# SEMI-SUPERVISED MULTIPLE FEATURE FUSION FOR VIDEO PREFERENCE ESTIMATION

*Akira Toyoda, Takahiro Ogawa and Miki Haseyama*

Graduate School of Information Science and Technology, Hokkaido University
N-14, W-9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan

## ABSTRACT

This paper presents a new method to estimate user preferences for videos based on multiple feature fusion via semi-supervised Multiview Local Fisher Discriminant Analysis (sMvLFDA). The proposed method first extracts multiple visual features from videos and functional near-infrared spectroscopy (fNIRS) features from fNIRS signals recorded during watching videos. Next, we apply Locality Preserving Canonical Correlation Analysis (LPCCA) to each visual feature and fNIRS features and project each visual feature to the new feature spaces (fNIRS-based visual feature spaces). Consequently, since the correlation between each visual feature and fNIRS features which reflect user preferences is maximized, we can transform visual features into features which also reflect user preferences. In addition, we newly introduce sMvLFDA and fuse multiple fNIRS-based visual features via sMvLFDA. sMvLFDA fuses features while using labeled samples and unlabeled samples simultaneously to reduce overfitting to the labeled samples. Furthermore, sMvLFDA adequately uses complementary properties in multiple features. Therefore, it can be expected that the fused features are more effective for estimation of user preferences than each fNIRS-based visual feature. The main contribution of this paper is the new derivation of sMvLFDA. Consequently, by using the fused features, it becomes feasible to estimate user preferences for videos successfully.

***Index Terms*—** Personal preference, functional near-infrared spectroscopy (fNIRS), canonical correlation analysis, Fisher discriminant analysis.

## 1. INTRODUCTION

With increasing the number of videos uploaded on the Web, it has become important to recommend videos and help users to find their favorite videos. To realize such recommendations, it is necessary to accurately estimate user preferences for videos [1]. Some methods have analyzed brain signals which represent the degree of brain activity to estimate user emotions and preferences [2, 3, 4]. From these methods, we consider that it is effective for preference estimation to introduce features extracted from brain signals. In addition, many methods have focused on using functional near-infrared spectroscopy (fNIRS) and measuring the degree of brain activity as fNIRS signals [5]. fNIRS detects hemoglobin (HbO2) and deoxygenated hemoglobin (HbR) changes, which are functional indicator of brain activity in regional cerebral blood flow, by using light in the near-infrared range (700-900nm). As in the case which uses brain signals measured with other methods, some studies have shown effectiveness of using fNIRS signals to estimate user emotions and preferences [6, 7].

In addition, for improvement of preference estimation, we previously proposed to collaboratively use features extracted from videos and fNIRS signals [8]. In this method, by using Locality Preserving Canonical Correlation Analysis (LPCCA) [9], we first project visual features extracted from videos into the new feature spaces and maximize the correlation between fNIRS features and visual features. Consequently, since the correlation between fNIRS features reflecting user preferences and visual features is maximized, the new features (fNIRS-based visual features) also reflect user preferences. Furthermore, by applying this projection to multiple visual features, multiple fNIRS-based visual features are computed, and these computed features are fused via Multiview Local Fisher Discriminant Analysis (MvLFDA) [10]. MvLFDA can fuse features and optimize the class separation of the fused feature space while considering different contributions and adequately using complementary properties in multiple features.

In our previous method, MvLFDA is the supervised feature fusion method. However, on the Web, there are very few videos which users rate [11, 12]. In addition, when only a small number of labeled samples are available, supervised methods tend to find new feature spaces which are overfitted to the labeled samples [13]. In order to reduce the influence of this problem, semi-supervised methods have been proposed [13, 14]. These methods use unlabeled samples together with labeled samples. In particular, Song *et.al.* proposed to optimize class separation in a feature space using labeled samples and simultaneously preserve the local structure of the whole feature space using both labeled and unlabeled samples. Consequently, this method realized the reduction of overfitting to the labeled samples. Thus, by introducing the framework to use unlabeled samples together with labeled samples and preserve the local structure of the whole feature space to MvLFDA, it becomes possible to extract features effective for preference estimation when it is difficult to obtain a number of labeled samples.

This paper presents a novel method to estimate user preferences for videos via multiple feature fusion. First, we extract fNIRS features from fNIRS signals and multiple visual features from videos. Next, we project visual features into the space of the fNIRS-based visual features via LPCCA. In the fNIRS-based feature space, the correlation between fNIRS features reflecting user preferences and each visual feature is maximized. In addition, to fuse multiple fNIRS-based visual features, we newly derive a novel method, *i.e.*, semi-supervised MvLFDA (sMvLFDA). sMvLFDA can compute the fused features while optimizing class separation in the fused feature space by using labeled samples and preserving the local structure of the feature space by using both labeled and unlabeled samples. The derivation of sMvLFDA is the main contribution of this paper. Consequently, our method realizes successful estimation of user preferences based on the above non-conventional approach.

This work was partly supported by JSPS KAKENHI Grant Numbers JP17H01744, JP15K12023.

## 2. MULTIPLE FEATURE FUSION FOR VIDEO PREFERENCE ESTIMATION

Our method consists of three stages. In the first stage, we extract fNIRS features from fNIRS signals recorded while users are watching videos and multiple visual features from videos. In the second stage, we transform each visual feature into the fNIRS-based visual features reflecting user preferences by projecting each visual feature to the new feature spaces and maximizing the correlation between fNIRS features and each visual feature via LPCCA. In the third stage, we fuse multiple fNIRS-based visual features via sMvLFDA. The estimation of user preferences then becomes feasible by using the fused features. The details of each stage are shown below.

### 2.1. Feature Extraction

This subsection shows the method to extract fNIRS features and visual features used in our method.

#### *2.1.1. fNIRS Feature Extraction*

First, fNIRS signals are divided into overlapping segments. Next, from signals in each segment, the following two types of fNIRS features are extracted based on [6]:

**Time-domain features**: Mean, variance, zero crossings, root mean squared, skewness, and kurtosis.

**Time-frequency features**: The wavelet coefficients are computed by applying wavelet decomposition to fNIRS signals. Next, from these coefficients, the energy values of some frequency bands are computed. Finally, the relative energies between the energy value of each frequency band and the total energy value are computed.

Consequently, the means and standard deviations of each feature are computed over all segments and used as fNIRS features. In addition, by using Structural Feature Selection with Sparsity [15], we select only fNIRS features related to preference estimation. Finally, by using the selected features, we define a fNIRS feature vector $\boldsymbol{x}_f \in \mathbb{R}^{d_{x_f}}$, where $d_{x_f}$ denotes the number of selected features.

#### *2.1.2. Visual Feature Extraction*

First, the following visual features are extracted from each frame of videos:

**CNN features**: A 4096 dimensional output from the fc6 layer of AlexNet [16] pre-trained on ImageNet database [17].

**Hand-crafted features**: HSV histograms (64 dimension) and Scale Invariant Feature Transform (100 dimension) [18].

In addition, the mean vector of each visual feature over all frames of the target video is computed and used as each visual feature for each video. Finally, we define $p$ th ($p = 1,\ 2,\ \ldots\ , P$) visual feature vector $\boldsymbol{x}_v^{(p)} \in \mathbb{R}^{d_{x_v^{(p)}}}$, where $d_{x_v^{(p)}}$ denotes the dimension of $p$ th feature vector, and $P$ is the kind of visual features, *i.e.*, $P = 3$.

### 2.2. Feature Transformation via LPCCA

In this subsection, we explain the method to transform visual features into the fNIRS-based visual features via LPCCA. From fNIRS features and visual features, we first define $\boldsymbol{X}_F = [\boldsymbol{x}_{f,1}\boldsymbol{x}_{f,2}\cdots\boldsymbol{x}_{f,n}]$ and $\boldsymbol{X}_V^{(p)} = [\boldsymbol{x}_{v,1}^{(p)}\boldsymbol{x}_{v,2}^{(p)}\cdots\boldsymbol{x}_{v,n}^{(p)}]$, where $n$ is the number of samples. Note that we assume that there exist labeled samples and unlabeled samples in these samples. Next, from the two features, we construct similarity matrices $\boldsymbol{A}_{X_F}$ and $\boldsymbol{A}_{X_V}^{(p)}$ based on [19], which realized similarity graph construction which is robust to data noise and parameter-free by solving a $\ell_1$ norm optimization problem. Specifically, we first solve the following optimization problem with respect to a coefficient vector $\boldsymbol{a}_{x_f,i} \in \mathbb{R}^{n+d_{x_f}-1}$:

$$\arg\min_{\boldsymbol{a}_{x_f,i}} ||\boldsymbol{a}_{x_f,i}||_1, \quad \text{s.t. } \boldsymbol{x}_{f,i} = \boldsymbol{B}_{x_f,i}\boldsymbol{a}_{x_f,i}. \tag{1}$$

In Eq. (1), $\boldsymbol{B}_{x_f,i} = [\boldsymbol{x}_{f,1}\ \boldsymbol{x}_{f,2}\ \cdots\ \boldsymbol{x}_{f,i-1}\ \boldsymbol{x}_{f,i+1}\ \cdots\ \boldsymbol{x}_{f,n}\ \boldsymbol{I}_{d_{x_f}}]$, where $\boldsymbol{I}_{d_{x_f}} \in \mathbb{R}^{d_{x_f}\times d_{x_f}}$ is the identity matrix. From the obtained coefficient vector $\boldsymbol{a}_{x_f,i}$, we define the $(i,j)$ th ($j = 1,\ 2,\ \ldots\ , n$) element of $\boldsymbol{A}_{X_F}$ as follows:

$$(\boldsymbol{A}_{X_F})_{i,j} = \begin{cases} |(\boldsymbol{a}_{x_f,i})_j| & i > j \\ 0 & i = j \\ |(\boldsymbol{a}_{x_f,i})_{j-1}| & i < j \end{cases}, \tag{2}$$

where $(\cdot)_{i,j}$ denotes the $(i,j)$ th element of the matrix, and $(\cdot)_j$ denotes the $j$ th element of the vector. Finally, we redefine the similarity matrix $\boldsymbol{A}_{X_F}$ as follows: $\boldsymbol{A}_{X_F} = \frac{(\boldsymbol{A}_{X_F} + \boldsymbol{A}_{X_F}^{\mathrm{T}})}{2}$. The similarity matrix $\boldsymbol{A}_{X_V}^{(p)}$ is computed in the same manner as the computation of $\boldsymbol{A}_{X_F}$. From the similarity matrices $\boldsymbol{A}_{X_F}$ and $\boldsymbol{A}_{X_V}^{(p)}$, we compute the following matrices:

$$\boldsymbol{A}_{X_F X_V}^{(p)} = \boldsymbol{A}_{X_F} \circ \boldsymbol{A}_{X_V}^{(p)}, \tag{3}$$

$$\boldsymbol{A}_{X_F X_F} = \boldsymbol{A}_{X_F} \circ \boldsymbol{A}_{X_F}, \tag{4}$$

$$\boldsymbol{A}_{X_V X_V}^{(p)} = \boldsymbol{A}_{X_V}^{(p)} \circ \boldsymbol{A}_{X_V}^{(p)}, \tag{5}$$

where "$\circ$" denotes the Hadamard product. Consequently, we compute the projection matrix $\boldsymbol{U}_{cca}^{(p)}$, which projects each visual feature into the new feature space, by solving the following optimization problem:

$$\arg\max_{\boldsymbol{u}_{x_f},\boldsymbol{u}_{x_v}^{(p)}} \frac{\boldsymbol{u}_{x_f}^{\mathrm{T}} \boldsymbol{L}_{X_F X_V}^{(p)} \boldsymbol{u}_{x_v}^{(p)}}{\sqrt{\boldsymbol{u}_{x_f}^{\mathrm{T}} \boldsymbol{L}_{X_F X_F} \boldsymbol{u}_{x_f}} \sqrt{\boldsymbol{u}_{x_v}^{(p)\mathrm{T}} \boldsymbol{L}_{X_V X_V}^{(p)} \boldsymbol{u}_{x_v}^{(p)}}}, \tag{6}$$

where $\boldsymbol{L}_{X_F X_V}^{(p)}$, $\boldsymbol{L}_{X_F X_F}$ and $\boldsymbol{L}_{X_V X_V}^{(p)}$ are Laplacian matrices derived from $\boldsymbol{A}_{X_F X_V}^{(p)}$, $\boldsymbol{A}_{X_F X_F}$ and $\boldsymbol{A}_{X_V X_V}^{(p)}$ as follows:

$$\boldsymbol{L}_{X_F X_V}^{(p)} = \boldsymbol{D}_{X_F X_V}^{(p)} - \boldsymbol{A}_{X_F X_V}^{(p)}, \tag{7}$$

$$\boldsymbol{L}_{X_F X_F} = \boldsymbol{D}_{X_F X_F} - \boldsymbol{A}_{X_F X_F}, \tag{8}$$

$$\boldsymbol{L}_{X_V X_V}^{(p)} = \boldsymbol{D}_{X_V X_V}^{(p)} - \boldsymbol{A}_{X_V X_V}^{(p)}, \tag{9}$$

where $\boldsymbol{D}_{X_F X_V}^{(p)}$ is a diagonal matrix, and its entries are column sum of $\boldsymbol{A}_{X_F X_V}^{(p)}$. In addition, $\boldsymbol{D}_{X_F X_F}$ and $\boldsymbol{D}_{X_V X_V}^{(p)}$ are defined in the same manner. By using the Lagrange multiplier method, Eq. (6) is converted to the generalized eigen value problem. Thus, by solving the generalized eigen value problem, the projection matrix $\boldsymbol{U}_{cca}^{(p)}$ is computed as follows:

$$\boldsymbol{U}_{cca}^{(p)} = [\hat{\boldsymbol{u}}_{x_v,1}^{(p)}\ \hat{\boldsymbol{u}}_{x_v,2}^{(p)}\ \cdots\ \hat{\boldsymbol{u}}_{x_v,k}^{(p)}\ \cdots\ \hat{\boldsymbol{u}}_{x_v,d_{\hat{x}_v^{(p)}}}^{(p)}], \tag{10}$$

where $\hat{\boldsymbol{u}}_{x_v,k}^{(p)}$ are the generalized eigenvectors which were sorted with respect to descending eigenvalue order, and $d_{\hat{x}_v^{(p)}}$ is the dimension of each fNIRS-based visual feature. Finally, each fNIRS-based

visual feature $\hat{\boldsymbol{x}}_{v,k}^{(p)}$ is computed by projecting $\boldsymbol{x}_{v,k}^{(p)}$ via $\boldsymbol{U}_{cca}^{(p)}$ as follows:

$$\hat{\boldsymbol{x}}_{v,k}^{(p)} = \boldsymbol{U}_{cca}^{(p)^{\mathrm{T}}} \boldsymbol{x}_{v,k}^{(p)}. \tag{11}$$

Consequently, by maximizing the correlation between each visual feature and fNIRS features which reflect user preferences, we can transform each visual feature to the fNIRS-based visual features. In our method, since we cannot obtain fNIRS features corresponding to visual features extracted from test samples, we transform those visual features into the new features by using $\boldsymbol{U}_{cca}^{(p)}$ computed from training samples.

### 2.3. Feature Fusion via sMvLFDA

In this subsection, we explain the method to fuse multiple fNIRS-based visual features. We first define the fused features as follows: $\boldsymbol{Y} = [\boldsymbol{y}_1\ \boldsymbol{y}_2\ \cdots\ \boldsymbol{y}_n] \in \mathbb{R}^{d_y \times n}$, where $d_y$ denotes the dimension of the fused features. In our proposed method, $\boldsymbol{Y}$ consists of the fused features computed from both training samples and test samples which are defined as unlabeled samples. Next, by using the similarity matrix $\boldsymbol{A}^{(p)}$ constructed from each fNIRS-based visual feature in the same manner as the computation of $\boldsymbol{A}_{X_F}$ in 2.2, we compute the $(i,j)$ th elements of the matrices $\boldsymbol{A}_w^{(p)}$ and $\boldsymbol{A}_b^{(p)}$ as follows:

$$(\boldsymbol{A}_w^{(p)})_{i,j} = \begin{cases} \frac{(\boldsymbol{A}^{(p)})_{i,j}}{n_c} & \text{label}(\hat{\boldsymbol{x}}_{v,i}^{(p)}) = \text{label}(\hat{\boldsymbol{x}}_{v,j}^{(p)}) = c \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

$$(\boldsymbol{A}_b^{(p)})_{i,j} = \begin{cases} (\boldsymbol{A}^{(p)})_{i,j}(\frac{1}{n} - \frac{1}{n_c}) & \text{label}(\hat{\boldsymbol{x}}_{v,i}^{(p)}) = \text{label}(\hat{\boldsymbol{x}}_{v,j}^{(p)}) = c \\ \frac{1}{n} & \text{label}(\hat{\boldsymbol{x}}_{v,i}^{(p)}) \neq \text{label}(\hat{\boldsymbol{x}}_{v,j}^{(p)}) \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

where label$(\cdot)$ denotes a label presented to each sample. From the above matrices $\boldsymbol{A}_w^{(p)}$, $\boldsymbol{A}_b^{(p)}$ and $\boldsymbol{A}^{(p)}$, we compute Laplacian matrices $\boldsymbol{L}_w^{(p)}$, $\boldsymbol{L}_b^{(p)}$ and $\boldsymbol{L}^{(p)}$ in the same manner as the computation of $\boldsymbol{L}_{X_F X_V}^{(p)}$ in 2.2, respectively. Note that the Laplacian matrix $\boldsymbol{L}^{(p)}$ is normalized as follows:

$$\boldsymbol{L}^{(p)} = \boldsymbol{I} - \boldsymbol{D}^{(p)^{-1/2}} \boldsymbol{A}^{(p)} \boldsymbol{D}^{(p)^{-1/2}}, \tag{14}$$

where $\boldsymbol{I} \in \mathbb{R}^{n \times n}$ is the identity matrix. In addition, $\boldsymbol{D}^{(p)}$ is a diagonal matrix, and its entries are column sum of $\boldsymbol{A}^{(p)}$. From the obtained Laplacian matrices, we define the following optimization problem for each feature:

$$\arg\min_{\boldsymbol{Y}} (1-\beta)[\text{Tr}\{\boldsymbol{Y}(\boldsymbol{L}_w^{(p)} - \lambda_1 \boldsymbol{L}_b^{(p)})\boldsymbol{Y}^{\mathrm{T}}\}] + \beta\{\text{Tr}(\boldsymbol{Y}\boldsymbol{L}^{(p)}\boldsymbol{Y}^{\mathrm{T}})\}, \tag{15}$$

where $\beta$ is a trade-off parameter, and $\lambda_1$ is a manually set parameter. In the first term of Eq. (15), the labeled samples are used to optimize class separation of the fused feature space, and in the second term, all samples are used to preserve the local structure in the whole feature space and reduce overfitting to the labeled samples. In addition, let $(1-\beta)(\boldsymbol{L}_w^{(p)} - \lambda_1 \boldsymbol{L}_b^{(p)}) + \beta \boldsymbol{L}^{(p)}$ be $\bar{\boldsymbol{L}}^{(p)}$, and we rewrite Eq. (15) as follows:

$$\arg\min_{\boldsymbol{Y}} \boldsymbol{Y}\bar{\boldsymbol{L}}^{(p)}\boldsymbol{Y}^{\mathrm{T}}, \tag{16}$$

In sMvLFDA, we add weights to each objective function in order to consider different contributions to preference estimation and complementary properties in features and sum over all functions. Consequently, the following optimization problem is defined:

$$\arg\min_{\boldsymbol{Y},\boldsymbol{w}} \sum_{p=1}^{P} w_p(\boldsymbol{Y}\bar{\boldsymbol{L}}^{(p)}\boldsymbol{Y}^{\mathrm{T}}) + \lambda_2||\boldsymbol{w}||^2, \tag{17}$$

$$\text{s.t. } \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} = \boldsymbol{I},\ \sum_{p=1}^{P} w_p = 1,\ w_p \geq 0,$$

where $w_p$ is the weight coefficient to consider the different contributions of each feature to preference estimation and explore the complementary properties in features, and $\boldsymbol{w} = [w_1,\ w_2,\ \ldots\ , w_p]$. The second term is used for regularization.

To solve the above optimization problem, sMvLFDA alternatively optimizes the above objective function with respect to the fused features $\boldsymbol{Y}$ and the weight coefficients $\boldsymbol{w}$ as follows.

[**Update of $\boldsymbol{Y}$**]

We compute the optimal solution of $\boldsymbol{Y}$ by fixing $\boldsymbol{w}$. Specifically, we first define the following optimization problem:

$$\arg\min_{\boldsymbol{Y}} \text{Tr}(\boldsymbol{Y}\bar{\boldsymbol{L}}\boldsymbol{Y}^{\mathrm{T}}),\ \text{ s.t. } \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} = \boldsymbol{I}, \tag{18}$$

where $\bar{\boldsymbol{L}} = \sum_{p=1}^{P} w_p \bar{\boldsymbol{L}}^{(p)}$. Consequently, we solve Eq. (18) via the generalized eigen value problem and compute the optimal solution of $\boldsymbol{Y}$ as follows:

$$\boldsymbol{Y} = [\hat{\boldsymbol{y}}_1\ \hat{\boldsymbol{y}}_2\ \cdots\ \hat{\boldsymbol{y}}_k\ \cdots\ \hat{\boldsymbol{y}}_{d_y}], \tag{19}$$

where $\hat{\boldsymbol{y}}_k$ are the generalized eigenvectors of $\bar{\boldsymbol{L}}$, which were sorted with respect to descending eigenvalue order.

[**Update of $\boldsymbol{w}$**]

We compute the optimal solution of $\boldsymbol{w}$ by fixing $\boldsymbol{Y}$. We first define the following Lagrangian function:

$$F(\boldsymbol{w},\lambda) = \sum_{p=1}^{P} w_p \text{Tr}(\boldsymbol{Y}\bar{\boldsymbol{L}}^{(p)}\boldsymbol{Y}^{\mathrm{T}}) - \lambda_2||\boldsymbol{w}||^2 - \lambda(\sum_{p=1}^{P} w_p - 1), \tag{20}$$

where we differentiate $F(\boldsymbol{w},\lambda)$ with respect to $w_p$ and $\lambda$ and set the results to zero as follows:

$$\begin{aligned} \frac{\partial F(\boldsymbol{w},\lambda)}{\partial w_p} &= \text{Tr}(\boldsymbol{Y}\bar{\boldsymbol{L}}^{(p)}\boldsymbol{Y}^{\mathrm{T}}) - 2\lambda_2 w_p - \lambda \\ &= 0, \qquad\qquad p = 1,\ 2,\ \ldots, P, \end{aligned} \tag{21}$$

$$\begin{aligned} \frac{\partial F(\boldsymbol{w},\lambda)}{\partial \lambda} &= \sum_{p=1}^{P} w_p - 1 \\ &= 0. \end{aligned} \tag{22}$$

From Eqs. (21) and (22), $w_p$ can be obtained as follows:

$$w_p = \frac{P\mathrm{Tr}(\boldsymbol{Y}\bar{\boldsymbol{L}}^{(p)}\boldsymbol{Y}^{\mathrm{T}}) - \sum_{p=1}^{P}\mathrm{Tr}(\boldsymbol{Y}\bar{\boldsymbol{L}}^{(p)}\boldsymbol{Y}^{\mathrm{T}}) - 2\lambda_2}{2P\lambda_2}. \quad (23)$$

Consequently, from the above alternative optimization, we can obtain the optimal solution of $\boldsymbol{Y}$. Finally, we train an estimator by using labeled samples in $\boldsymbol{Y}$ and estimate user preferences for test samples.

## 3. EXPERIMENTAL RESULTS

This section presents the experimental results to verify the effectiveness of our proposed method. We used movie trailers of four genres ("action", "comedy", "drama" and "horror") as stimuli, and the number of movie trailers in each genre was eight. Thus, the total number of movie trailers was 32. All movie trailers were collected from YouTube[1]. We then separated the movie trailers into training and testing dataset. Both training and testing dataset consisted of 16 movie trailers, *i.e.*, the number of movie trailers in each genre was four in each dataset.

In this experiment, five healthy volunteers participated. We acquired fNIRS signals by the fNIRS device (LIGHTNIRS, Shimadzu Corporation, Japan) and placed 20 channels on the frontal and occipital regions of the scalp. fNIRS signals were acquired duration of the following experimental task: (1) relaxing video clip period (20s); (2) a fixation white cross on a black background (10s); (3) the movie trailer period; (4) the subjective rating (four levels). Note that while fNIRS signals were being acquired, participants watched movie trailers in the training dataset. In addition, participants watched movie trailers in the testing dataset and presented ratings to those movie trailers when fNIRS signals were not acquired. Consequently, we distributed movle trailers rated three or four into class "Like" and movie trailers rated one or two into class "Dislike".

In our experiment, we separated fNIRS signals and movie trailers into nonoverlapping segments and used these segments as each sample. We set the length of the segments to 10s. Note that we distributed each sample into the same class as corresponding movie trailers and fNIRS signals.

In our experiment, we used Support Vector Machine (SVM) [20] as the estimator. We then adopted F-measure as the evaluation measure. We randomly selected two movie trailers from each class and defined samples corresponding to selected movie trailers as labeled samples. On the other hand, we defined samples corresponding to not selected movie trailers as unlabeled samples. Consequently, we ran random selections of the labeled samples five times and averaged F-measure over all selections.

In our experiment, we conducted the following comparison to confirm the performance of our proposed method.

(a) Comparison with the method which used only visual features

To confirm the effectiveness of using fNIRS signals, we compared our proposed method with the method which used only visual features and fused these features via sMvLFDA. Results of this comparison are shown in Table 1. From the results, we can confirm that fNIRS-based visual features are effective for the preference estimation.

(b) Comparison with other feature fusion methods

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

**Table 1**. Experimental results to confirm the effectiveness of using fNIRS signals.

| | Only visual features | **fNIRS features and visual features** |
|---|---|---|
| Participant A | 0.752 | **0.791** |
| Participant B | 0.541 | **0.624** |
| Participant C | 0.810 | **0.832** |
| Participant D | 0.509 | **0.601** |
| Participant E | 0.586 | **0.648** |
| Average | 0.640 | **0.699** |

**Table 2**. Experimental results to compare the performance of our method with those of other feature fusion methods

| | MvLFDA | MSE | SMSE | **sMvLFDA** |
|---|---|---|---|---|
| Participant A | 0.749 | 0.784 | 0.780 | **0.791** |
| Participant B | 0.569 | **0.632** | 0.582 | 0.624 |
| Participant C | 0.809 | 0.827 | 0.825 | **0.832** |
| Participant D | 0.511 | 0.567 | 0.590 | **0.601** |
| Participant E | 0.604 | 0.618 | 0.636 | **0.648** |
| Average | 0.648 | 0.686 | 0.683 | **0.699** |

In this comparison, we first computed multiple fNIRS-based visual features via LPCCA. To confirm the effectiveness of fusing fNIRS-based visual features via sMvLFDA, we compared sMvLFDA with MvLFDA, Multiview Spectral Embedding (MSE) [21], which is the unsupervised method, and Supervised MSE (SMSE) [22]. Note that since SMSE can fused features while using labeled samples and unlabeled samples, we used SMSE as the semi-supervised method. In addition, in all methods, we constructed similarity matrices based on [19]. Results of this comparison are shown in Table 2. From the comparison between the results of the methods which used sMvLFDA and the other feature fusion methods, we can confirm that sMvLFDA outperforms the other feature fusion methods. Note that the p-value of a one-sided paired t-test performed between the results of sMvLFDA and MSE was 0.0815. Consequently, the effectiveness of fusing fNIRS-based visual features via sMvLFDA were shown.

## 4. CONCLUSIONS

In this paper, we have presented a new method to compute features effective for preference estimation. Our proposed method first extracts fNIRS features and multiple visual features. Next, we compute the fNIRS-based visual features which reflect user preferences via LPCCA. In addition, via sMvLFDA, we fuse multiple fNIRS-based visual features while using both labeled and unlabeled samples. Consequently, the experimental results have shown the effectiveness of the proposed method.

## 5. REFERENCES


[1] H. M. Blanken, H. E. Blok, L. Feng, and A. P. Vries, "Multimedia retrieval," 2007.

[2] SM. H. Hosseini, Y. Mano, M. Rostami, M. Takahashi, M. Sugiura, and R. Kawashima, "Decoding what one likes or dislikes from single-trial fNIRS measurements," *Neuroreport*, vol. 22. no. 6, pp. 269 – 273, 2011.

[3] A. Yazdani, J. S. Lee, J. M. Vesin, and T. Ebrahimi, "Affect recognition based on physiological changes during the watching of music videos," *ACM Transactions on Interactive Intelligent Systems*, vol. 2. no. 1, pp. 7:1 – 7:26, 2012.

[4] J, Han, X. Ji, X. Hu, L. Guo, and T. Liu, "Arousal recognition using audio-visual features and fmri-based brain response," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 337-347, 2015.

[5] D. A. Boas, C. E. Elwell, M. Ferrari, and G. Taga, "Twenty years of functional near-infrared spectroscopy: Introduction for the special issue," *NeuroImage*, vol. 85, no. 1, pp. 1-5, 2014.

[6] K. Tai and T. 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. 39-52, 2009.

[7] D. Heger, C. Herff, F. Putze, R. Mutter, and T. Schultz, "Continuous affective states recognition using functional near infrared spectroscopy," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 113-125, 2014.

[8] A. Toyoda, T. Ogawa, and M. Haseyama, "Video preference estimation using fNIRS signals," in *Proceedings of the IEEE Global Conference on Consumer Electronics*, pp. 297-298, 2017.

[9] T. Sun and S. Chen, "Locality preserving CCA with applications to data visualization and pose estimation," *Image and Vision Computing*, vol. 25, no. 5, pp. 531-543, 2007.

[10] A. Toyoda, T. Ogawa, and M. Haseyama, "MvLFDA-based video preference estimation using complementary properties of features," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 635-639, 2017.

[11] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[12] D. Jannach, Z. Karakaya, and F. Gedikli, "Accuracy improvements for multi-criteria recommender systems," in *Proceedings of the ACM Conference on Electronic Commerce*, pp. 674–689, 2012.

[13] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese, "Semi-supervised local Fisher discriminant analysis for dimensionality reduction," *Machine Learning*, vol. 78, no. 1, pp. 35–61, 2010.

[14] Y. Song, F. Nie, and C. Zhang, "Semi-supervised sub-manifold discriminant analysis," *Pattern Recognition Letters*, vol. 29, no. 13, pp. 1806–1813, 2008.

[15] Z. Ma, F. Nie, Y. Yang, J. Uijlings, N. Sebe, and A. G. Hauptmann, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1662-1672, 2012.

[16] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012.

[17] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 248-255, 2009.

[18] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE International Conference on Computer Vision,* pp. 1150-1157, 1999.

[19] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, "Learning with $\ell^1$-graph for image analysis," *IEEE Transactions on Image Processing*, vol. 19, no. 4, pp. 858–866, 2010.

[20] C. Cortes, and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[21] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 6, pp. 1438-1446, 2010.

[22] S. Liu, L. Zhang, W. Cai, Y. Song, Z. Wang, L. Wen, and D. D. Feng, "A supervised multiview spectral embedding method for neuroimaging classification," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 601-605, 2013.