H} = [\boldsymbol{h}(\tilde{\boldsymbol{x}}_1), \boldsymbol{h}(\tilde{\boldsymbol{x}}_2), \ldots, \boldsymbol{h}(\tilde{\boldsymbol{x}}_N)]^\top, \tag{18}$$

where $\tilde{\boldsymbol{x}}_n$ is obtained by vectorizing the third-order BGIT $\tilde{\boldsymbol{X}}_n^{3\mathrm{rd}}$. We calculate $\boldsymbol{h}(\tilde{\boldsymbol{x}}_n)$ by adopting an activation function g to $\tilde{\boldsymbol{x}}_n$ as follows:

$$\boldsymbol{h}(\tilde{\boldsymbol{x}}_n) = [g(\boldsymbol{a}_1, b_1, \tilde{\boldsymbol{x}}_n), g(\boldsymbol{a}_2, b_2, \tilde{\boldsymbol{x}}_n), \ldots, g(\boldsymbol{a}_Q, b_Q, \tilde{\boldsymbol{x}}_n)]^\top \tag{19}$$

Generally, the following sigmoid function is used as the activation function.

$$g(\boldsymbol{a}_q, b_q, \tilde{\boldsymbol{x}}_n) = \frac{1}{1 + \exp\left(\boldsymbol{a}_q^\top \tilde{\boldsymbol{x}}_n + b_q\right)}, \tag{20}$$

where $q = 1, 2, \ldots, Q$ (Q being the number of hidden nodes), and $\boldsymbol{a}_q$ and b_q are random values following a uniform distribution.

For a test vector $\tilde{\boldsymbol{x}}$, the outputs $f(\tilde{\boldsymbol{x}})$ of the ELM-based classifier are calculated as follows:

$$f(\tilde{\boldsymbol{x}}) = \boldsymbol{h}(\tilde{\boldsymbol{x}})^\top \boldsymbol{\beta}. \tag{21}$$

Finally, the proposed method estimates emotions by comparing elements of $f(\tilde{\boldsymbol{x}})$ and detecting the highest value. In this way, the proposed method can estimate human emotions by applying the tensor analysis and the simple machine learning to the third-order BGIT $\tilde{\mathcal{X}}_n^{3\mathrm{rd}}$.

III. EXPERIMENTAL RESULTS

In this section, we show experimental results for verifying the effectiveness of our method for the emotion estimation.

A. EXPERIMENTAL SETTINGS

In this experiment, we used Tobii Eye tracker 4C[1] for obtaining eye gaze data and LIGHTNIRS[2] for recording fNIRS signals. Then we used 20 channels, including 10 channels on the front of the head and 10 channels on the back of the head, as shown in Fig. 5. The participants gazed at the image on a 15-inch display at a distance of 70 cm. Moreover, the participants wore a head cap in order to measure fNIRS signals, and the gaze sensor were placed on the display.

As viewed images, we used the art photo dataset published in [50]. This dataset includes images that are given to a single label of eight emotional labels (*Amusement, Awe, Contentment, Excitement, Sad, Fear, Anger and Disgust*), and we used 10 images belonging to each emotional label, totally 80 images. Moreover, we randomly selected 64 images as training images and used the remaining images as test images.

There were 10 participants (Pars. 1-10), including seven healthy men and three healthy women, in this experiment.[3]

[1] https://tobiigaming.com/eye-tracker-4c/

[2] http://www.shimadzu.com/

[3] This human research was conducted with the approval by the ethical committee in Hokkaido University.

TABLE 1. Numbers of emotions for participants.

	Par1	Par2	Par3	Par4	Par5	Par6	Par7	Par8	Par9	Par10
Training Image (Positive)	29	28	30	28	36	23	31	35	28	39
Training Image (Negative)	35	37	34	36	28	41	33	29	36	25
Test Image (Positive)	8	7	8	8	9	7	8	7	7	7
Test Image (Negative)	8	9	8	8	7	9	8	9	9	9

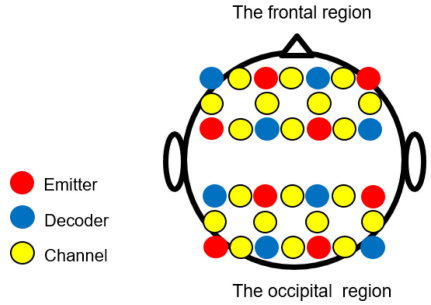


FIGURE 5. Positions of channels. We obtained fNIRS signals from 20 channels by using emitters and decoders.

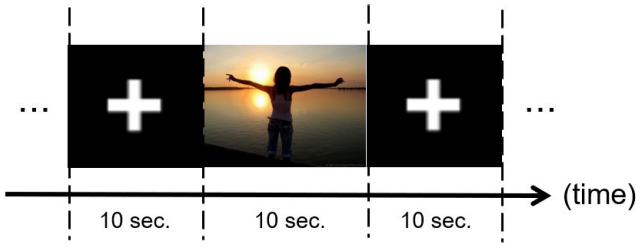


FIGURE 6. Our experimental design. Participants gazed at each image for ten seconds with an inter-stimulus interval of ten seconds. Then we obtained eye gaze data and fNIRS signals from each participant.

The participants were instructed to gaze at each image for ten seconds with an inter-stimulus interval of ten seconds for preventing fNIRS signals from being influenced by the previous image as shown in Fig. 6. In the interval, we showed an image that had a cross mark in the center in order to avoid the influence of the previous image and to lead the gaze to the center of the monitor. After the task, subjects provide feedbacks (positive/negative) as ground truths about their emotion induced by viewing the images. We adopted two emotional states for increasing the number of images belonging to each category when we obtained the feedbacks of participants. Table 1 shows the number of emotions for each participant. It was confirmed that there was not a large difference between the numbers of emotions. For verifying the effectiveness of the proposed method, we used eight comparative methods (CMs1-8). Their details are shown below.

CM1. This method is similar to our method but uses only one CNN feature instead of multiple CNN features.

CM2. This method estimates emotions by using only gaze information of our method.

CM3. This method uses fNIRS features calculated as shown in Sec. II-B. Moreover, fNIRS features are input into the ELM-based classifier to estimate emotions.

CM4. This method treats two modalities, gaze and fNIRS features. Then a novel gaze feature [51] is adopted. Moreover, the fNIRS features [10] are extracted in the same manner as our method. Finally, two ELM-based classifiers corresponding to each modality are constructed, and emotions are estimated based on late fusion. Note that late fusion is one of the major fusion methods, and several researchers have used this method with multimodal analysis [52]–[54].

In addition to the above comparative methods which consist of the parts of our method, we compared our method with the following state-of-the-art methods.

CM5. This is an emotion estimation method [17] that uses Deep Canonical Correlation Analysis (Deep CCA) [55] between gaze and brain information. Then three kinds of CNN features obtained by the GIT are adopted as gaze features, and Deep CCA is applied to each CNN features and fNIRS features. Finally, Support Vector Machine (SVM) [56] is trained by using the projected features via Deep CCA.

CM6. This is an emotion estimation method [18] that uses bimodal deep autoencoder (BDAE). The BDAE is trained for reconstracting the inputs which are fNIRS and gaze features, and we can extract the combined high-level features. Finally, SVM estimates the human emotions by inputting the combined high-level features obtained from the hidden layer of BDAE. Note that fNIRS and gaze features are calculated in the same manner as our method without consideration of the changes with time, that is, $d_4 = 1$, and these features are vectorized.

CM7. This is an emotion estimation method [19] which is an extended version of CM6. This method performs the multilayer perceptron-based regression from gaze features to the combined high-level features calculated from CM6. Finally, SVM estimates the human emotions. Note that gaze features are calculated in the same manner as CM6.

CM8. This is an emotion estimation method [20] that uses long short-term memory (LSTM) [57]. The inputs of LSTM are gaze and fNIRS features, and LSTM extracts the combined high-level features with consideration of the changes with time. Finally,

TABLE 2. Average values for participants. The average of all participants and its standard deviation are also shown.

	Par1	Par2	Par3	Par4	Par5	Par6	Par7	Par8	Par9	Par10	Average ± std	p-value
Ours (I-D-X)	0.80	0.71	0.82	0.71	0.77	0.63	**0.80**	0.84	0.75	0.71	0.76 ± 0.06	
Ours (I-X-D)	**0.88**	**0.80**	0.71	**0.80**	0.77	0.63	0.75	0.86	**0.88**	0.67	$\mathbf{0.77 \pm 0.08}$	
Ours (X-I-D)	0.82	0.62	**0.88**	0.62	**0.88**	0.56	0.77	**0.89**	**0.88**	0.71	0.76 ± 0.12	
CM1 (X)	0.78	0.62	0.74	0.50	0.63	0.59	0.50	0.74	0.59	0.35	0.60 ± 0.12	$p < 0.01$
CM1 (I)	0.67	0.59	0.70	0.46	0.67	0.53	0.57	0.74	0.63	0.63	0.62 ± 0.08	$p < 0.01$
CM1 (D)	0.53	0.50	0.53	0.56	0.67	0.53	0.56	0.67	0.46	0.67	0.57 ± 0.07	$p < 0.01$
CM2 (I-D-X)	0.67	0.63	0.43	0.31	0.33	0.59	0.46	0.71	0.63	0.71	0.54 ± 0.14	$p < 0.01$
CM2 (I-X-D)	0.67	0.53	0.53	0.36	0.63	0.43	0.47	0.63	0.63	0.67	0.55 ± 0.10	$p < 0.01$
CM2 (X-I-D)	0.67	0.71	0.57	0.36	0.31	0.40	0.62	0.56	0.70	0.44	0.53 ± 0.14	$p < 0.01$
CM3	0.53	0.47	0.53	0.22	0.50	0.59	0.53	0.74	0.63	0.57	0.53 ± 0.12	$p < 0.01$
CM4 [51]	0.53	0.43	0.47	0.22	0.36	0.44	0.14	0.63	0.50	0.59	0.43 ± 0.15	$p < 0.01$
CM5 [17]	0.40	0.44	0.40	0.22	0.46	0.27	0.20	0.63	0.40	0.61	0.40 ± 0.14	$p < 0.01$
CM6 [18]	0.63	0.62	0.62	0.67	0.62	**0.67**	0.50	0.75	0.62	0.63	0.63 ± 0.06	$p < 0.01$
CM7 [19]	0.71	0.77	0.71	0.57	0.33	0.57	0.67	0.71	0.75	**0.78**	0.66 ± 0.13	$p < 0.02$
CM8 [20]	0.50	0.50	0.50	0.50	0.40	0.50	0.50	0.72	0.61	0.61	0.53 ± 0.08	$p < 0.01$

SVM estimates the human emotion by inputting the combined high-level features obtained from the last sequence of LSTM. Note that gaze and fNIRS features are calculated in the same manner as our method.

We adopted SVM instead of ELM by following [17]–[20] in CMs5-8. In CMs5-7, we calculate features without consideration of the changes with time since these methods do not have the mechanism which can consider the time changes.

In this experiment, as an evaluation index, we adopted F1-measure calculated as follows:

$$\text{F1-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}, \tag{22}$$

where Recall and Precision are calculated by using the obtained classification results as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{23}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{24}$$

TP, FN and FP mean the numbers of images estimated to be true positive, false negative and false positive, respectively.

B. PERFORMANCE EVALUATION

The experimental results are shown in Table 2. This table shows the results of F1-measures, and it is confirmed that our method outperforms all of the CMs in average. In this experiment, we attempted several combinations of CNN features and the order of CNN features is expressed as "ours (a-b-c)". Note that a, b and c represent one of X (Xception), I (InceptionResnet-v2) and D (Densenet201), and ours (a-b-c) represents our method and the order of CNN features. As shown in Table 2, our method is superior to CMs in any combination. However, a difference depending on the combination order of CNN features was not confirmed.

A comparison of our method and CM1 showed that consideration of the changes with time in visual attention and brain activity is effective for the emotion estimation. Moreover, a comparison of our method with CM2 and CM3 showed that the collaborative use of gaze and brain information is effective for the emotion estimation. A comparison of our method with CM4, which uses the novel gaze features [51] and a traditional feature fusion method [21] showed that our method is more effective as a multimodal method. CM4 uses novel gaze and fNIRS features without considering the changes with time. Thus, we can confirm that the GIT-based extraction method for gaze features and the CCA-based fusion method are effective for the emotion estimation. Finally, comparisons of our method with CMs5-8 showed that our method is more effective to the human emotion estimation than the state-of-the-art methods. Although CMs5, 8 were some of the state-of-the-art emotion estimation methods based on the collaborative use of gaze and brain information, the results of CMs5, 8 were not good. We guess that the amount of training data belonging to each category was too small to train Deep CCA or LSTM, which had a large number of training parameters to be optimized. Thus, Deep CCA or LSTM seemed not to be trained sufficiently, and its estimation accuracy was not high. On the other hand, the BDAE used by CMs6, 7 was one of unsupervised learning methods and the amount of training data was larger than Deep CCA and LSTM. Thus, BDAE was considered to be relatively optimized. Moreover, we performed Welch's t-test [58] between ours(I-D-X) and CMs, and confirmed the statistical superiority.

Figure 7 shows examples of estimation results for one of the participants. Figures 7 (a) and (b) show images for which ours (I-X-D) estimated true emotions and Fig. 7 (c) and (d) show images for which ours (I-X-D) estimated false emotions. Clearly, the images in which ours (I-X-D) estimated true emotions have differences in brightness and the objects. The image in Fig. 7 (a) is bright overall and the object in the image would evoke a positive emotion in most humans, whereas the image in Fig. 7 (b) is dark overall and


(a) Example of true positive


(b) Example of true negative


(c) Example of false positive


(d) Example of false negative

FIGURE 7. Examples of estimation results for Par1. (a) and (b) show that ours (I-X-D) estimated true emotion. (c) and (d) show that ours (I-X-D) estimated false emotion.

the object in the image would evoke a negative emotion in most humans. Thus, these images are easily classified as true emotions. On the other hand, the image in Fig. 7 (c) is dark overall and the object in the image resembles a flower. Generally, flowers are related to positive emotions. Thus, ours (I-X-D) may estimate positive emotion from the characteristics of a flower. The image in Fig. 7 (d) mainly consists of white and black colors and includes a fox. Since the monotone composition may evoke negative emotion in humans, ours (I-X-D) estimates negative emotion when the participant gazes at the image in Fig. 7 (d).

From the above-described qualitative and quantitative evaluations, we can verify that our method is effective for the human emotion estimation and can also find its limitations.

IV. CONCLUSION

In this paper, we have proposed a human-centric emotion estimation method based on correlation maximization that considers the changes with time in both visual attention and brain activity. In the proposed method, we focus on two signals that represent the changes with time of visual attention with respect to objects in images and brain activity. Then we construct two feature extraction networks in order to consider the above two signals. We project the gaze-based features by maximizing the correlations with fNIRS features. Also, we can perform the projection with consideration of the changes over time by focusing on the fourth mode of GIT. This projection based on CCA between the gaze-based CNN features and fNIRS features is the biggest contribution of this paper. Thus, our method realizes the emotion estimation by using only gaze information, that is, it does not need brain activity data in the test phase. We have realized the human-centric emotion estimation, and its effectiveness has been verified from the experimental results. We will consider the use of the data obtained from other users by clarifying the relationship of the data obtained from some users in the future work.

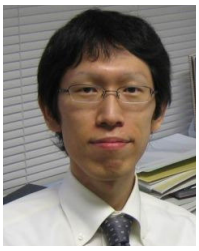
REFERENCES

[1] R. W. Picard, *Affective Computing*, vol. 167. Cambridge, MA, USA: MIT Press, 1997, p. 170.
[2] W. Wang and Q. He, "A survey on emotional semantic image retrieval," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 117–120.
[3] 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.
[4] 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.
[5] H. Ye, M. Malu, U. Oh, and L. Findlater, "Current and future mobile and wearable device use by people with visual impairments," in *Proc. 32nd Annu. ACM Conf. Hum. factors Comput. Syst. CHI*, 2014, pp. 3123–3132.
[6] J. Hernandez, Y. Li, J. Rehg, and R. Picard, "BioGlass: Physiological parameter estimation using a head-mounted wearable device," in *Proc. 4th Int. Conf. Wireless Mobile Commun. Healthcare -Transforming Healthcare Through Innov. Mobile Wireless Technol.*, 2014, pp. 55–58.
[7] S. Ishimaru, K. Kunze, K. Kise, J. Weppner, A. Dengel, P. Lukowicz, and A. Bulling, "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. AH*, 2014, p. 15.
[8] K. Sugata, T. Ogawa, and M. Haseyama, "Selection of significant brain regions based on MvGTDA and TS-DLF for emotion estimation," *IEEE Access*, vol. 6, pp. 32481–32492, 2018.
[9] 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, Dec. 2013.
[10] K. Tai and T. Chau, "Single-trial classification of NIRS signals during emotional induction tasks: Towards a corporeal machine interface," *J. NeuroEng. Rehabil.*, vol. 6, no. 1, p. 39, Dec. 2009.
[11] P. Vuilleumier, "How brains beware: Neural mechanisms of emotional attention," *Trends Cognit. Sci.*, vol. 9, no. 12, pp. 585–594, Dec. 2005.
[12] R. J. Compton, "The interface between emotion and attention: A review of evidence from psychology and neuroscience," *Behav. Cognit. Neurosci. Rev.*, vol. 2, no. 2, pp. 115–129, Jun. 2003.
[13] H. Zheng, T. Chen, Q. You, and J. Luo, "When saliency meets sentiment: Understanding how image content invokes emotion and sentiment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 630–634.
[14] K. Pasupa, P. Chatkamjuncharoen, C. Wuttilertdeshar, and M. Sugimoto, "Using image features and eye tracking device to predict human emotions towards abstract images," in *Proc. Pacific–Rim Symp. Image Video Technol. (PSIVT)*, 2015, pp. 419–430.
[15] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7521–7531.
[16] N. Naseer and K.-S. Hong, "FNIRS-based brain-computer interfaces: A review," *Frontiers Hum. Neurosci.*, vol. 9, p. 3, Jan. 2015.
[17] J. Qiu, W. Liu, and B. Lu, "Multi-view emotion recognition using deep canonical correlation analysis," in *Proc. Int. Conf. Neural Inf. Process. (ICONIP)*, 2018, pp. 221–231.
[18] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *Proc. Int. Conf. Neural Inf. Process. (ICONIP)*, 2016, pp. 521–529.
[19] H.-F. Jiang, X.-Y. Guan, W.-Y. Zhao, L.-M. Zhao, and B.-L. Lu, "Generating multimodal features for emotion classification from eye movement signals," in *Proc. Int. Conf. Neural Inf. Process. (ICONIP)*, 2019, pp. 59–66.
[20] H. Tang, W. Liu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep neural networks," in *Proc. Int. Conf. Neural Inf. Process. (ICONIP)*, 2017, pp. 811–819.

[21] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, Dec. 1936.
[22] Y. Moroto, K. Maeda, T. Ogawa, and M. Haseyama, "Estimation of emotion labels via tensor-based spatiotemporal visual attention analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4105–4109.
[23] M. Liu, Y. Fu, and T. S. Huang, "An audio-visual fusion framework with joint dimensionality reducton," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 4437–4440.
[24] Q.-S. Sun, Z.-D. Liu, P.-A. Heng, and D.-S. Xia, "A theorem on the generalized canonical projective vectors," *Pattern Recognit.*, vol. 38, no. 3, pp. 449–452, Mar. 2005.
[25] K.-H. Pong and K.-M. Lam, "Multi-resolution feature fusion for face recognition," *Pattern Recognit.*, vol. 47, no. 2, pp. 556–567, Feb. 2014.
[26] E. H. El-Shazly, M. M. Abdelwahab, A. Shimada, and R.-I. Taniguchi, "Real time algorithm for efficient HCI employing features obtained from MYO sensor," in *Proc. IEEE 59th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Oct. 2016, pp. 1–4.
[27] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2012, pp. 1097–1105.
[28] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.
[29] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jul. 2004, pp. 985–990.
[30] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 478–487.
[31] S. Jialin Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[32] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2016, *arXiv:1610.02357*. [Online]. Available: http://arxiv.org/abs/1610.02357
[33] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[35] A. Girouard, E. T. Solovey, L. M. Hirshfield, E. M. Peck, K. Chauncey, A. Sassaroli, S. Fantini, and R. J. Jacob, "From brain signals to adaptive interfaces: Using fNIRS in HCI," in *Brain-Computer Interfaces*. London, U.K.: Springer, 2010, pp. 221–237.
[36] Y. Hoshi, "Near-infrared spectroscopy for studying higher cognition," in *Neural Correlates of Thinking*. Berlin, Germany: Springer, 2009. pp. 83–93.
[37] D. Heger, R. Mutter, C. Herff, F. Putze, and T. Schultz, "Continuous recognition of affective states by functional near infrared spectroscopy signals," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 832–837.
[38] D. Bandara, S. Velipasalar, S. Bratt, and L. Hirshfield, "Building predictive models of emotion with functional near-infrared spectroscopy," *Int. J. Hum.-Comput. Stud.*, vol. 110, pp. 75–85, Feb. 2018.
[39] T. Gruber, C. Debracque, L. Ceravolo, K. Igloi, B. M. Bosch, S. Frühholz, and D. Grandjean, "Human discrimination and categorization of emotions in voices: A functional near-infrared spectroscopy (fNIRS) study," *Frontiers Neurosci.*, vol. 14, p. 570, Jun. 2020.
[40] Y. Kita, A. Gunji, K. Sakihara, M. Inagaki, M. Kaga, E. Nakagawa, and T. Hosokawa, "Scanning strategies do not modulate face identification: Eye-tracking and near-infrared spectroscopy study," *PLoS ONE*, vol. 5, no. 6, pp. 1–10, Jun. 2010.
[41] Y. Suzuki, K. Shirahada, M. Kosaka, and A. Maki, "A new marketing methodology by integrating brain measurement, eye tracking, and questionnaire analysis," in *Proc. ICSSSM*, Jul. 2012, pp. 770–773.
[42] K. Fujiwara, N. Kiyota, K. Kunita, M. Yasukawa, K. Maeda, and X. Deng, "Eye movement performance and prefrontal hemodynamics during saccadic eye movements in the elderly," *J. Physiolog. Anthropology*, vol. 29, no. 2, pp. 71–78, 2010.
[43] M. J. Shensa, "The discrete wavelet transform: Wedding the a trous and mallat algorithms," *IEEE Trans. Signal Process.*, vol. 40, no. 10, pp. 2464–2482, Oct. 1992.
[44] K. Pearson, "On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Phil. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.
[45] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman, "Canonical correlation analysis when the data are curves," *J. Roy. Statist. Soc. Ser. B, Methodol.*, vol. 55, no. 3, pp. 725–740, 1993.
[46] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Amsterdam, The Netherlands: Elsevier, 2013.
[47] W. F. Schmidt, M. A. Kraaijveld, and R. P. W. Duin, "Feedforward neural networks with random weights," in *Proc. 11th IAPR Int. Conf. Pattern Recognit. Conf. B, Pattern Recognit. Methodol. Syst.*, vol. 2, 1992, pp. 1–4.
[48] Y.-H. Pao, G.-H. Park, and D. J. Sobajic, "Learning and generalization characteristics of the random vector functional-link net," *Neurocomputing*, vol. 6, no. 2, pp. 163–180, Apr. 1994.
[49] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, May 1993.
[50] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. Int. Conf. Multimedia - MM*, 2010, pp. 83–92.
[51] N. Karessli, Z. Akata, B. Schiele, and A. Bulling, "Gaze embeddings for zero-shot image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017.
[52] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017.
[53] G. Cai and B. X. Cai, "Convolutional neural networks for multimedia sentiment analysis," in *Natural Language Processing and Chinese Computing*. Cham, Switzerland: Springer, 2015, pp. 159–167.
[54] M. Soleymani, M. Riegler, and P. Halvorsen, "Multimodal analysis of image search intent: Intent recognition in image search from user behavior and visual content," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2017, pp. 251–259.
[55] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 1247–1255.
[56] V. Vapnik, "Pattern recognition using generalized portrait method," *Automat. Remote Control*, vol. 24, pp. 774–780, 1963.
[57] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[58] B. L. Welch, "The significance of the difference between two means when the population variances are unequal," *Biometrika*, vol. 29, nos. 3–4, pp. 350–362, Feb. 1938.

YUYA MOROTO (Graduate Student Member, IEEE) received the B.S. degree in electronics and information engineering from Hokkaido University, Japan, in 2019, where he is currently pursuing the master's degree with the Graduate School of Information Science and Technology. His research interests include biological processing and multimedia retrieval. He is a Student Member of the IEICE.

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



TAKAHIRO OGAWA (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Japan, in 2003, 2005, and 2007, respectively. He joined the Graduate School of Information Science and Technology, Hokkaido University, in 2008, where he is currently an Associate Professor with the Faculty of Information Science and Technology. His research interests include AI, the IoT, and big data analysis for multimedia signal processing and its applications. He is a member of ACM, IEICE, and ITE. He was the Special Session Chair of the IEEE ISCE2009, the Doctoral Symposium Chair of ACM ICMR2018, the Organized Session Chair of the IEEE GCCE2017–2019, the TPC Vice Chair of the IEEE GCCE2018, and the Conference Chair of the IEEE GCCE2019. He has also been an Associate Editor of *ITE Transactions on Media Technology and Applications*.

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