MULTI-FEATURE FUSION BASED ON SUPERVISED MULTI-VIEW MULTI-LABEL CANONICAL CORRELATION PROJECTION

Keisuke Maeda†, Sho Takahashi††, Takahiro Ogawa† and Miki Haseyama†††

Graduate School of Information Science and Technology, Hokkaido University† †††
Faculty of Engineering, Hokkaido University††
N-14, W-9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan† †††
N-13, W-8, Kita-ku, Sapporo, Hokkaido, 060-8628, Japan††
E-mail: {maeda, sho, ogawa}@lmd.ist.hokudai.ac.jp† ††, miki@ist.hokudai.ac.jp†††

ABSTRACT

This paper presents multi-feature fusion based on supervised multi-view multi-label canonical correlation projection (sM2CP). The proposed method applies sM2CP-based feature fusion to multiple features obtained from various convolutional neural networks (CNNs) whose characteristics are different. Since new fused features with high representation ability can be obtained, performance improvement of multi-label classification is realized. Specifically, in order to tackle the multi-label problem, sM2CP introduces a label similarity information of label vectors into the objective function of supervised multi-view canonical correlation analysis. Thus, sM2CP can deal with complex label information such as multi-label annotation. The main contribution of this paper is the realization of feature fusion of multiple CNN features for the multi-label problem by introducing multi-label similarity information into the canonical correlation analysis-based feature fusion approach. Experimental results show the effectiveness of sM2CP, which enables effective fusion of multiple CNN features.

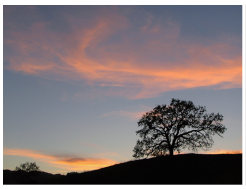
Index Terms— Multi-label, multi-view, feature fusion, canonical correlation, convolutional neural network.

1. INTRODUCTION

In order to improve the performance of various tasks such as image classification and retrieval, deep learning-based approaches have been proposed [1]. Beginning with AlexNet [2], various convolutional neural networks (CNNs) [3, 4] have been proposed. Since recent CNNs can effectively train a large number of hidden layers, they can achieve high performance for basic tasks. Recently, transfer learning such as fine-tuning of pre-trained networks and use of CNN features calculated from pre-trained networks has been drawing attention [5]. Especially, since CNN features with high representation ability can be easily calculated without training, they are used in various tasks in various fields [6].

In the case that CNNs pre-trained from a large scale dataset such as ImageNet [7] are used as feature extractors, even though the same dataset is used for training of each CNN, classification results are often different for each CNN since their network structures are different. For example, as shown in Fig. 1, when an image is input into two CNNs, the results obtained from different networks do not necessarily become the same. That is, there exist differences between representation ability of CNNs. In order to improve the performance, the following two solutions can be considered: "selection of the optimal CNN suitable for the task from a large number of CNNs" and "feature fusion which can consider the characteristics of different CNN features". Since the number of CNN architectures will increase, selection of CNNs is unrealistic. On the other hand, it was reported in traditional studies that classification and retrieval accuracy was improved by multi-feature fusion of local visual features with different characteristics such as HoG and GIST [8]. Therefore, it is expected that by applying a feature fusion approach to CNN features calculated from different networks, it is possible to break limitations of only using a single CNN feature.

In general feature fusion approaches, canonical correlation anal-



	DenseNet-201	Inception-ResNet v2
1st	volcano	beacon
2nd	barn	promontory
3rd	valley	cliff
4th	lakeside	rapeseed
5th	seashore	seashore

	DenseNet-201	Inception-ResNet v2
1st	prison	paintbrush
2nd	gondola	umbrella
3rd	paintbrush	swing
4th	miniskirt	swab
5th	rifle	cellular_telephone

Fig. 1. Classification results of images obtained from two CNNs, DenseNet201 [3] and Inception-ResNet-v2 [4] pre-trained from ImageNet. "1st" represents the class with the highest output probability.

Fig. 2. Examples of images including multiple labels. Although these images include both "people" and "dog", the left image belongs to "dog", and the right image belongs to "people", respectively.

This work was partly supported by JSPS KAKENHI Grant Numbers JP18J10373 and JP17H01744 and Global Station for Big Data and Cyber Security, a project of Global Institution for Collaborative Research and Education at Hokkaido University.

ysis (CCA) [9] is commonly used [10]. Recently, various methods, which extend CCA, have been proposed, and many of these methods are constructed for single label annotation [11, 12]. Although the single-label annotation roughly represents semantic meaning of images, high-level semantics cannot be extracted via only the single-label annotation. As shown in Fig. 2, although images from Caltech-256 [13] belong to one class, they include various meanings other than a single object. Thus, it is necessary to consider the situation in the multi-label annotation since there are many cases where not only a single object but also multiple objects exist in actual data [14].

In this paper, we newly propose multi-feature fusion based on supervised multi-view multi-label canonical correlation projection (sM2CP). The main contribution of this paper is the realization of feature fusion of multiple CNN features for the multi-label problem by introducing multi-label similarity information into the CCA-based feature fusion approach. In the situation where labels are given and multiple kinds of features are provided, it has been reported that supervised multi-view canonical correlation analysis (sMVCCA) [15], which introduces the concept of both supervised learning and multi-view, is effective. Since sMVCCA uses label information for improving discriminability, it is effective for simple tasks such as the single-label problem. In the multi-label problem, there is an issue that the number of labels assigned to each sample is different. Due to the above problem, imbalance of the number of labels assigned to each sample affects the performance for calculating the optimal projection of sMVCCA, and it cannot extract highly semantic features. In order to solve this problem, we construct sM2CP by introducing the multi-label similarity into the objective function of sMVCCA. This approach is inspired by multi-label CCA (mlCCA) [14], and the advantages of sMVCCA and mlCCA are included in sM2CP. Consequently, we apply sM2CP to multiple CNN features and can calculate the optimal projection for the multi-label problem. Experimental results show the effectiveness of sM2CP by comparing our method with various fusion methods including sMVCCA and mlCCA.

2. MULTI-FEATURE FUSION BASED ON SM2CP

In this section, we show the details of sM2CP. sM2CP includes the following two points: (i) dealing with multi-label problems based on label similarities and (ii) supervised multi-view learning which can handle several modalities. Given training images n ($n = 1, 2, ..., N$; N being the number of training images), we extract multiple visual features $\boldsymbol{x}_m^n \in \mathbb{R}^{d_m}$ ($m = 1, ..., M$). Note that d_m is the number of dimensions of $\boldsymbol{x}_m^n$ obtained by inputting an image n into mth CNN architecture. Furthermore, multiple class labels are assigned to each image, and we obtain class label vectors $\boldsymbol{y}^n \in \mathbb{R}^C$ (C being the number of classes). The elements corresponding to their own classes are one, and the others are zero. In the proposed method, we regard multiple CNN architectures as the modalities. In addition, we also regard class label vectors $\boldsymbol{y}^n$ as $(M+1)$th modality in the proposed method, i.e., we newly define $\boldsymbol{x}_{M+1}^n = \boldsymbol{y}^n (n = 1, 2, ..., N)$.

We calculate the optimal projection $\boldsymbol{w}_m \in \mathbb{R}^{d_m}$ which can effectively integrate these multiple CNN features by maximizing the following objective function:

$$\arg \max_{\boldsymbol{w}_1,...,\boldsymbol{w}_{M+1}} \sum_{m_1=1}^{M+1} \sum_{m_2=1,m_2\neq m_1}^{M+1} \frac{\boldsymbol{w}_{m_1}^\top \overline{\boldsymbol{P}}_{m_1,m_2} \boldsymbol{w}_{m_2}}{\sqrt{\boldsymbol{w}_{m_1}^\top \boldsymbol{P}_{m_1,m_1} \boldsymbol{w}_{m_1}} \sqrt{\boldsymbol{w}_{m_2}^\top \boldsymbol{P}_{m_2,m_2} \boldsymbol{w}_{m_2}}}, \tag{1}$$

where $\boldsymbol{P}_{m_1,m_2}$ is the covariance matrix between modalities m_1 and m_2. Furthermore, $\overline{\boldsymbol{P}}_{m_1,m_2}$ is the covariance matrix dealing with class information. Since the scaling of $\boldsymbol{w}$ does not have influence on the optimization, Eq. (1) is rewritten as

$$\begin{aligned} \arg \max_{\boldsymbol{w}_1,...,\boldsymbol{w}_{M+1}} & \sum_{m_1=1}^{M+1} \sum_{m_2=1,m_2\neq m_1}^{M+1} \boldsymbol{w}_{m_1}^\top \overline{\boldsymbol{P}}_{m_1,m_2} \boldsymbol{w}_{m_2} \\ \text{s.t.} \quad & \boldsymbol{w}_{m_1}^\top \boldsymbol{P}_{m_1,m_1} \boldsymbol{w}_{m_1} = 1 \ (m_1 = 1, 2, ..., M+1). \end{aligned} \tag{2}$$

By defining $\boldsymbol{W} = [\boldsymbol{W}_1^\top, \boldsymbol{W}_2^\top, ..., \boldsymbol{W}_{M+1}^\top]^\top \in \mathbb{R}^{(d_1+...+d_M+C)\times(d_p\times(M+1))}$, where d_p is the dimension of the projection, Eq. (2) can be rewritten as

$$\arg \max_{\boldsymbol{W}} \quad \text{trace}(\boldsymbol{W}^\top \overline{\boldsymbol{P}} \boldsymbol{W}) \quad \text{s.t.} \quad \boldsymbol{W}^\top \boldsymbol{P}_d \boldsymbol{W} = \boldsymbol{I}, \tag{3}$$

where

$$\overline{\boldsymbol{P}} = \begin{bmatrix} \boldsymbol{0} & \overline{\boldsymbol{P}}_{1,2} & \cdots & \overline{\boldsymbol{P}}_{1,M} & \overline{\boldsymbol{P}}_{1,M+1} \\ \overline{\boldsymbol{P}}_{2,1} & \boldsymbol{0} & \cdots & \overline{\boldsymbol{P}}_{2,M} & \overline{\boldsymbol{P}}_{2,M+1} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \overline{\boldsymbol{P}}_{M,1} & \cdots & \cdots & \cdots & \overline{\boldsymbol{P}}_{M,M+1} \\ \overline{\boldsymbol{P}}_{M+1,1} & \cdots & \cdots & \overline{\boldsymbol{P}}_{M+1,M} & \boldsymbol{0} \end{bmatrix}, \tag{4}$$

$$\boldsymbol{P}_d = \begin{bmatrix} \boldsymbol{P}_{1,1} & \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_{2,2} & \cdots & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \cdots & \boldsymbol{P}_{M,M} & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} & \boldsymbol{P}_{M+1,M+1} \end{bmatrix}. \tag{5}$$

The matrix $\overline{\boldsymbol{P}}_{m_1,m_2}$ is detailed below. In multi-label classification problems, since one sample often has some labels, we focus on label information for each class based on multi-label linear discriminant analysis [16] in order to calculate the covariance matrix. Thus, given a mean-normalized $\boldsymbol{X}_m = [\boldsymbol{x}_m^1, \boldsymbol{x}_m^2, \cdots, \boldsymbol{x}_m^N]$, the covariance matrix $\overline{\boldsymbol{P}}_{m_1,m_2}$ is obtained as follows:

$$\overline{\boldsymbol{P}}_{m_1,m_2} = \sum_{i=1}^{C} \sum_{k=1}^{N} \sum_{l=1}^{N} y_i^k y_i^l \boldsymbol{x}_{m_1}^k (\boldsymbol{x}_{m_2}^l)^\top, \tag{6}$$

where $y_i^k \in \{0, 1\}$. If kth sample has ith class label, $y_i^k = 1$, otherwise $y_i^k = 0$. Furthermore, $\overline{\boldsymbol{P}}_{m_1,m_2}$ in Eq. (6) is rewritten as

$$\begin{aligned} \overline{\boldsymbol{P}}_{m_1,m_2} &= \sum_{i=1}^{C} \boldsymbol{X}_{m_1} \boldsymbol{A}_i^{multi} \boldsymbol{X}_{m_2}{}^\top \\ &= \boldsymbol{X}_{m_1} \boldsymbol{A}^{multi} \boldsymbol{X}_{m_2}{}^\top, \end{aligned} \tag{7}$$

where $\boldsymbol{A}^{multi} = \boldsymbol{A}_1^{multi} + \boldsymbol{A}_2^{multi} + \cdots + \boldsymbol{A}_C^{multi} \in \mathbb{R}^{N\times N}$. If samples n_1 and n_2 have ith class label, (n_1, n_2)th element of $\boldsymbol{A}_i^{multi}$ is one. Thus, each element of $\boldsymbol{A}^{multi}$ represents the number of labels shared by two samples. Equation (7) is formulated based on the definition of the within-class covariance matrix in discriminative CCA (DCCA) [17], which is a traditional CCA-based method, and the definition is very simple. However, the definition is not suitable for the multi-label problem since DCCA can deal with only the single-label problem. Thus, it is necessary to improve the definition of Eq. (7), and its solution is explained below.

Since the number of labels assigned to each sample is different, imbalance of the number of labels assigned to each sample affects $\boldsymbol{A}^{multi}$. For example, when sample n_1 has three labels, the $(n_1, *)$ element of $\boldsymbol{A}^{multi}$ has a maximum of three. On the other hand, when sample n_2 has five labels, the $(n_2, *)$ element of $\boldsymbol{A}^{multi}$ has a max-

imum of five. Therefore, since the range of values is different between different samples, it is difficult to consider the class information correctly.

In order to solve the above problem, we introduce the mlCCA-based approach into the proposed method. Specifically, we use the cosine similarity between multi-label vectors of samples n_1 and n_2 for calculating $\boldsymbol{A}^{multi}$ as follows:

$$\boldsymbol{A}^{multi} = \begin{bmatrix} f(\boldsymbol{y}^1, \boldsymbol{y}^1) & f(\boldsymbol{y}^1, \boldsymbol{y}^2) & \cdots & \cdots & f(\boldsymbol{y}^1, \boldsymbol{y}^N) \\ f(\boldsymbol{y}^2, \boldsymbol{y}^1) & f(\boldsymbol{y}^2, \boldsymbol{y}^2) & \cdots & \cdots & f(\boldsymbol{y}^2, \boldsymbol{y}^N) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ f(\boldsymbol{y}^N, \boldsymbol{y}^1) & f(\boldsymbol{y}^N, \boldsymbol{y}^2) & \cdots & \cdots & f(\boldsymbol{y}^N, \boldsymbol{y}^N) \end{bmatrix}, \quad (8)$$

where

$$f(\boldsymbol{y}^{n_1}, \boldsymbol{y}^{n_2}) = \frac{\boldsymbol{y}^{n_1\top}\boldsymbol{y}^{n_2}}{\|\boldsymbol{y}^{n_1}\| \, \|\boldsymbol{y}^{n_2}\|}. \quad (9)$$

Finally, we solve the following generalized eigenvalue problem:

$$\overline{\boldsymbol{P}}\boldsymbol{W} = \lambda(\boldsymbol{P}_d + \gamma \boldsymbol{I})\boldsymbol{W}, \quad (10)$$

where γ is a regularization parameter. Then we can obtain the optimal projection $\hat{\boldsymbol{W}}_m \in \mathbb{R}^{d_m \times d_p}$ for fusing multiple CNN features. The matrix $\hat{\boldsymbol{W}}_m$ consists of the eigenvectors of the d_p-largest eigenvalues, and d_p becomes the number of dimensions of the projected features as explained before. Note that $d_p \leq \min(d_1, ..., d_M, C)$. By using the optimal projection matrix, we can calculate the projected features as follows:

$$\boldsymbol{Z}_m = \hat{\boldsymbol{W}}_m^\top \boldsymbol{X}_m \in \mathbb{R}^{d_p \times N}, \quad (11)$$

where $\boldsymbol{Z}_m = [\boldsymbol{z}_m^1, \boldsymbol{z}_m^2, ..., \boldsymbol{z}_m^N]$. Then we can obtain the projected CNN features $\boldsymbol{z}^n = [(\boldsymbol{z}_1^n)^\top, (\boldsymbol{z}_2^n)^\top, ..., (\boldsymbol{z}_M^n)^\top]^\top$ of nth sample and can construct a classifier by inputting these features. Consequently, sM2CP can perform the feature fusion of multiple CNN features for the multi-label problem by introducing the similarities of multi-label vectors into the objective function of sMVCCA.

3. EXPERIMENTAL RESULTS

In this section, we show the effectiveness of our method. The details of the dataset, CNN features and classifiers, and comparative methods used in this experiment are explained in **3.1**, **3.2** and **3.3**, respectively. Furthermore, we show evaluation indices, parameter settings and performance evaluation in **3.4**, **3.5** and **3.6**, respectively.

3.1. Details of Dataset

In this experiment, we used MIRFlickr-25K dataset [18]. The MIRFlickr-25K dataset consists of 25,000 images collected from the social photography website, Flickr. All images are annotated with 24 semantic concepts including various scenes and objects categories such as dog, sunset, sky and river. For the 25,000 images, we performed five-fold cross validation. In order to search parameters used in the proposed method and other comparative methods, we divided the training dataset into validation dataset for training and that for test.

3.2. Details of CNN Features and Classifiers

In order to verify the effectiveness and robustness of the proposed method, we adopt the three kinds of CNN architectures. Specifically, DenseNet-201 [3], Inception-ResNet-v2 [4] and ResNet50 [19] pre-traind by using ImageNet are used in this experiment. We extracted visual features from the middle layer of the network. The dimensions d_1, d_2 and d_3 of DenseNet-201, Inception-ResNet-v2 and ResNet50 were 1920, 1536 and 2048, respectively. We applied the proposed method and other comparative methods to all combination of these features. Furthermore, we trained Extreme Learning Machine (ELM) [20] as a classifier by using the projected features $\boldsymbol{z}^n$ via sM2CP and obtained class labels via thresholding the output values of ELM. When output values of ELM are larger than the threshold value, indices corresponding to these output values become estimated labels.

3.3. Comparative Methods

For comparing the classification performance, we adopted some comparative methods. Specifically, we adopted CCA [9], Cluster CCA [21], fast multi-label CCA (fast-mlCCA) [14] which is an extended version of mlCCA, supervised locality preserving CCA (SLPCCA) [22] and supervised multi-view CCA (sMVCCA) [15] for fusion of two sets of CNN features. In addition, we adopted multiset CCA (MCCA) [23], MVCCA [24], Laplacian multiset canonical correlations (LapMCCs) [12], graph regularized multiset canonical correlations (GrMCCs) [11] and multi-view discriminant analysis (MvDA) [25] for fusion of three sets of CNN features. Note that since LapMCCs and GrMCCs are approaches for the single-label classification problem, samples with the same multi-label combination are considered to be the same class in this experiment. For comparison of CCA-based methods in the multi-label problem, replacing the multi-label annotation with the single-label annotation is a general approach [26]. SLPCCA also deal with the single-label problem. Specifically, SLPCCA is trained in such a way that within-class samples become minimum and between-class samples become maximum. Thus, in this experiment, SLPCCA can be applied to the multi-label problem by defining the above class information based on a similarity between label vectors. Consequently, we can apply these methods to the multi-label problem.

3.4. Evaluation Indices

We adopted five evaluation indices, exact matching rates (EMR), accuracy (ACC), hamming loss (Ham loss), macro averaged F-measure (Macro-F) and micro averaged F-measure (Micro-F). EMR is a ratio of correctly classifying all the labels of a sample. ACC is a ratio of the total number of correctly classified labels over all samples. Ham loss evaluates how many times, on average, an example-label pair is misclassified. The lower the value, the better the classification performance is. Micro-F aggregates true positives/negatives and false positives/negatives over labels and is calculated from them. Macro-F is calculated for each label and takes the average over labels. The details of these indices are shown in [27].

3.5. Parameter Settings

We determined parameters in such a way that each method outputs the best classification performance by using the validation dataset. Specifically, the searching range of each parameter is shown as follows: the number of hidden neurons of ELM: $\{300, 500, 700, 1000, 1500\}$, the threshold value for calculating the final class labels: $\{0.01, 0.015, ..., 0.1\}$, the number of clusters of Cluster CCA: $\{50, 100, ..., 250\}$, the number of neighbors of SLPCCA: $\{100, 1100, ..., 9100\}$ and those of LapMCCs and GrMCCs: $\{10, 60, ..., 260\}$. Since the dimension $d_p \leq \min(d_1, ..., d_M, C)$, if methods such as sMVCCA and sM2CP use class label vectors as one modality, we searched d_p from $\{15, 18, 21, 24\}$. Otherwise, we searched d_p from $\{100, 200, 500, 800\}$.

Table 1. Comparison between sM2CP and the comparative methods. This table shows results of fusion of DenseNet201 (Dense) [3] and Inception-ResNet-v2 (Inception) [4].

Methods	EMR	ACC	Ham loss	Macro-F	Micro-F
Dense [3]	0.317	0.611	0.0802	0.606	0.728
Inception [4]	0.305	0.591	0.0852	0.625	0.709
CCA [9]	0.303	0.610	0.0836	0.646	0.731
Cluster CCA [21]	0.196	0.379	0.155	0.385	0.527
fast-mlCCA [14]	**0.355**	0.653	0.0728	0.669	0.761
SLPCCA [22]	0.304	0.610	0.0830	0.641	0.731
sMVCCA [15]	0.323	0.631	0.0790	0.681	0.746
sM2CP (Ours)	**0.355**	**0.666**	**0.0703**	**0.706**	**0.771**

Table 2. Comparison between sM2CP and the comparative methods. This table shows results of fusion of Inception-ResNet-v2 (Inception) [4] and ResNet50 [19].

Methods	EMR	ACC	Ham loss	Macro-F	Micro-F
Inception [4]	0.305	0.591	0.0852	0.625	0.709
ResNet50 [19]	0.311	0.594	0.0850	0.595	0.712
CCA [9]	0.301	0.602	0.0854	0.632	0.723
Cluster CCA [21]	0.189	0.359	0.162	0.355	0.507
fast-mlCCA [14]	0.349	0.644	0.0748	0.659	0.752
SLPCCA [22]	0.299	0.603	0.0849	0.639	0.725
sMVCCA [15]	0.317	0.620	0.0817	0.669	0.737
sM2CP (Ours)	**0.356**	**0.656**	**0.0725**	**0.685**	**0.763**

Note that a parameter dealing with similarity of SLPCCA was experimentally set to 0.2. We experimentally set the regularization parameter γ in our method to 0.01. The activation function of ELM is a sigmoid function.

3.6. Performance Evaluation

Tables 1, 2 and 3 show the classification results of fusing two CNN features. Table 4 shows the classification results of fusing all of the three CNN features. From these tables, the effectiveness of sM2CP is verified since almost all indices among EMR, ACC, Macro-F and Micro-F of the proposed method are higher than those of all comparative methods. Furthermore, since Ham loss of sM2CP is lower than those of comparative method, sM2CP is better than the other comparative methods.

Firstly, we discuss these results obtained from Tables 1, 2 and 3. Although the locality preserving approach is generally effective for fusion of other modalities' features such as visual and text features [26], the performance of CCA are close to those of SLPCCA in this experiment. Thus, it is suggested that the locality preserving approach is not suitable for fusion of CNN features. However, since the proposed method outperforms these methods, sM2CP can perform the effective feature fusion of CNN features. By comparing sM2CP with fast-mlCCA, we can confirm that the performance of sM2CP is higher than that of fast-mlCCA which is a generic and strong method for the multi-label problem. Furthermore, by comparing sM2CP with sMVCCA, the effectiveness of introducing the multi-label similarity information into a supervised multi-view approach is verified.

Secondly, we discuss the results shown in Table 4. As with the results from Tables 1, 2 and 3, since the performance of sM2CP is higher than that of sMVCCA, we can confirm that sM2CP can also calculate the optimal projection performing feature fusion of multiple CNN features. Furthermore, by comparing sM2CP with

Table 3. Comparison between sM2CP and the comparative methods. This table shows results of fusion of DenseNet201 (Dense) [3] and ResNet50 [19].

Methods	EMR	ACC	Ham loss	Macro-F	Micro-F
Dense [3]	0.317	0.611	0.0802	0.606	0.728
ResNet50 [19]	0.311	0.594	0.0850	0.595	0.712
CCA [9]	0.291	0.593	0.0881	0.601	0.718
Cluster CCA [21]	0.186	0.354	0.163	0.340	0.501
fast-mlCCA [14]	**0.354**	0.642	0.0740	0.641	0.750
SLPCCA [22]	0.292	0.594	0.0871	0.600	0.718
sMVCCA [15]	0.306	0.612	0.0841	0.648	0.731
sM2CP (Ours)	0.348	**0.653**	**0.0707**	**0.675**	**0.762**

Table 4. Comparison between sM2CP and the comparative methods. This table shows results of fusion of DenseNet201, Inception-ResNet-v2 and ResNet50.

Methods	EMR	ACC	Ham loss	Macro-F	Micro-F
MCCA [23]	0.304	0.614	0.0826	0.638	0.734
MVCCA [24]	0.301	0.599	0.0869	0.595	0.720
LapMCCs [12]	0.273	0.551	0.0946	0.563	0.684
GrMCCs [11]	0.310	0.620	0.0811	0.654	0.740
MvDA [25]	0.256	0.556	0.0986	0.561	0.692
sMVCCA [15]	0.356	0.662	0.0715	0.697	0.768
sM2CP (Ours)	**0.357**	**0.668**	**0.0698**	**0.706**	**0.773**

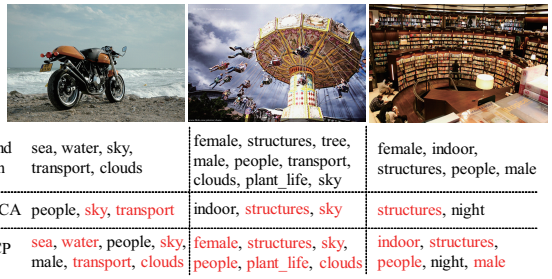


Fig. 3. Examples of multi-label classification results obtained from the experiments in Table 4. Correctly classified labels are marked in red.

LapMCCs and GrMCCs which are extended version of MCCA, it is verified that the label similarity information is more useful than local structure, discriminative and intrinsic geometrical structure for solving the multi-label problem. Moreover, as shown in Fig. 3, sM2CP can extract higher semantic meanings than sMVCCA. Although the central image includes "tree" as ground truth, it is considered that "tree" is not representative meanings of the image. Since sM2CP does not estimate "tree", sM2CP can effectively extract meanings by considering objects in images. Consequently, since sM2CP outperforms several CCA-based methods, the optimal fusion of multiple CNN features for the multi-label problem can be realized via sM2CP.

4. CONCLUSIONS

In this paper, we have presented multi-feature fusion based on sM2CP. sM2CP can perform effective feature fusion of CNN features and deal with the multi-label problem. Specifically, sM2CP introduces the label similarity information into the objective function of sMVCCA. This is the biggest advantage of sM2CP. Consequently, the experimental results show the effectiveness of sM2CP, which enables the successful fusion of the multiple CNN features for the multi-label problem.

5. REFERENCES

[1] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in neural information processing systems*, 2012, pp. 1097–1105.

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[4] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conference on Artificial Intelligence*, 2017, pp. 4278–4284.

[5] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, pp. 1–40, 2016.

[6] Y. Gao and K. M. Mosalam, "Deep transfer learning for image-based structural damage recognition," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 9, pp. 748–768, 2018.

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

[8] Z. Ren, Y. Deng, and Q. Dai, "Local visual feature fusion via maximum margin multimodal deep neural network," *Neurocomputing*, vol. 175, pp. 427–432, 2016.

[9] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[10] W. Zuobin, M. Kezhi, and G.-W. Ng, "Feature regrouping for cca-based feature fusion and extraction through normalized cut," in *Proc. International Conference on Information Fusion*, 2018, pp. 2275–2282.

[11] Y.-H. Yuan and Q.-S. Sun, "Graph regularized multiset canonical correlations with applications to joint feature extraction," *Pattern Recognition*, vol. 47, no. 12, pp. 3907–3919, 2014.

[12] Y.-H. Yuan, Y. Li, X.-B. Shen, Q.-S. Sun, and J.-L. Yang, "Laplacian multiset canonical correlations for multiview feature extraction and image recognition," *Multimedia Tools and Applications*, vol. 76, no. 1, pp. 731–755, 2017.

[13] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.

[14] V. Ranjan, N. Rasiwasia, and C. Jawahar, "Multi-label cross-modal retrieval," in *Proc. IEEE International Conference on Computer Vision*, 2015, pp. 4094–4102.

[15] G. Lee, A. Singanamalli, H. Wang, M. D. Feldman, S. R. Master, N. N. Shih, E. Spangler, T. Rebbeck, J. E. Tomaszewski *et al.*, "Supervised multi-view canonical correlation analysis (smvcca): integrating histologic and proteomic features for predicting recurrent prostate cancer," *IEEE Trans. Medical Imaging*, vol. 34, no. 1, pp. 284–297, 2015.

[16] H. Wang, C. Ding, and H. Huang, "Multi-label linear discriminant analysis," in *Proc. European Conference on Computer Vision*, 2010, pp. 126–139.

[17] T.-K. Sun, S.-C. Chen, Z. Jin, and J.-Y. Yang, "Kernelized discriminative canonical correlation analysis," in *Proc. International Conference on Wavelet Analysis and Pattern Recognition*, vol. 3, 2007, pp. 1283–1287.

[18] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proc. ACM International Conference on Multimedia Information Retrieval*, 2008, pp. 39–43.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[20] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proc. IEEE International Joint Conference on Neural Networks*, vol. 2, 2004, pp. 985–990.

[21] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2014, pp. 823–831.

[22] J. Yang and X. Zhang, "Feature-level fusion of fingerprint and finger-vein for personal identification," *Pattern Recognition Letters*, vol. 33, no. 5, pp. 623–628, 2012.

[23] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Trans. image processing*, vol. 11, no. 3, pp. 293–305, 2002.

[24] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Proc. Conference on Data Mining and Data Warehouses*, 2010, pp. 1–4.

[25] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 188–194, 2016.

[26] Y. Hua, J. Du, Y. Zhu, and P. Shi, "Semantics and locality preserving correlation projections," in *Proc. IEEE International Conference on Multimedia and Expo*, 2017, pp. 913–918.

[27] E. Gibaja and S. Ventura, "A tutorial on multilabel learning," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, p. 52, 2015.