NOVEL IMAGE CLASSIFICATION BASED ON DECISION-LEVEL FUSION OF EEG AND VISUAL FEATURES

Takuya Kawakami, 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
E-mail: {kawakami, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

ABSTRACT

This paper presents a novel image classification based on decision-level fusion of EEG and visual features. In the proposed method, we extract the EEG features from EEG signals recorded while users stare at images, and the visual features are computed from these images. Then the classification of images is performed based on Support Vector Machine (SVM) by separately using the EEG and visual features. Furthermore, we merge the above classification results based on Supervised Learning from Multiple Experts to obtain the final classification result. This method focuses on the classification accuracy calculated from each classification result. Therefore, although classification accuracy based on EEG and visual features are different from each other, our method realizes effective integration of these classification results. In addition, we newly derive a kernelized version of the method in order to realize more accurate integration of the classification results. Consequently, our method realizes successful multimodal classification of images by the object categories that they contain.

Index Terms— electroencephalogram, image classification, multimodal scheme, decision-level fusion.

1. INTRODUCTION

Image classification is an important task for image semantic analysis. Therefore, various methods which perform classification of images according to object categories that they contain have intensively been proposed [1–3].

Generally, most image classification methods try to improve the performance of visual features and classifiers [2, 3]. In particular, various local features (*e.g.*, Scale Invariant Feature Transform (SIFT) [4], Histograms of Oriented Gradients (HOG) [5], etc), have recently been proposed in order to improve the classification accuracy. Furthermore, the bag-of-features (BoF) [6] representation from these local features becomes one of the most widely used features in recent years.

Although the classification performance was improved by utilizing these visual features, the improvement of the classification accuracy based on discovery of new visual features tends to be saturated. Therefore, it is necessary to introduce a new idea such as solving the problem by multimodal schemes. In this paper, we newly utilize EEG features for image classification. Thus, we propose a multimodal image classification method which utilizes both EEG features and visual features collaboratively. We previously proposed a multimodal estimation method of segments where singing voices exist in musical pieces [7]. This method realized more accurate estimation of these segments than the method using only audio features. Therefore, by using EEG and visual features, more accurate classification than the method utilizing only visual features can be also expected.

In the proposed method, we first extract the EEG features from EEG signals recorded while users stare at images and their visual features. Next, we perform image classification using EEG and visual features separately to obtain multiple classification results. In our method, we employ Supervised Learning from Multiple Experts [8] in order to merge the above classification results, *i.e.*, decision-level fusion. Although the linear discriminating function is adopted in the original classifier [8], we newly derive the kernelized version of Supervised Learning from Multiple Experts in order to realize more accurate integration of classification results. Consequently, successful image classification becomes feasible by using the above non-conventional approach.

2. IMAGE CLASSIFICATION BASED ON DECISION-LEVEL FUSION

In this section, we explain the proposed method. Our method is composed of two stages. In the first stage, we extract the EEG features from EEG signals recorded while users stare at images, and the visual features are computed from these images. Then we perform image classification based on Support Vector Machine (SVM) [9] by inputting EEG and visual features into the classifiers, separately. Thus, multiple classification results are obtained. Furthermore, in the second stage, we employ the kernelized decision-level fusion approach, *i.e.,* merging the above classification results, considering their classification accuracy.

2.1. Feature Extraction and Single Feature-Based Image Classification

In this subsection, we explain the EEG features and the visual features used in the proposed method. Furthermore, the single feature-based image classification method in the first stage is presented.

EEG Feature Extraction

We calculate the EEG features from observed EEG signals and the power spectrum computed by applying short-time Fourier transform (STFT) to each channel's EEG signal. The details are shown below.

First, segmentation of each channel's EEG signal is performed at fixed intervals with an overlapped Hamming window. In this paper, $f_j (j = 1, 2, \cdots, F$; F is the total number of EEG segments) denote EEG segments. Next, we compute the EEG features shown in Table 1 from each EEG segment. Note that C and P denote the number of channels of EEG signals and the number of symmetric electrode pairs placed on the scalp, respectively. Thus, the dimension of EEG features becomes $6C + 10P$. In this table, we calculate Zero Crossing Rate [10] in the time domain, and the other features are computed in the frequency domain. The details of EEG features in our method are shown in [7].

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Table 1. Features used for EEG signals in the proposed method. Note that C denotes the number of channels of EEG signals and P shows the number of symmetric electrode pairs placed on the scalp.

EEG Features		Num. of Dimension
Zero Crossing Rate		C
Content percentage of the power spectrum	θ wave (4-7Hz)	C
	slow-α wave (7-9Hz)	C
	mid-α wave (9-11Hz)	C
	slow-α wave (11-13Hz)	C
	β wave (13Hz-)	C
Power spectrum of the hemispheric asymmetry [11]	θ wave (4-7Hz)	$2P$
	slow-α wave (7-9Hz)	$2P$
	mid-α wave (9-11Hz)	$2P$
	slow-α wave (11-13Hz)	$2P$
	β wave (13Hz-)	$2P$

Visual Feature Extraction

We utilize three kinds of visual features: SIFT [4], Pyramid Histogram of Oriented Gradients (PHOG) [12] and GIST descriptor [13].

SIFT: From each image, 128-dimensional SIFT descriptors are extracted, and BoF approach is applied to the obtained results to generate a feature vector $\boldsymbol{x}^{\mathrm{vSIFT}} (\in \mathbb{R}^{300})$.

PHOG: 40 bins of histogram is extracted at each resolution level l ($l = 1, 2, 3$), and the dimension of PHOG extracted from an image becomes 3400, where $\boldsymbol{x}^{\mathrm{vPHOG}} (\in \mathbb{R}^{3400})$ denotes the feature vector obtained by PHOG.

GIST: After 4×4 grid segmentation of an image, orientation histograms are extracted from each segment, and the number of bins of each histogram is 20. Then the dimension of a feature vector obtained by GIST descriptors becomes $360 (= 4 \times 4 \times 20 \times 3)$, where $\boldsymbol{x}^{\mathrm{vGIST}} (\in \mathbb{R}^{360})$.

Due to the limitation of pages, we only show the above overview of the visual features. The details can be found in [4], [12] and [13].

Single Feature-Based Image Classification

Next, we explain the method to classify images in the first stage. First, since relationships between "stimuli to human beings from the outside" and "which parts of the human brain are affected by these stimuli" are not well-known, we employ the feature selection in order to obtain EEG feature vectors. This means we reduce the dimension of the features shown in Table 1 to select only features useful for the classification. Specifically, we apply the feature selection method based on Max-Relevance and Min-Redundancy (mRMR) algorithm proposed in [14] to the EEG features calculated from each segment in order to obtain an efficient feature set for the classification. After this procedure, $\boldsymbol{x}_i^{f_j} \in \mathbb{R}^{d_{f_j}} (i = 1, 2, \cdots, N$; N is the number of images included in training data; d_{f_j} is the number of the selected features based on mRMR algorithm for EEG segment f_j) are obtained as EEG feature vectors for each EEG segment $f_j (j = 1, 2, \cdots, F)$. As for visual feature vectors, we directly use the vectors $\boldsymbol{x}_i^{\mathrm{vSIFT}}$, $\boldsymbol{x}_i^{\mathrm{vPHOG}}$ and $\boldsymbol{x}_i^{\mathrm{vGIST}}$, separately.

In the first stage of the proposed method, we employ SVM as the classifier to classify images. Although SVM is a two class classifier, image classification is generally a multi-class problem. Fortunately, since the two class classification can be easily expanded into multi-class classification based on one vs. one approach [15] or one vs. all approach [16], we focus on the improvement of the two class classification performance.

We train classifiers by separately using EEG feature vectors calculated from each EEG segment and visual feature vectors. This means multiple classifiers ($F + 3$ classifiers) are respectively obtained based on EEG features $\boldsymbol{x}_i^{f_1}, \boldsymbol{x}_i^{f_2}, \cdots, \boldsymbol{x}_i^{f_F}$ and visual features $\boldsymbol{x}_i^{\mathrm{vSIFT}}, \boldsymbol{x}_i^{\mathrm{vPHOG}}, \boldsymbol{x}_i^{\mathrm{vGIST}}$ by using each feature vector for training. Therefore, we can classify images based on EEG and visual features by inputting feature vectors extracted from test data into each trained classifier. Finally, $F + 3$ kinds of classification results are obtained.

2.2. Multiple Feature-Based Image Classification

In this subsection, we explain the method to obtain the final classification result in the second stage. In the proposed method, we merge the classification results obtained in the first stage based on Supervised Learning from Multiple Experts [8] to determine the final classification result. Whereas a linear discriminating function is adopted in [8], we newly derive its kernelized version in the proposed method. Therefore, it is expected that our method realizes more efficient classification than that of the original model proposed in [8]. Since this method has come from the research field of computer-aided diagnosis (CAD), they merge multiple classification results from each human *annotator*, *e.g.,* radiologist. In the proposed method, we regard the $F + 3$ classifiers based on EEG features extracted from each EEG segment and visual features as $F + 3$ *annotators*. In order to merge the multiple classification results, we focus on the classification accuracy of each annotator and assign higher weights to classification results of annotators which have higher classification accuracy. The details of the second stage are shown below.

2.2.1. Each annotator's classification accuracy and classification model

We explain the classification accuracy of each annotator and the classification model defined in our method. Let $y^j \in \{0, 1\}$ be the label assigned to the feature vector $\boldsymbol{x}$ by annotator $j \in \mathcal{J}$, where $\mathcal{J} = \{f_1, f_2, \cdots, f_F, \mathrm{v_{SIFT}}, \mathrm{v_{PHOG}}, \mathrm{v_{GIST}}\}$ is a set of annotators, and f and v correspond to Frame (EEG segment) and Visual, respectively. Furthermore, the details of $\boldsymbol{x}$ are shown in the following paragraph. Given the actual label $y \in \{0, 1\}$, *i.e.,* ground truth, the classification accuracy of each annotator, P_{se}^j (sensitivity) and P_{sp}^j (specificity) are respectively defined as follows:

$$P_{se}^j := \Pr[y^j = 1 | y = 1], \tag{1}$$

$$P_{sp}^j := \Pr[y^j = 0 | y = 0]. \tag{2}$$

In our method, classification model is specifically written as follows:

$$f_w(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}). \tag{3}$$

where $\boldsymbol{w}$ is a weight.

In the second stage of our method, the feature vector $\boldsymbol{x} \in \mathbb{R}^r$ is generated by applying the feature selection method based on mRMR algorithm to the EEG features obtained by calculating the average and standard deviation of each feature from all EEG segments $(f_1, f_2, \cdots, f_F)$ and all visual features calculated in the previous subsection (2.1). This means r-dimensional features are selected by mRMR algorithm from $(300 + 3400 + 360 + 2 \times (6C + 10P))$-dimensional features. In Eq. (3), $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^{r'}$ ($r' \gg r$) is obtained by mapping the feature vector $\boldsymbol{x}$ into a high-dimensional feature space. The final classification result $\hat{y}$ is obtained as follows:

$$\hat{y} = \begin{cases} 1 & f_w(\boldsymbol{x}) \geq Th \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where Th is a predetermined threshold. Given the training dataset $\mathcal{D}$ consisting of N feature vectors $\boldsymbol{x}_i \in \mathbb{R}^r (i = 1, 2, \cdots, N)$, a weight $\boldsymbol{w}$ is specifically written as follows:

$$\begin{aligned} \boldsymbol{w} &= \sum_{i=1}^{N} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i) \\ &= \boldsymbol{\Xi} \boldsymbol{\alpha}, \end{aligned} \tag{5}$$

where $\Xi = [\boldsymbol{\phi}(\boldsymbol{x}_1), \boldsymbol{\phi}(\boldsymbol{x}_2), \cdots, \boldsymbol{\phi}(\boldsymbol{x}_N)]$ and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_N]^\top$. Therefore, by using $\boldsymbol{\alpha}$ in Eq. (5), the discriminating function in Eq. (3) is rewritten as follows:

$$\begin{aligned} f_w(\boldsymbol{x}) &= \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) \\ &= \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}), \end{aligned} \tag{6}$$

where $k(\cdot,\cdot)$ is a kernel function of $\boldsymbol{\phi}(\cdot)$, and we specifically employ the Gaussian kernel. In order to determine the discriminating function $f_w(\cdot)$, we have to obtain the coefficients $\alpha_i (i = 1, 2, \cdots, N)$ from training data by using each annotator's classification accuracy defined in Eqs. (1) and (2). The details are shown below.

2.2.2. Training Phase

Given the training data $\mathcal{D}$ consisting of N feature vectors with the classification results by $F + 3$ annotators and their actual labels, $\mathcal{D} = \{y_i, \boldsymbol{\phi}(\boldsymbol{x}_i), y_i^{f_1}, y_i^{f_2}, \cdots, y_i^{f_F}, y_i^{\text{vSIFT}}, y_i^{\text{vPHOG}}, y_i^{\text{vGIST}}\}_{i=1}^N$, where y_i is the actual label, the estimation target is the coefficients α_i $(i = 1, 2, \cdots, N)$ in Eq. (6). From the training data $\mathcal{D}$, the likelihood of the coefficient vector $\boldsymbol{\alpha}$ is defined as:

$$\Pr[\mathcal{D}|\boldsymbol{\alpha}] = \prod_{i=1}^{N} \Pr\left[y_i^{f_1}, y_i^{f_2}, \cdots, y_i^{f_F}, y_i^{\text{vSIFT}}, y_i^{\text{vPHOG}}, y_i^{\text{vGIST}} | \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\alpha}\right]. \tag{7}$$

By using the values of sensitivity $\boldsymbol{P_{se}} = \{P_{se}^j | j \in \mathcal{J}\}$ obtained from each annotator and those of specificity $\boldsymbol{P_{sp}} = \{P_{sp}^j | j \in \mathcal{J}\}$, the above equation is rewritten as follows:

$$\begin{aligned} &\Pr[\mathcal{D}|\boldsymbol{\alpha}] \\ &= \prod_{i=1}^{N} \Big\{ \Pr[y_i^{f_1}, y_i^{f_2}, \cdots, y_i^{f_F}, y_i^{\text{vSIFT}}, y_i^{\text{vPHOG}}, y_i^{\text{vGIST}} | y_i = 1, \boldsymbol{P_{se}}] \\ &\qquad \times \Pr[y_i = 1 | \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\alpha}] \\ &\qquad + \Pr[y_i^{f_1}, y_i^{f_2}, \cdots, y_i^{f_F}, y_i^{\text{vSIFT}}, y_i^{\text{vPHOG}}, y_i^{\text{vGIST}} | y_i = 0, \boldsymbol{P_{sp}}] \\ &\qquad \times \Pr[y_i = 0 | \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\alpha}] \Big\}. \end{aligned} \tag{8}$$

If it is assumed that each annotator $j \in \mathcal{J}$ is independent each other, $\Pr[y_i^{f_1}, y_i^{f_2}, \cdots, y_i^{f_F}, y_i^{\text{vSIFT}}, y_i^{\text{vPHOG}}, y_i^{\text{vGIST}} | y_i = 1, \boldsymbol{P_{se}}]$ can be rewritten as follows:

$$\begin{aligned} &\Pr[y_i^{f_1}, y_i^{f_2}, \cdots, y_i^{f_F}, y_i^{\text{vSIFT}}, y_i^{\text{vPHOG}}, y_i^{\text{vGIST}} | y_i = 1, \boldsymbol{P_{se}}] \\ &= \prod_{j \in \mathcal{J}} [P_{se}^j]^{y_i^j} [1 - P_{se}^j]^{1 - y_i^j}. \end{aligned} \tag{9}$$

Similarly, $\Pr[y_i^{f_1}, y_i^{f_2}, \cdots, y_i^{f_F}, y_i^{\text{vSIFT}}, y_i^{\text{vPHOG}}, y_i^{\text{vGIST}} | y_i = 0, \boldsymbol{P_{sp}}]$ can be rewritten as follows:

$$\begin{aligned} &\Pr[y_i^{f_1}, y_i^{f_2}, \cdots, y_i^{f_F}, y_i^{\text{vSIFT}}, y_i^{\text{vPHOG}}, y_i^{\text{vGIST}} | y_i = 0, \boldsymbol{P_{sp}}] \\ &= \prod_{j \in \mathcal{J}} [P_{sp}^j]^{1 - y_i^j} [1 - P_{sp}^j]^{y_i^j}. \end{aligned} \tag{10}$$

Then the likelihood in Eq. (8) is rewritten as

$$\Pr[\mathcal{D}|\boldsymbol{\alpha}] = \prod_{i=1}^{N} [a_i p_i + b_i (1 - p_i)]. \tag{11}$$

Note that

$$\begin{aligned} p_i &= \Pr[y_i = 1 | \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\alpha}] \\ &= \frac{1}{1 + \exp(-\boldsymbol{\alpha}^\top \boldsymbol{k}_i)}, \end{aligned} \tag{12}$$

where $\boldsymbol{k}_i = [k(\boldsymbol{x}_1, \boldsymbol{x}_i), k(\boldsymbol{x}_2, \boldsymbol{x}_i), \cdots, k(\boldsymbol{x}_N, \boldsymbol{x}_i)]^\top$, and

$$a_i = \prod_{j \in \mathcal{J}} [P_{se}^j]^{y_i^j} [1 - P_{se}^j]^{1 - y_i^j}, \tag{13}$$

$$b_i = \prod_{j \in \mathcal{J}} [P_{sp}^j]^{1 - y_i^j} [1 - P_{sp}^j]^{y_i^j}. \tag{14}$$

The maximum-likelihood estimator is found by maximizing the following log-likelihood:

$$\hat{\boldsymbol{\alpha}}_{ML} = \arg\max_{\boldsymbol{\alpha}} \Big\{ \ln \Pr[\mathcal{D}|\boldsymbol{\alpha}] \Big\}. \tag{15}$$

Let $\boldsymbol{y} = [y_1, \cdots, y_N]$ be the set of the actual labels, and the complete data log-likelihood can be written as

$$\ln \Pr[\mathcal{D}, \boldsymbol{y}|\boldsymbol{\alpha}] = \sum_{i=1}^{N} \{ y_i \ln p_i a_i + (1 - y_i) \ln (1 - p_i) b_i \}. \tag{16}$$

In order to maximize this likelihood, the following Expectation-Maximization (EM) algorithm [17] is adopted.

E-step

In the E-step, when the training data $\mathcal{D}$ and the current estimate of the coefficient vector $\boldsymbol{\alpha}$ are given, the conditional expected value of log-likelihood is computed as follows:

$$\mathrm{E}\Big\{ \ln \Pr[\mathcal{D}, \boldsymbol{y}|\boldsymbol{\alpha}] \Big\} = \sum_{i=1}^{N} \Big\{ \mu_i \ln p_i a_i + (1 - \mu_i) \ln (1 - p_i) b_i \Big\}, \tag{17}$$

where μ_i is computed as follows:

$$\begin{aligned} \mu_i &\propto \Pr[y_i^{f_1}, y_i^{f_2}, \cdots, y_i^{f_F}, y_i^{\text{vSIFT}}, y_i^{\text{vPHOG}}, y_i^{\text{vGIST}} | y_i = 1, \boldsymbol{\alpha}] \\ &\quad \times \Pr[y_i = 1 | \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\alpha}] \\ &= \frac{a_i p_i}{a_i p_i + b_i (1 - p_i)}. \end{aligned} \tag{18}$$

M-step

In the M-step, based on the current estimate μ_i and the training data $\mathcal{D}$, the coefficient vector $\boldsymbol{\alpha}$ is estimated by maximizing the conditional expected value in Eq. (17). Specifically, we obtain the estimated coefficient vector $\boldsymbol{\alpha}$ by solving equation $\frac{\partial}{\partial \boldsymbol{\alpha}} \{ \ln \Pr[\mathcal{D}, \boldsymbol{y}|\boldsymbol{\alpha}] \} = 0$.

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta \boldsymbol{H}^{-1} \boldsymbol{g}. \tag{19}$$

In Eq. (19), $\boldsymbol{g}$ is a gradient vector, $\boldsymbol{H}$ is a Hessian matrix and η is a step length. The gradient vector $\boldsymbol{g}$ and the Hessian matrix $\boldsymbol{H}$ are respectively computed as follows:

$$\boldsymbol{g} = \sum_{i=1}^{N} [\mu_i - \sigma(\boldsymbol{\alpha}^\top \boldsymbol{k}_i)] \boldsymbol{k}_i, \tag{20}$$

$$\boldsymbol{H} = -\sum_{i=1}^{N} [\sigma(\boldsymbol{\alpha}^\top \boldsymbol{k}_i)][1 - \sigma(\boldsymbol{\alpha}^\top \boldsymbol{k}_i)] \boldsymbol{k}_i \boldsymbol{k}_i^\top, \tag{21}$$

where $\sigma(\boldsymbol{\alpha}^\top \boldsymbol{k}_i) = \frac{1}{1 + \exp(-\boldsymbol{\alpha}^\top \boldsymbol{k}_i)}$.

2.2.3. Testing Phase

Give the test data, the final classification result can be obtained as follows. In the previous phase, we essentially solved a regular logistic regression problem with probabilistic labels μ_i. Thus, we obtain the final classification result by applying a threshold to μ calculated from a test data $\{\boldsymbol{\phi}(\boldsymbol{x}), y^{f_1}, y^{f_2}, \cdots, y^{f_F}, y^{\text{vSIFT}}, y^{\text{vPHOG}}, y^{\text{vGIST}}\}$, where its label y is unknown, instead of directly using $\boldsymbol{\alpha}$. The value of μ is computed by using the estimated coefficient vector $\boldsymbol{\alpha}$ and a, b calculated from the training data. Specifically, $p = \frac{1}{1 + \exp(-\boldsymbol{\alpha}^\top \boldsymbol{k})}$ is calcu-

Table 2. Classification accuracy: As for the results of only visual features, since we utilize the same images, these values do not depend on each subject.

	Only Visual Features			Only EEG Features			Proposed Method		
	SIFT	PHOG	GIST	subject A	subject B	subject C	subject A	subject B	subject C
panda	78 ± 0.2%	72 ± 0.2%	76 ± 0.2%	60 ± 0.3%	64 ± 0.2%	68 ± 0.2%	80 ± 0.2%	82 ± 0.2%	78 ± 0.2%
soccer ball	76 ± 0.2%	72 ± 0.3%	62 ± 0.1%	52 ± 0.2%	70 ± 0.2%	68 ± 0.2%	78 ± 0.2%	82 ± 0.2%	76 ± 0.2%
strawberry	66 ± 0.2%	58 ± 0.2%	66 ± 0.2%	62 ± 0.1%	74 ± 0.2%	46 ± 0.2%	92 ± 0.1%	88 ± 0.2%	86 ± 0.2%
Ave.	73 ± 0.05%	67 ± 0.07%	68 ± 0.06%	58 ± 0.04%	69 ± 0.04%	61 ± 0.10%	83 ± 0.06%	84 ± 0.03%	80 ± 0.04%

lated, where $\boldsymbol{k} = [k(\boldsymbol{x}_1, \boldsymbol{x}), k(\boldsymbol{x}_2, \boldsymbol{x}), \cdots, k(\boldsymbol{x}_N, \boldsymbol{x})]^\top$. Furthermore, $a = \prod_{j \in \mathcal{J}} [P^j_{se}]^{y^j} [1 - P^j_{se}]^{1-y^j}$ and $b = \prod_{j \in \mathcal{J}} [P^j_{sp}]^{1-y^j} [1 - P^j_{sp}]^{y^j}$ are obtained, where P^j_{se} and P^j_{sp} are accuracy of annotator j calculated from training data and y^j is classification result of the test data. Therefore, we obtain the final classification result considering each annotator's accuracy. Then $\mu = \frac{ap}{ap+b(1-p)}$ is computed by using p, a and b. Finally, we obtain the final classification result as follows:

$$y = \begin{cases} 1 & \mu \geq \gamma \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

where γ is a predetermined threshold. The value of μ is the posterior probability.

3. EXPERIMENTAL RESULTS

In this section, we show experimental results to verify the effectiveness of the proposed method. First, we explain EEG signal collection and the experimental procedures in 3.1. Furthermore, the results of image classification are shown in 3.2.

3.1. EEG Signal Collection and Experimental Procedures

In this subsection, we first explain how to collect EEG signals in this experiment. In this study, three healthy subjects participated, and EEG recordings were conducted during staring at images. The age of each subject was 22 or 23 years old. We recorded EEG signals from 12 channels (Fp1, Fp2, F7, F8, T3, T4, C3, C4, P3, P4, O1 and O2) according to the international 10-20 system. All leads were referenced to linked earlobes, and a ground electrode was located in the forehead. Since EEG signals are weak, we amplified these signals by using an amplifier (MEG-6116M, NIHON KOHDEN). We also applied a band-pass filter to recorded EEG signals to avoid artifacts, and set the filter bandwidth to 0.04-30Hz. and the sampling rate is 2kHz. Each subject stared at images displayed on a monitor. During experiment, subject sat comfortably on a chair and kept relaxing. Subjects were instructed that they stared at images without blinking.

We collected EEG signals while subjects stared at various kinds of images. In order to recognize the objects in images easily, we let subjects know what kinds of categories are used for this experiment. Specifically, we presented each subject the two kinds of images which are used for image classification (target images) and not used for the classification (non-target images) sequentially. In addition, EEG signals were recorded while subjects performed the task in a scheme similar to an oddball paradigm. The number of times each subject performed the task was the same as the number of target image categories. The time length of staring at each image was three seconds. Three second silence was inserted between every two images to remove the effect of a previous image. In this experiment, we performed image classification by using EEG signals recorded while subjects were staring at target images. In addition, we set the time length of an EEG segment and an overlapping to 100ms and 50ms, respectively.

3.2. Experimental Results

In this subsection, we show the experimental results in order to verify the performance of the proposed method. In this experiment, we utilized Caltech101 dataset [18]. This dataset consists of images from 101 categories. The significance of this database is its large inter-class variability. Specifically, we used the images included in "panda", "soccer ball" and "strawberry" in the database as the target images, and the number of images was 35 per category. These images were randomly selected in advance. We also used the images included in "airplane", "elephant", "joshua tree", "pyramid" and "stapler" in the same database for the non-target images.

In this experiment, we performed the multi-class classification of images by the object categories that they contain based on one vs. all approach [16]. Therefore, the final classification was determined according to the posterior probability obtained from the testing phase (2.2.3). We followed [2, 3] for our experimental setup. Specifically, we randomly selected 30 training images per class and test on the remaining images. Then we calculated the classification accuracy which was normalized according to the number of test images per class. We repeated the random selection 10 times and show the average classification accuracy per class.

We show the results of image classification in Table 2. In this table, we also show the results of the comparative methods. From the obtained results, the proposed method realizes the most accurate classification. Therefore, the effectiveness of our method can be verified. In Table 2, the most significant improvement of classification performance by only using visual features is 6% (67% to 73%). However, the classification accuracy based on the proposed method, which utilizes both EEG features and visual features, is higher than that of SIFT, which is the highest classification accuracy in only visual features, by average 9%. Therefore, our multimodal approach is more effective than the improvement of visual features. In addition, while only subject B's classification accuracy based on only EEG features is higher than the accuracy of PHOG and GIST, our method realizes accurate classification for all subjects. Therefore, the effectiveness of the proposed method can be verified.

4. CONCLUSION

In this paper, we have proposed a novel image classification. Our method realized the improvement of the classification performance based on decision-level fusion of EEG and visual features. Experimental results show the effectiveness of the proposed method.

5. ACKNOWLEDGMENT

This work was partly supported by Grant-in-Aid for Scientific Research (B) 25280036 from JSPS, and Grant-in-Aid for Scientific Research on Innovative Areas 24120001 from the MEXT.

6. REFERENCES

[1] M. Haseyama, T. Ogawa, and N. Yagi, "A review of video retrieval based on image and video semantic understanding," *ITE Transactions on Media Technology and Applications*, vol. 1, no. 1, pp. 2–9, 2013.

[2] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International journal of computer vision*, vol. 73, no. 2, pp. 213–238, 2007.

[3] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2005, vol. 2, pp. 1458–1465.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 1, pp. 886–893.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proceedings of European Conference on Computer Vision*, 2004, vol. 1, pp. 1–22.

[7] T. Kawakami, T. Ogawa, and M. Haseyama, "Vocal segment estimation in music pieces based on collaborative use of EEG and audio features," in in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 1197–1201.

[8] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 889 –896, 2009.

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

[10] Alexander A. Borbely and HansUlrich Neuhaus, "Sleep-deprivation: Effects on sleep and EEG in the rat," *Journal of comparative physiology*, vol. 133, no. 1, pp. 71–87, 1979.

[11] Y. P. Lin, C. H. Wang, T. P. Jung, T. L. Wu, S. K. Jeng, J. R. Duann, and J. H. Chen, "EEG-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798 –1806, 2010.

[12] Anna Bosch, Andrew Zisserman, and Xavier Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, 2007, pp. 401–408.

[13] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226 –1238, 2005.

[15] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *The annals of statistics*, vol. 26, no. 2, pp. 451–471, 1998.

[16] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *The Journal of Machine Learning Research*, vol. 5, pp. 101–141, 2004.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, pp. 1–38, 1977.

[18] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2004, pp. 178–178.