HUMAN EMOTION RECOGNITION USING MULTI-MODAL BIOLOGICAL SIGNALS BASED ON TIME LAG-CONSIDERED CORRELATION MAXIMIZATION

Yuya Moroto[†], *Keisuke Maeda*[††], *Takahiro Ogawa*[†††] *and Miki Haseyama*[†††]

[†]Graduate School of Information Science and Technology, Hokkaido University, Japan
[††] Office of Institutional Research, Hokkaido University, Japan
[†††] Faculty of Information Science and Technology, Hokkaido University, Japan
E-mail: {moroto, maeda, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

ABSTRACT

A human emotion recognition using multi-modal biological signals based on time lag-considered correlation maximization is presented in this paper. Various multi-modal emotion recognition methods for visual stimuli have been studied and they focus on gaze and brain activity data. The visual stimuli captured by human eyes are sent to the brain by neurotransmitters. Thus, there is a time lag between gaze data, which record where humans gaze at, and brain activity data. However, most of the previous methods only integrate features obtained from each data without considering such a time lag. The proposed method newly introduces the mechanism to consider the time lag into the canonical correlation analysis scheme by assuming that the influence of the visual stimuli on brain activity data follows the Poisson distribution. The contribution of this paper is the construction of a recognition method with considering the time lag for getting truly close to the realization of the occurrence mechanism of human emotions. Experimental results show the effectiveness of considering the time lag between gaze and brain activity data.

Index Terms— Visual attention, brain activity, multi-modal human emotion recognition, sequential data, time lag.

1. INTRODUCTION

Human emotions are the essential but enigmatic nature, and clarification of their occurrence mechanism can contribute to several fields. For instance, multimedia content recommendation, which reflects personalized preferences based on the occurrence mechanism of human emotions, can be realized [1, 2]. In addition, in the field of human-computer interaction, as humans communicate with each other, agents in robots or computers can communicate affectively with humans by using such a mechanism [3, 4]. The attempt to implement this mechanism on computers is called affective computing [5]. Computers need to recognize human emotions in the first step of affective computing, but this is still a challenging task due to the subjectivity of human emotions.

In the field of signal processing, brain activities obtained from humans while viewing images/videos or listening to music have been analyzed for recognizing human emotions [6, 7]. Emotions obtained from such brain activity analysis, however, are not necessarily related to the target stimuli since the human brain processes huge amounts of information received from all over the body. On the other hand, various non-verbal cues such as eye gaze, facial expression, and so on have been studied as clues for recognizing human emotions [8]. Such non-verbal cues may reflect the subconscious response since they are governed by the sympathetic nervous system regardless of human intention. Therefore, the use of these cues besides brain activity enables to capture of more reliable emotion-related information. In fact, multi-modal human emotion recognition methods, which use multiple biological signals, have achieved higher accuracy than uni-modal methods [9–12]. Most of those multi-modal approaches adopt brain activity and eye gaze as the modalities for capturing implicit and explicit information, respectively. The previous methods collaboratively use multi-modal biological signals by integrating features calculated from each modality, and researchers have sought the more emotion-friendly feature integration methods. Specifically, in [9] and [10], bi-modal deep autoencoder (BDAE) [13] and deep canonical correlation analysis (DeepCCA) [14] are used for extracting the common factors of all features, respectively. Although biological signals are sequential data, these methods cannot consider the changes in biological signals with time, which are one of the most significant characteristics for human emotion recognition [15, 16]. Then, in [11, 12], bi-modal long-short term memory (BLSTM) [17] and gaze and image tensor (GIT) with CCA [18] are used for extracting the common factors with considering time changes. These methods have succeeded in introducing the mechanism for considering time changes by associating each modality at the same timestep.

In the case of the human emotion recognition for visual stimuli, humans acquire information through the eyes and process it in the brain. The visual stimuli captured by human eyes are sent to the brain by neurotransmitters. Thus, there is a time lag between gaze data, which record where humans gaze at, and brain activity data as shown in Fig. 1 [19]. However, the previous studies cannot consider such a time lag since they just integrate features obtained from gaze and brain activity data. Therefore, these studies integrate features not for the same visual stimuli but for different visual stimuli between gaze and brain activity data, and there is limitation to the expressive power of the integrated features for human emotions. Under these circumstances, the integration method that can consider the time lag between gaze and brain activity data is necessary for realizing the occurrence mechanism of human emotions. In this study, we treat the image as visual stimuli for the first attempt to consider the time lag between multiple biological signals. Moreover, the acquisition of

This work was partly supported by the KAKENHI Grant Number JP17H01744 and JP21H03456.

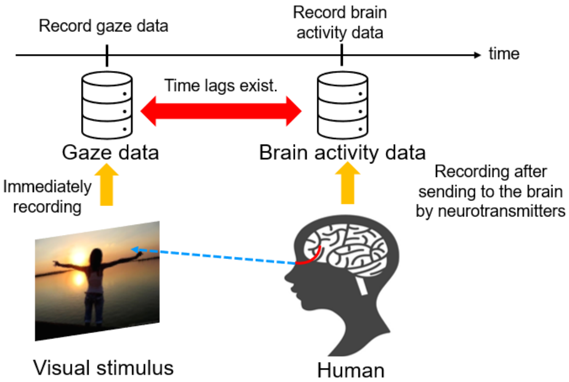

Fig. 1: Concept of the time lag between gaze and brain activity data. Since the visual stimuli captured by human eyes are sent to the brain by neurotransmitters, there is a time lag between gaze and brain activity data [19].

brain activity data is a heavy burden on humans in their daily lives, and we use brain activity data only in the training phase.

We propose a multi-modal human emotion recognition method based on the time lag-considered correlation maximization in this paper. The time lag between gaze and brain activity data depends on neuron's reaction time that may vary among humans, and the proposed method should efficiently deal with the time lag for modeling the neuron's reaction time. Then we focus on maximizing the weighted correlation for realizing the feature integration with considering the time lag. Concretely, we extract gaze and brain activity features and calculate transform vectors that project each feature into a common space. Then transform vectors are trained by maximizing correlations weighted in response to the time lags. For realizing the time lag-considered correlation maximization, we expand the CCA scheme, which is a simple linear transformation but can be extended with a structure that explicitly takes into account the time lag that is the linear shift between gaze and brain activity. In the CCA, the latent features are calculated with the optimization of the transform vectors by using the original correlation. We focus on the original correlation and introduce the weights by considering the time lag into this correlation. In the proposed method, we assume that the time lag follows a certain distribution, and the weights are generated from this distribution. Once we obtain the transform vectors from training data, the brain activity data, which are burdens to acquire, are not necessarily required in the inference phase. Finally, the proposed method recognizes human emotions using the features integrated by the transform vectors. The contribution of this paper is that we newly construct the human emotion recognition method with considering the time lag for getting truly close to the realization of the occurrence mechanism of human emotions.

2. HUMAN EMOTION RECOGNITION BASED ON TIME LAG-CONSIDERED CORRELATION MAXIMIZATION

The multi-modal human emotion recognition based on time lag-considered correlation maximization is explained in this section. First, we calculate sequential features from gaze and brain activity data. As the gaze features, we adopt the GIT-based method [12] for capturing visual information obtained by humans. Moreover, as brain activity data, we adopt functional near-infrared spectroscopy (fNIRS), which is well known for its higher temporal resolution than functional magnetic resonance imaging (fMRI). In particular, fNIRS data are considered to be more robust than the electroencephalogram (EEG) to the effects of external activities such as eye blinks that occur when viewing images [20]. Several researchers, therefore, have studied the relationship between human emotions and fNIRS data [21–23]. Thus, we use the gaze and fNIRS features as multi-modal features. Note that the proposed method uses the fNIRS features for only calculating the transform vectors in the training phase. Next, the gaze and fNIRS features are integrated by using our time lag-considered CCA (TlCCA), and we calculate the latent features including the common factor of two types of features. The effectiveness of correlation with a time lag in feature integration of multimedia data and tweets of Twitter has been verified [24].

2.1. Feature Extraction

In this subsection, we explain the way to extract gaze and fNIRS features. In the proposed method, we extract gaze features based on the GIT. Meanwhile, the fNIRS features are extracted based on the statistical and wavelet transform [25]-based features since the fNIRS data consist of multiple one-dimensional signals.

2.1.1. Gaze Features based on GIT

To capture temporal visual information obtained by humans, we adopt the GIT method and calculate sequential visual features from the GIT as gaze features. For representing the image $\boldsymbol{x} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ ($d_1 \times d_2$ and d_3 being the image size and the number of color channels, respectively) and the corresponding gaze data, simultaneously, we construct the GIT $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3 \times d_t}$ (d_t being the number of timesteps). Then gaze data include the gaze location corresponding to the image while viewing the image. Concretely, we calculate the gaze weight $\boldsymbol{W}_t^{\mathrm{gaze}} \in \mathbb{R}^{d_1 \times d_2}$ at timestep $t (= 1, 2, \ldots, d_t)$ by applying the Gaussian blur into the gray-scale image representing gazed location. In order to represent the gaze weight $\boldsymbol{W}_t^{\mathrm{gaze}}$ and the image $\boldsymbol{x}$ with the time changes, we calculate a gaze and image weight $\boldsymbol{W}_t^{\mathrm{GIW}} \in \mathbb{R}^{d_1 \times d_2}$ at timestep t as follows:

$$\boldsymbol{W}_t^{\mathrm{GIW}} = d_t \frac{\boldsymbol{W}_t^{\mathrm{gaze}}}{\sum_{t=1}^{d_t} \boldsymbol{W}_t^{\mathrm{gaze}}} + \boldsymbol{E}, \tag{1}$$

where $\boldsymbol{E} \in \mathbb{R}^{d_1 \times d_2}$ is the matrix whose elements are all one for representing the pixel values that are not gazed in the GIT. Without the introduction of $\boldsymbol{E}$, the GIT loses information on images that the target human does not gaze at. Next, the GIT $\mathcal{X}$ is constructed as follows:

$$\mathcal{X}_{ch,t} = \boldsymbol{x}_{ch} \circ \boldsymbol{W}_t^{\mathrm{GIW}}, \tag{2}$$

where $\mathcal{X}_{ch,t} \in \mathbb{R}^{d_1 \times d_2}$ and $\boldsymbol{x}_{ch} \in \mathbb{R}^{d_1 \times d_2}$ ($ch = 1, 2, \cdots, d_3$) are gray-scale images obtained from each channel of the GIT $\mathcal{X}$ and the image $\boldsymbol{x}$, respectively. It should be noted that "$\circ$" is the Hadamard product operator. In this way, we construct the GIT for obtaining the sequential data including the image and gaze data.

As the gaze features, we extract visual features from the GIT $\mathcal{X}_t$ at each timestep. In order to obtain high semantic gaze features, we use the convolutional neural network (CNN). Concretely, we pretrain the Xception model [26] with the ImageNet [27] as the transfer learning scheme [28]. Moreover, the outputs of the last convolution layer with the average pooling of the Xception model are used as d_{gaze} dimensional gaze features $\boldsymbol{y}_{\mathrm{gaze},t}$ at timestep t.

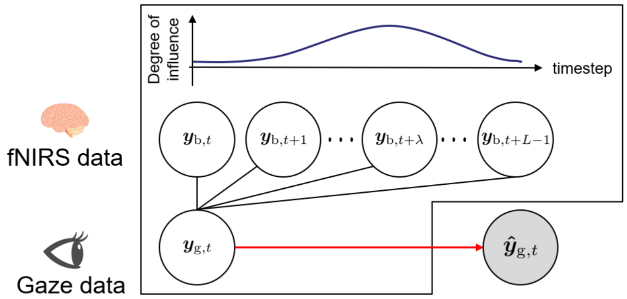
Degree of influence
timestep
fNIRS data
$\boldsymbol{y}_{\mathrm{b},t}$
$\boldsymbol{y}_{\mathrm{b},t+1}$
$\boldsymbol{y}_{\mathrm{b},t+\lambda}$
$\boldsymbol{y}_{\mathrm{b},t+L-1}$
Gaze data
$\boldsymbol{y}_{\mathrm{g},t}$
$\hat{\boldsymbol{y}}_{\mathrm{g},t}$

Fig. 2: Overview of TlCCA. We introduce weights considering the time lag into the original correlation. In this figure, $\boldsymbol{y}_{\mathrm{g},t}$ and $\boldsymbol{y}_{\mathrm{b},t}$ mean $\boldsymbol{y}_{\mathrm{gaze},t}$ and $\boldsymbol{y}_{\mathrm{brain},t}$, respectively. Moreover, white and gray circles are observed and unobserved variables. λ and L mean the peak and range of the time lag.

2.1.2. *fNIRS Features*

As the fNIRS features, we extract two types of features that are statistical and wavelet transform-based features by referring to [29]. When recording fNIRS data, we use several channels and extract the following features for each channel. As the statistical features, we calculate average, variance, skewness, kurtosis, zero-crossing-rate and root-mean-square of fNIRS data at each timestep. As the wavelet transform-based features, we calculate the proportion of the energy to the total energy in the high- or low- frequency domain at each timestep. Finally, we calculate d_{brain} dimensional fNIRS features $\boldsymbol{y}_{\mathrm{brain},t}$ by concatenating these features.

2.2. Human Emotion Recognition based on TlCCA

The human emotion recognition based on TlCCA is described in this subsection. The overview of the TlCCA is shown in Fig. 2. We define gaze and brain features as follows:

$$\boldsymbol{Y}_p = [\boldsymbol{y}_{p,1}, \boldsymbol{y}_{p,2}, \ldots, \boldsymbol{y}_{p,d_t}], \;\; p = \{\mathrm{gaze}, \mathrm{brain}\}. \qquad (3)$$

We assume that the fNIRS data are recorded a few seconds after gaze data are recorded for one stimulus, that is, there is the time lag of fNIRS data with respect to the gaze data. In addition to the time lag, the influence of visual stimuli can continue to the fNIRS data at the next timestep. Under these circumstances, we assume that the visual stimuli obtained by humans are immediately recorded in the gaze data, while the influence of the visual stimuli on fNIRS data follows the Poisson distribution. The TlCCA enables the calculation of the latent features from gaze and brain activity features with considering the assumptions of the time lag. Concretely, the transform vector set $\boldsymbol{w} = \{\boldsymbol{w}_{\mathrm{gaze}}, \boldsymbol{w}_{\mathrm{brain}}\} \in (\mathbb{R}^{d_{\mathrm{gaze}}}, \mathbb{R}^{d_{\mathrm{brain}}})$ is optimized with training feature sets $\{\boldsymbol{Y}_{\mathrm{gaze},n}, \boldsymbol{Y}_{\mathrm{brain},n}\}_{n=1}^{N}$ ($n = 1, 2, \ldots, N$; N being the number of training data) as follows:

$$\begin{aligned} \hat{\boldsymbol{w}} &= \arg\max_{\boldsymbol{w}} \boldsymbol{w}_{\mathrm{gaze}}^{\top} \sum_{n=1}^{N} \boldsymbol{C}_n^{\mathrm{b}} \boldsymbol{w}_{\mathrm{brain}} \\ \text{s.t. } & \boldsymbol{w}_{\mathrm{gaze}}^{\top} \boldsymbol{C}_n^{\mathrm{gaze}} \boldsymbol{w}_{\mathrm{gaze}} = \boldsymbol{w}_{\mathrm{brain}}^{\top} \boldsymbol{C}_n^{\mathrm{brain}} \boldsymbol{w}_{\mathrm{brain}} = 1, \forall n, \qquad (4) \end{aligned}$$

where $\boldsymbol{C}_n^{\mathrm{gaze}}$ and $\boldsymbol{C}_n^{\mathrm{brain}}$ are the variance matrices of gaze and fNIRS features, respectively. These variance matrices are calculated as follows:

$$\boldsymbol{C}_n^{p} = \boldsymbol{Y}_{p,n} \boldsymbol{Y}_{p,n}^{\top}, \;\; p = \{\mathrm{gaze}, \mathrm{brain}\}. \qquad (5)$$

Moreover, $\boldsymbol{C}_n^{\mathrm{b}}$ is the covariance matrix considering the time lag between gaze and fNIRS features as follows:

$$\boldsymbol{C}_n^{\mathrm{b}} = \frac{1}{\sum_{l=0}^{L} e^{-\lambda} \lambda^{l} / l!} \sum_{l=0}^{L} \frac{e^{-\lambda} \lambda^{l}}{l!} \boldsymbol{Y}_{\mathrm{gaze},n,l} \boldsymbol{Y}_{\mathrm{brain},n,0}^{\top}, \qquad (6)$$

where λ is a shape parameter of the Poisson distribution and determines the strongest point of visual stimuli that affect the fNIRS features. L is a hyperparameter and decides the number of timesteps that influence visual stimuli on fNIRS features. It should be noted that the features are mean-normalized, and we prepare the new feature set $\boldsymbol{Y}_{p,n,l} = [\boldsymbol{y}_{p,n,L-l}, \boldsymbol{y}_{p,n,L+1-l}, \ldots, \boldsymbol{y}_{p,n,d_t-l}]$ ($l = 0, 1, \ldots, L-1$). In order to solve Eq. (4), we use the method of Lagrange multiplier as follows:

$$F = \boldsymbol{w}_{\mathrm{gaze}}^{\top} \sum_{n=1}^{N} \boldsymbol{C}_n^{\mathrm{b}} \boldsymbol{w}_{\mathrm{brain}} - \sum_{p} \beta_p \left(\sum_{n=1}^{N} \boldsymbol{w}_p^{\top} \boldsymbol{C}_n^{p} \boldsymbol{w}_p - 1 \right), \qquad (7)$$

where $p = \{\mathrm{gaze}, \mathrm{brain}\}$, β_{gaze} and β_{brain} are the Lagrange coefficients. By calculating $\partial F / \partial \boldsymbol{w}_{\mathrm{gaze}} = 0$ and $\partial F / \partial \boldsymbol{w}_{\mathrm{brain}} = 0$, we obtain the eigenvalue problem as follows:

$$\begin{aligned} &\begin{bmatrix} \boldsymbol{O} & \sum_{n=1}^{N} \boldsymbol{C}_n^{\mathrm{b}\,\top} \\ \sum_{n=1}^{N} \boldsymbol{C}_n^{\mathrm{b}} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_{\mathrm{gaze}} \\ \boldsymbol{w}_{\mathrm{brain}} \end{bmatrix} = \\ &\qquad \beta \begin{bmatrix} \sum_{n=1}^{N} \boldsymbol{C}_n^{\mathrm{gaze}} & \boldsymbol{O} \\ \boldsymbol{O} & \sum_{n=1}^{N} \boldsymbol{C}_n^{\mathrm{brain}} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_{\mathrm{gaze}} \\ \boldsymbol{w}_{\mathrm{brain}} \end{bmatrix}, \qquad (8) \end{aligned}$$

where $\beta = 2\beta_{\mathrm{gaze}} = 2\beta_{\mathrm{brain}}$. By simply solving Eq. (8), the optimal transform vector pair $(\hat{\boldsymbol{w}}_{\mathrm{brain}}, \hat{\boldsymbol{w}}_{\mathrm{brain}})$ is calculated. It should be noted that although some solution sets $(\beta, \hat{\boldsymbol{w}}_{\mathrm{brain}}, \hat{\boldsymbol{w}}_{\mathrm{brain}})$ become candidates, β represents the efficiency of transform vectors $(\hat{\boldsymbol{w}}_{\mathrm{brain}}, \hat{\boldsymbol{w}}_{\mathrm{brain}})$. Thus, we sort eigenvalues β, and then adopt the top d_{latent} $(\leq \min(d_{\mathrm{gaze}}, d_{\mathrm{brain}}))$ eigenvalues and their corresponding transform vectors. The transform matrices $\hat{\boldsymbol{W}}_{\mathrm{gaze}} \in \mathbb{R}^{d_{\mathrm{gaze}} \times d_{\mathrm{latent}}}$ and $\hat{\boldsymbol{W}}_{\mathrm{brain}} \in \mathbb{R}^{d_{\mathrm{brain}} \times d_{\mathrm{latent}}}$ are used for the actual transformation.

In the inference phase that we only use the gaze data, the transformed features $\hat{\boldsymbol{Y}}_{\mathrm{gaze}} \in \mathbb{R}^{d_{\mathrm{latent}} \times d_t}$ are calculated with the transform matrix $\hat{\boldsymbol{W}}_{\mathrm{gaze}}$ as $\hat{\boldsymbol{Y}}_{\mathrm{gaze}} = \hat{\boldsymbol{W}}_{\mathrm{gaze}}^{\top} \boldsymbol{Y}_{\mathrm{gaze}}$. In this way, we can obtain the transformed features with considering the time lag as shown in Eq. (6). Finally, the proposed method recognizes human emotions by obtaining d_{emotion} dimensional one-hot vectors $\boldsymbol{e} \in \mathbb{R}^{d_{\mathrm{emotion}}}$ whose elements correspond to human emotions as $\boldsymbol{e} = \boldsymbol{f}(\hat{\boldsymbol{Y}}_{\mathrm{gaze}})$, where $\boldsymbol{f}(\cdot)$ is the classifier trained by using the transformed features obtained from the training gaze data.

3. EXPERIMENTS

In this experiment, Tobii eye tracker 4C[1] and LIGHTNIRS[2] were used for obtaining gaze and fNIRS data. The eye tracker was installed at a 15-inch display with about 70 cm distance from the participants. When recording fNIRS data, we used the head cap with 10 channels on the front of the head and 10 channels on the back of the head. In the fNIRS data, the changes in both oxy/deoxygenated hemoglobin levels were recorded. Furthermore, 80 images from the art photo dataset [30] were used, and we randomly selected 64 images and the remaining images as test images.

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

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

Table 1: Numbers of images in each emotion.

	Par1	Par2	Par3	Par4	Par5	Par6	Par7	Par8	Par9	Par10
Number of positive images	37	35	38	36	45	30	39	42	35	46
Number of negative images	43	46	42	44	35	50	41	38	45	34

Table 2: Properties of our method and comparative methods.

	Features		Time	
	Gaze	fNIRS	Change	Lag
Abbreviation 1		✓	✓	
Abbreviation 2	✓		✓	
Deep CCA [10]	✓	✓		
BDAE [9]	✓	✓		
BLSTM [11]	✓	✓	✓	
MVAE [32]	✓	✓	✓	
CCA with GIT [12]	✓	✓	✓	
Our method	✓	✓	✓	✓

Table 3: Average results of each method.

	Recall	Precision	F1-score	Accuracy
Abbreviation 1	0.30	0.48	0.65	0.52
Abbreviation 2	**0.83**	0.74	0.76	0.77
Deep CCA [10]	0.57	0.54	0.53	0.58
BDAE [9]	0.29	0.64	0.55	0.57
BLSTM [11]	0.37	0.31	0.44	0.44
MVAE [32]	0.49	0.55	0.52	0.57
CCA with GIT [12]	0.63	**0.85**	0.67	0.74
Our method	0.75	0.84	**0.78**	**0.81**

There were 10 participants (Pars. 1-10) including seven men and three women, in this experiment. We instructed participants to gaze at each image for ten seconds with a ten-second interval for preventing fNIRS data from being influenced by the previous image. In the interval, an image with a cross mark in the center for leading the gaze to the center of the monitor was shown. After the task, participants answered the questionnaire for giving feedbacks about their emotion (positive/negative, that is $d_{\mathrm{emotion}} = 2$) induced by viewing the images. Table 1 shows the number of images in each emotion for each participant.

As comparative methods, we used the following seven methods. As shown in Table 2, two methods that adopted gaze or fNRIS features were used as the abbreviation studies (Hereafter, Abbreviations 1 and 2). That is, they are uni-modal methods. Note that these two methods applied the principal component analysis [31] in order to reduce the dimensions of features to d_{latent}. The rest five methods were other human emotion recognition methods proposed in [9–12, 32] that integrated multi-modal features by using Deep CCA [14], BDAE [13], BLSTM [17], CCA with GIT [12] and multi-view variational autoencoder (MVAE) [33], respectively. In [9, 10], there was no mechanism to consider not only the time lag but also the time changes, and we calculated each feature when $d_t = 1$ in Sec. 2.1. Moreover, in the inference phase of [9–11, 32], both gaze and brain activity features were necessarily required, while in that of [12] and the proposed method, only gaze data were required. It should be noted that SVM was used in each method as the final classifier $\boldsymbol{f}(\cdot)$. The hyperparameters L, λ, d_{brain}, d_{gaze} and d_{latent} were set to 5, 1 440, 2048 and 50, respectively. For evaluating each method, we used Recall, Precision, F1-score and Accuracy.

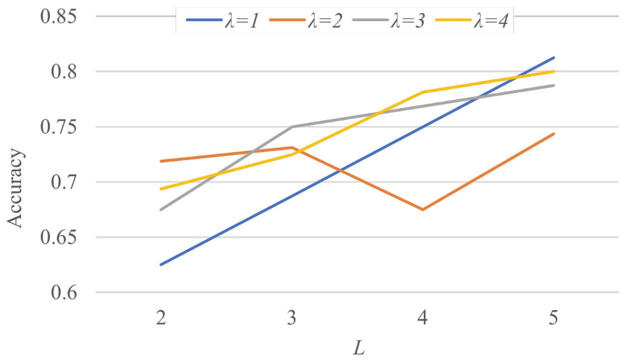

Fig. 3: Changes in average Accuracy of the proposed method with respect to λ and L.

Table 3 shows the average results of each method. By comparing our method with Abbreviations 1 and 2, we confirm that the collaborative use of gaze and fNIRS features is effective for human emotion recognition. In the evaluation index "Recall", Abbreviation 2 outperforms other methods, but the proposed method outperforms other methods in other evaluation indices. Then Precision of Abbreviation 2 is lower than that of the proposed method, and the recognition ability of the proposed method is higher since the proposed method achieves the higher F1-score, which is the harmonic mean of Recall and Precision, than that of Abbreviation 2. Moreover, by comparing our method with [9, 10], the mechanism to consider the time changes is confirmed to be effective. By comparing our method with [12], in the evaluation index "Precision", CCA with GIT outperforms other methods, but the effectiveness of our method is insisted in the same discussion as Recall of Abbreviation 1. Finally, by comparing our method with [32], the effectiveness of considering the time lag between gaze and fNIRS data, which is the main focus in this paper, is verified. In the evaluation index "Recall", Abbreviation 2 outperforms other methods, but the proposed method outperforms other methods in other evaluation indices.

Figure 3 shows the changes in average Accuracy of the proposed method with respect to λ and L. In this figure, for any λ, the accuracy is best at $L = 5$. Specifically, it is the highest value when $\lambda = 1$ that means that the peak of the time lag is one timestep corresponding to one second. Our results can be close to the results reported in previous studies [19, 34] in the field of brain computing. Therefore, the human cognition process can be well represented using the time lag in the proposed method.

4. CONCLUSIONS

We have proposed the multi-modal human emotion recognition method based on TlCCA for considering the time lag between gaze and fNRIS data. Concretely, we introduce the mechanism to consider the time lag into CCA scheme by assuming that the influence of the visual stimuli on fNIRS data follows the Poisson distribution. By using TlCCA, we newly construct the human emotion recognition method with considering the time lag.

5. REFERENCES

[1] Deger Ayata, Yusuf Yaslan, and Mustafa E Kamasak, "Emotion based music recommendation system using wearable physiological sensors," *IEEE Trans. Consumer Electronics*, vol. 64, no. 2, pp. 196–203, 2018.

[2] Ryosuke Sawata, Takahiro Ogawa, and Miki Haseyama, "Human-centered favorite music classification using eeg-based individual music preference via deep time-series cca," in *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1320–1324.

[3] Suhaila Najim Mohammed and Alia Karim Abdul Hassan, "A survey on emotion recognition for human robot interaction," *Journal of Computing and Information Technology*, vol. 28, no. 2, pp. 125–146, 2020.

[4] Satyajit Nayak, Bingi Nagesh, Aurobinda Routray, and Monalisa Sarma, "A human–computer interaction framework for emotion recognition through time-series thermal video sequences," *Computers & Electrical Engineering*, vol. 93, pp. 107280, 2021.

[5] Rosalind W Picard, "Affective computing," *The MIT Press Cambridge*, vol. 167, pp. 170, 1997.

[6] Kento Sugata, Takahiro Ogawa, and Miki Haseyama, "Emotion estimation via tensor-based supervised decision-level fusion from multiple brodmann areas," in *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 999–1003.

[7] Ningjie Liu, Yuchun Fang, Ling Li, Limin Hou, Fenglei Yang, and Yike Guo, "Multiple feature fusion for automatic emotion recognition using eeg signals," in *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 896–900.

[8] C Vinola and K Vimaladevi, "A survey on human emotion recognition approaches, databases and applications," *Electronic Letters on Computer Vision and Image Analysis (ELCVIA)*, pp. 00024–44, 2015.

[9] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu, "Emotion recognition using multimodal deep learning," in *Proc. Int'l Conf. Neural Information Processing (ICONIP)*, 2016, pp. 521–529.

[10] Jie Qiu, Wei Liu, and Bao Lu, "Multi-view emotion recognition using deep canonical correlation analysis," in *Proc. Int'l Conf. Neural Information Processing (ICONIP)*, 2018, pp. 221–231.

[11] Hao Tang, Wei Liu, Wei-Long Zheng, and Bao-Liang Lu, "Multimodal emotion recognition using deep neural networks," in *Proc. Int'l Conf. Neural Information Processing (ICONIP)*, 2017, pp. 811–819.

[12] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Human-centric emotion estimation based on correlation maximization considering changes with time in visual attention and brain activity," *IEEE Access*, vol. 8, pp. 203358–203368, 2020.

[13] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *Proc. Int'l Conf. Machine Learning (ICML)*, 2011.

[14] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *Proc. Int'l Conf. Machine Learning (ICML)*, 2013, pp. 1247–1255.

[15] Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu, "Emotional state classification from eeg data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, 2014.

[16] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan Koenig, Juan Xu, Mohan Kankanhalli, and Qi Zhao, "Emotional attention: A study of image sentiment and visual attention," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7521–7531.

[17] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3, pp. 321–377, 1936.

[19] Shiori Amemiya, Hidemasa Takao, and Osamu Abe, "Origin of the time lag phenomenon and the global signal in resting-state fMRI," *Frontiers in Neuroscience*, vol. 14, pp. 1141, 2020.

[20] Audrey Girouard, Erin Treacy Solovey, Leanne M Hirshfield, Evan M Peck, Krysta Chauncey, Angelo Sassaroli, Sergio Fantini, and Robert JK Jacob, "From brain signals to adaptive interfaces: using fnirs in hci," *Brain-Computer Interfaces*, pp. 221–237, 2010.

[21] Dominic Heger, Reinhard Mutter, Christian Herff, Felix Putze, and Tanja Schultz, "Continuous recognition of affective states by functional near infrared spectroscopy signals," in *Proc. IEEE Humaine Association Conf. Affective Computing and Intelligent Interaction*, 2013, pp. 832–837.

[22] Danushka Bandara, Senem Velipasalar, Sarah Bratt, and Leanne Hirshfield, "Building predictive models of emotion with functional near-infrared spectroscopy," *Int'l Journal of Human-Computer Studies*, vol. 110, pp. 75–85, 2018.

[23] Thibaud Gruber, Coralie Debracque, Leonardo Ceravolo, Kinga Igloi, Blanca Marin Bosch, Sascha Frühholz, and Didier Grandjean, "Human discrimination and categorization of emotions in voices: a functional near-infrared spectroscopy (fnirs) study," *Frontiers in Neuroscience*, vol. 14, pp. 570, 2020.

[24] Kaito Hirasawa, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Mvgan maximizing time-lag aware canonical correlation for baseball highlight generation," in *Proc. IEEE Int'l Conf. Multimedia & Expo Workshops (ICMEW)*, 2020, pp. 1–6.

[25] Mark Shensa, "The discrete wavelet transform: wedding the a trous and mallat algorithms," *IEEE Trans. Signal Processing (TSP)*, vol. 40, no. 10, pp. 2464–2482, 1992.

[26] François Chollet, "Xception: Deep learning with depthwise separable convolutions," *arXiv preprint arXiv: 1610.02357*, 2017.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[28] Sinno Jialin Pan, Qiang Yang, et al., "A survey on transfer learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[29] Kelly Tai and Tom Chau, "Single-trial classification of nirs signals during emotional induction tasks: towards a corporeal machine interface," *Journal of Neuroengineering and Rehabilitation*, vol. 6, no. 1, pp. 39, 2009.

[30] Jana Machajdik and Allan Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. ACM Int'l Conf. Multimedia (ACMMM)*, 2010, pp. 83–92.

[31] Karl Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[32] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Human emotion estimation using multi-modal variational autoencoder with time changes," in *Proc. IEEE Global Conf. Life Sciences and Technologies (LifeTech)*, 2021, pp. 67–68.

[33] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo, "Joint multimodal learning with deep generative models," *arXiv preprint arXiv:1611.01891*, 2016.

[34] Parthana Sarma and Shovan Barma, "Review on stimuli presentation for affect analysis based on eeg," *IEEE Access*, vol. 8, pp. 51991–52009, 2020.