Noname manuscript No.
(will be inserted by the editor)

Deterioration level estimation via neural network maximizing category-based ordinally supervised multi-view canonical correlation

Keisuke Maeda · Sho Takahashi · Takahiro Ogawa · Miki Haseyama

Received: date / Accepted: date

Abstract A deterioration level estimation method via neural network maximizing category-based ordinally supervised multi-view canonical correlation is presented in this paper. This paper focuses on real world data such as industrial applications and has two contributions. First, a novel neural network handling multi-modal features transforms original features into features effectively representing deterioration levels in transmission towers, which are one of the infrastructures, with consideration of only correlation maximization. It can be realized by setting projection matrices maximizing correlations between multiple features into weights of hidden layers. That is, since the proposed network has only a few hidden layers, it can be trained from a small amount of training data. Second, since there exist diverse characteristics and an ordinal scale in deterioration levels, the proposed method newly derives category-based ordinally supervised multi-view canonical correlation analysis (Co-sMVCCA). Co-sMVCCA enables estimation of effective projection considering both within-class divergence and the ordinal scale between classes. Experimental results showed that the proposed method realizes accurate deterioration level estimation.

Keywords

Neural network, within-class divergence, ordinal scale, canonical correlation, deterioration level estimation.

K. Maeda † · S. Takahashi ‡ · T. Ogawa †† · M. Haseyama † †
† Office of Institutional Research, Hokkaido University, Sapporo, Japan
Tel.: +81-11-706-6078
‡ Faculty of Engineering, Hokkaido University, Sapporo, Japan
†† Faculty of Information Science and Technology, Hokkaido University, Sapporo, Japan
Tel.: +81-11-706-6078
E-mail: † maeda@lmd.ist.hokudai.ac.jp
†† ogawa@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp
‡ stakahashi@eng.hokudai.ac.jp

1 Introduction

With the development of hardware devices and advent of the big data era, convolutional neural networks (CNNs) have been effectively trained by using a large-scale dataset [10,45] such as ImageNet and have achieved accurate image classification [12]. In the field of information science, although many researchers have focused on large-scale datasets, many recent studies have been conducted by using not images for generic object recognition but real data such as agricultural images [20], medical images [23] and images for infrastructure management [24] in order to efficiently support experts in several fields. In studies using images for infrastructure management, automatic detection of specific distresses [4] such as potholes and automatic deterioration level estimation [39] have attracted much attention. Since human errors often occur in manual deterioration level estimation due to ambiguity in the decisions of inspectors, there is an urgent need to realize automatic and quantitative analysis of the levels by using statistical approaches and machine learning technologies [38].

In image classification and object recognition targeting these real data, since preparation of a large number of training images is difficult, not full-scratch CNNs but some transfer learning-based approaches have been adopted [8, 9]. Specifically, transfer learning includes fine-tuning and CNN-based feature extraction. In order to fine-tune CNNs for a target domain, parameters optimized by using a large-scale dataset, which is often called a source domain, are used as initial values of the network parameters. On the other hand, CNN-based features calculated from an intermediate layer of CNNs pre-trained by the large-scale dataset are also useful for several tasks. Unfortunately, there is a limitation to improvement of the performance of these transfer learning-based approaches due to the following two problems.

- Since visual characteristics of real data such as images for infrastructure management (distress images) are more different than those of images used for generic object detection, simple transfer learning frameworks have difficulty in extracting discriminant features for deterioration level estimation [6].
- Estimation of deterioration levels is not a basic classification task. Concretely, there exists an ordinal scale between deterioration levels, and diverse visual characteristics exist within the same levels. Therefore, in order to obtain discriminant ability, consideration of the ordinal scale between classes and the within-class divergence is essential.

To solve the above problems, construction of a trainable framework from a small number of training images that can calculate high discriminant features with consideration of the ordinal scale and the divergence is needed.

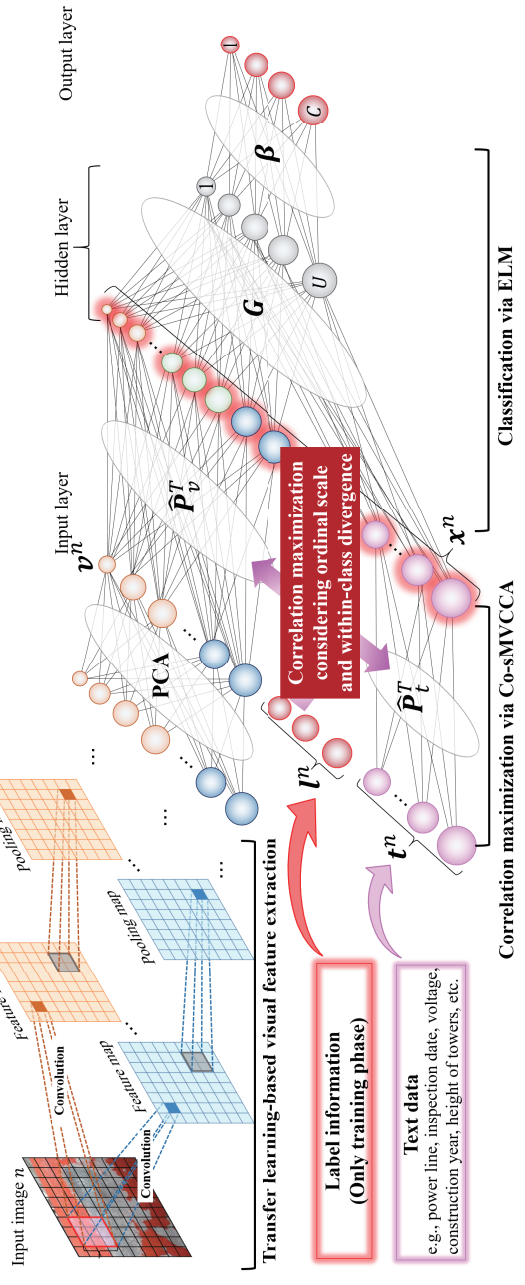


Fig. 1 Overview of our novel neural network. The proposed network has three procedures. First, we extract visual features via simple transfer learning of the CNN and calculate text features from text data. Second, we maximize the canonical correlation between these heterogeneous features via Co-sMVCCA. Finally, we construct an extreme learning machine (ELM) [16], which is a feedforward neural network-based classifier.

In many applications, there exist not only image data but also their corresponding metadata. For example, in a web application such as twitter, there are images, text description and movies uploaded by users. Moreover, multi-modal data have also been widely used in RGB and depth saliency fields [7,46]. Also, in the field of infrastructure maintenance, engineers not only take distress images but also record their corresponding text data including supplemental information related to deterioration levels such as materials of structures and distress locations when they manually decide the deterioration levels [25]. From the above, the multi-modal information, distress images and text data, becomes effective evidence for manual judgment in the actual maintenance inspection. Several multi-modal approaches for maintenance inspection of infrastructures have been proposed. For example, Im et al. proposed a method for crack direction detection by using visual and audio information obtained by sensors [17], and Kasahara et al. proposed an unsupervised learning approach for automation of a hammering test [21]. Since determination of the existence of a defect becomes feasible by using sounds returned after a hammer strike on a structure's surface, they collaboratively used audio and position information. However, methods target only one distress such as a crack. On the other hand, there exist some multi-modal methods corresponding to more complex tasks such as classification of multiple types of distress and classification of road surface conditions. In [27], by using images and text information recorded in the inspection, multiple types of distress including cracks were classified on the basis of deep learning. Furthermore, Jonsson et al. performed spectral analysis by using infrared images that enabled classification of area segments of weather-related road surface conditions such as icy, wet, or snowy, that is, they treated with multiple wavelength data [19]. In addition, the method in [2] enabled prediction of the quality of a road using a triaxial accelerometer and a gyroscope. Furthermore, they created a real-time android application called "RoadSense" and that is useful for road manager to evaluate the states of their road networks. In analysis of transmission towers, which this paper focuses on, the method proposed in [37] used an unmanned aerial vehicle and images recorded by the vehicle. Detection of regions of transmission towers was realized by object detection approaches, Faster R-CNN [32] and Yolo-v3 [31]. Although several uni-modal approaches [36] for deterioration level estimation based on only visual features such as color information obtained from images have been proposed, our approach proposed in this work and [28] is the first work focusing on multi-modal information for deterioration level estimation in transmission towers.

It has been reported in the fields of multi-modal signal processing that consideration of the canonical correlation between heterogeneous features obtained from multi-modal information enables calculation of features with high discriminant ability [40]. Furthermore, in the case of treatment with a small number of training images, it has been verified that a neural network maximizing canonical correlation was more effective than feature transformation by using many hidden layers such as general deep learning approaches [26]. Therefore, we can realize enhancement of the discriminant ability by constructing a neural network that can train projections maximizing canonical correlation with consideration of both the ordinal scale and the within-class divergence.

In this paper, we propose a novel neural network maximizing category-based ordinally supervised multi-view canonical correlation. Figure 1 shows an overview of

Table 1 Example of text data. Distress images taken by inspectors have inspection records such as office: A and transmission line: C.

Inspection item	Inspection record	Num. of dimension
Office	A, B, ...	D_1
Transmission line	C, D, ...	D_2
Area	E, F, ...	D_3
Salt damage	A1, A2, ...	D_4
Type of towers	Angle towers, pipe towers, ...	D_5
Voltage (kV)	275, 66, ...	1
Height of towers (m)	120, 50, ...	1
Inspection date	2006, 1995, ...	1
Coating year	2005, 1980, ...	1
Latitude	34.8, 35.0, ...	1
Longitude	139.0, 138.9, ...	1
Sum	-	$d_t = \sum_{q=1}^{5} D_q + 6$

the proposed neural network which consists of the following three procedures: feature extraction, correlation maximization via category-based ordinally supervised multi-view canonical correlation (Co-sMVCCA) and neural network-based classification. The contributions of our method are twofold.

- In order to tackle the first issue, we extract CNN-based visual features via the simplest transfer learning and text features from text data. Furthermore, we estimate projection matrices that maximize canonical correlation between heterogeneous features. The derivation of features with high discriminant ability becomes feasible by setting the projection matrices to the intermediate layer of our novel network.
- In order to tackle the second issue, we derive a novel canonical correlation technique that can consider the ordinal scale between classes and the within-class divergence. Specifically, we newly derive Co-sMVCCA, which is an extended version of sMVCCA [22], in order to obtain discriminant features for deterioration level estimation. Co-sMVCCA can estimate the optimal projection matrices that can transform original features to high discriminant features since it not only adds a term considering the ordinal scale between classes to the objective function of our CCA but also calculates covariance matrices between samples belonging to the same classes.

In summary, construction of a neural network realizing effective transformation from a small amount of training data by maximizing canonical correlation is the first contribution of this paper. Derivation of Co-sMVCCA, which can consider the ordinal scale and the within-class divergence is the second contribution. The proposed neural network with a novel CCA-based approach, Co-sMVCCA, realizes accurate deterioration level estimation.

The rest of the paper is organized as follows. First, an explanation of the proposed method is presented in Section 2. In Section 3, we show experimental results. Finally, we show conclusions in Section 4.

2 Neural Network Maximizing Category-based Ordinally Supervised Multi-view Canonical Correlation

The neural network maximizing category-based ordinally supervised multi-view canonical correlation is shown in this section. As shown in Fig. 1, the proposed method consists of three procedures: heterogeneous feature extraction (2.1), maximization of canonical correlation (2.2) and classification (2.3).

2.1 Heterogeneous Feature Extraction

In this subsection, we explain extraction of visual features via transfer learning of CNNs and extraction of text features from recorded text data. Extraction of class label features used in the training phase for calculation of the projection matrices is also explained. Since the number of samples of distress images is small as mentioned in Section 1, we use pre-trained CNN-based visual features. In fact, detection and classification problems in the field of infrastructure management have often been solved by the use of CNN features pre-trained by using a large-scale dataset for generic object recognition [8]. Furthermore, in the detection of myocardial infarction using ECG images [1], fine-tuning of the CNN model and output features from the middle layer of the CNN model have been used when extracting visual features from the images. Thus, the feature calculation in studies using real data is considered to be valid. Given training images n ($n = 1, 2, ..., N$; N being the number of training images), we input them into a CNN model pre-trained by using a large-scale dataset, ImageNet. We extract visual features from an average pooling layer of the CNN model. In addition, we obtain visual features $\boldsymbol{v}^n \in \mathbb{R}^{d_v}$ by applying a simple dimension reduction approach, principal component analysis (PCA), to the obtained features in order to prevent over-fitting. Note that application of PCA to CNN features is generally used for dimension reduction [14, 41].

Next, given a training image n, we calculate text features $\boldsymbol{t}^n \in \mathbb{R}^{d_t}$ ($d_t = \sum_{q=1}^{5} D_q + 6$) from text information that the image has. There are various kinds of variables in text information as shown in Table 1, and we explain our encoding approach below. When we calculate text features from nominal qualitative variables such as office and transmission line, we extract one hot vector for each inspection item. Note that we regard office, transmission line, area, salt damage and type of tower as nominal qualitative variables in the proposed method. Thus, we calculate D_q-dimensional binary features from the q ($= 1, 2, ..., 5$) th record. Specifically, an element corresponding to the inspection record that the image n has becomes one. Otherwise, non-corresponding $D_q - 1$ elements become zero. On the other hand, when continuous numeric variables are given, we directly use their original values without specific processing. Then we extract six-dimensional features from numeric variables. Therefore, the total dimension of text features d_t is $\sum_{q=1}^{5} D_q + 6$.

Furthermore, in only the training phase, we also calculate class label features $\boldsymbol{l}^n \in \mathbb{R}^C$. Note that C is the number of class labels. The class label features consist of binary values, and an element corresponding to their own class is one. The other

elements become zero. Consequently, Co-sMVCCA adopts three modalities, visual, text and class label features.

CNN-based features are often used in the field of infrastructure management. However, detection performance and classification performance are limited due to the simplest transfer learning using only CNN-based features. Specifically, since these methods extract CNN-based features from distress images and perform classification, the performance depends on the representation ability of visual features, but there is a limitation of the potential of visual features calculated from the pre-trained network. Therefore, in order to obtain discriminant features, multi-modal feature integration via the proposed network including canonical correlation maximization is effective for deterioration level estimation.

2.2 Co-sMVCCA-based Correlation Maximization

Derivation of Co-sMVCCA, which can consider the ordinal scale and class information, for correlation maximization is shown in this subsection. Co-sMVCCA has the following two strong points: (i) dealing with continuously varying deterioration levels and (ii) considering the within-class divergence. Co-sMVCCA estimates the optimal projection $\boldsymbol{p}_k \in \mathbb{R}^{d_k}$. Note that $k \in \{v, t, l\}$ represents the modality. In Co-sMVCCA, we can integrate these heterogeneous features by maximizing the following objective function:

$$\arg\max_{\boldsymbol{p}_v, \boldsymbol{p}_t, \boldsymbol{p}_l} \sum_{k_1 \in \{v,t,l\}} \sum_{k_2 \in \{v,t,l\},\ k_2 \neq k_1} \frac{\boldsymbol{p}_{k_1}^\top \overline{\boldsymbol{C}}_{k_1,k_2} \boldsymbol{p}_{k_2}}{\sqrt{\boldsymbol{p}_{k_1}^\top \underline{\boldsymbol{C}}_{k_1,k_1} \boldsymbol{p}_{k_1}} \sqrt{\boldsymbol{p}_{k_2}^\top \underline{\boldsymbol{C}}_{k_2,k_2} \boldsymbol{p}_{k_2}}}, \tag{1}$$

where $\overline{\boldsymbol{C}}_{k_1,k_2}$ is a covariance matrix considering the ordinal scale and the within-class divergence between the modalities k_1 and k_2. Its details are shown below. In addition, $\underline{\boldsymbol{C}}_{k_1,k_2}$ is the covariance matrix between the modalities k_1 and k_2. Since the solution of the optimization problem does not depend on the scale of $\boldsymbol{p}_{k_1}$, Eq. (1) can be rewritten as follows:

$$\begin{aligned} &\arg\max_{\boldsymbol{p}_v, \boldsymbol{p}_t, \boldsymbol{p}_l} \sum_{k_1 \in \{v,t,l\}} \sum_{k_2 \in \{v,t,l\}, k_2 \neq k_1} \boldsymbol{p}_{k_1}^\top \overline{\boldsymbol{C}}_{k_1,k_2} \boldsymbol{p}_{k_2}. \\ &\quad \text{s.t.} \quad \boldsymbol{p}_{k_1}^\top \underline{\boldsymbol{C}}_{k_1,k_1} \boldsymbol{p}_{k_1} = 1 \ \ (k_1 \in \{v, t, l\}) \end{aligned} \tag{2}$$

In our method, we define $\boldsymbol{P} = [\boldsymbol{P}_v^\top, \boldsymbol{P}_t^\top, \boldsymbol{P}_l^\top]^\top \in \mathbb{R}^{(d_v+d_t+C)\times(d_p\times 3)}$. Note that d_p ($\leq \min(d_v, d_t, C)$) represents the dimension of the transformed features via Co-sMVCCA. Then Eq. (2) can be rewritten as

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

where

$$\overline{\boldsymbol{C}} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{C}_{VT} & \boldsymbol{C}_{VL} \\ \boldsymbol{C}_{TV} & \boldsymbol{0} & \boldsymbol{C}_{TL} \\ \boldsymbol{C}_{LV} & \boldsymbol{C}_{LT} & \boldsymbol{0} \end{bmatrix}, \tag{4}$$

$$\underline{\boldsymbol{C}} = \begin{bmatrix} \boldsymbol{V}\boldsymbol{V}^\top & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T}\boldsymbol{T}^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{L}\boldsymbol{L}^\top \end{bmatrix}. \tag{5}$$

Note that $\boldsymbol{C}_{VT} = \boldsymbol{C}_{TV}$, $\boldsymbol{C}_{VL} = \boldsymbol{C}_{LV}$ and $\boldsymbol{C}_{TL} = \boldsymbol{C}_{LT}$. In addition, we use mean-normalized feature matrices, $\boldsymbol{V} = [\boldsymbol{v}^1, \boldsymbol{v}^2, \cdots, \boldsymbol{v}^N]$, $\boldsymbol{T} = [\boldsymbol{t}^1, \boldsymbol{t}^2, \cdots, \boldsymbol{t}^N]$ and $\boldsymbol{L} = [\boldsymbol{l}^1, \boldsymbol{l}^2, \cdots, \boldsymbol{l}^N]$, for solving the above problem. In general CCA-based methods, the non-diagonal element of $\overline{\boldsymbol{C}}$ is simply calculated as general covariance $E_n(\boldsymbol{v}^n \boldsymbol{t}^{n\top})$, where $E_n(\cdot)$ performs the average. However, a general covariance matrix cannot consider both the ordinal scale of continuously varying deterioration levels and the divergence of samples belonging to the same label. We therefore apply a weight matrix $\boldsymbol{W}$ dealing with the ordinal scale to the non-diagonal covariance matrices of $\overline{\boldsymbol{C}}$. From the above, in order to deal with the ordinal scale, we define $\boldsymbol{C}_{VL}$ and $\boldsymbol{C}_{TL}$ as follows:

$$\boldsymbol{C}_{VL} = \boldsymbol{V}\boldsymbol{W}\boldsymbol{L}^\top, \tag{6}$$

$$\boldsymbol{C}_{TL} = \boldsymbol{T}\boldsymbol{W}\boldsymbol{L}^\top, \tag{7}$$

where

$$\boldsymbol{W}_{n_1,n_2} = \max(0, 1 - |c_{n_1} - c_{n_2}|/\epsilon), \tag{8}$$

where $\boldsymbol{W}_{n_1,n_2}$ represents the (n_1, n_2) th element of $\boldsymbol{W}$, and ϵ is a parameter. Furthermore, $c_{n_1} \in \{1, 2, ..., C\}$ is a class label (deterioration level) of the training image n_1. In addition, Eqs. (6) and (7) can be formulated due to the inspiration of the definition of the within-class covariance matrix of discriminative CCA (DCCA) [34]. The weight matrix with a size of $N \times N$ for construction of the within-class covariance matrix of DCCA consists of binary values. Specifically, each element of the weight matrix corresponding to the same class label becomes one, and the other elements become zero. Then, although DCCA can effectively consider the covariance between samples belonging to the same class, the weight matrix $\boldsymbol{W}$ of Co-sMVCCA affects the covariance matrix between two sets of samples when deterioration levels of the samples are similar. That is, since Co-sMVCCA sets fuzzy weights according to the distance between the levels as shown in Eq. (8), the matrix $\overline{\boldsymbol{C}}$ can be effectively obtained.

Furthermore, although these covariance matrices between class label features and other features can consider label information, the covariance matrix $\boldsymbol{C}_{VT}$ between visual and text features cannot consider label information. Thus, we adopt a novel category-based covariance matrix inspired by [42]. Specifically, in order to consider the within-class divergence, we newly derive the category-based covariance matrix $\boldsymbol{C}_{VT}$ as follows:

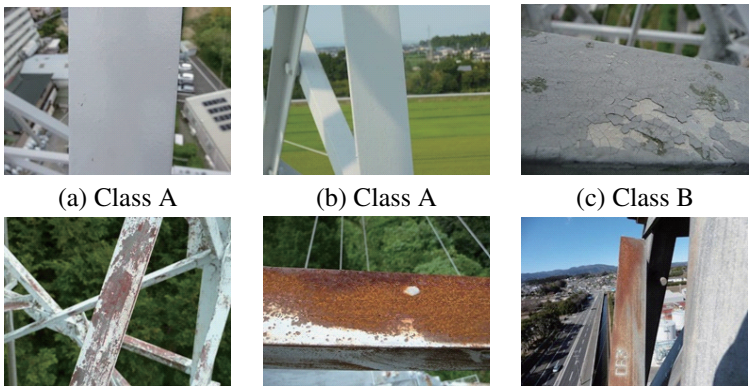
$$\boldsymbol{C}_{VT} = \beta E_{c \subset C}(\boldsymbol{C}^s_{VT}(c)) + (1 - \beta) E_{c \subset C}(\boldsymbol{C}^t_{VT}(c)), \tag{9}$$

where

$$\boldsymbol{C}^s_{VT}(c) = E_{n \in c}(\boldsymbol{v}^n \boldsymbol{W}(n, n) \boldsymbol{t}^{n\top}), \tag{10}$$

$$\boldsymbol{C}^t_{VT}(c) = E_{n_1, n_2 \in c, n_1 \neq n_2}(\boldsymbol{v}^{n_1} \boldsymbol{W}(n_1, n_2) \boldsymbol{t}^{n_2\top}). \tag{11}$$

Note that β $(0 \leq \beta \leq 1)$ is used as a parameter. Although general CCA-based methods adopt a covariance matrix between multiple features calculated from the same



(a) Class A (b) Class A (c) Class B

(d) Class B (e) Class C (f) Class C

Fig. 2 Examples of images [11] used in the proposed method.

sample, Co-sMVCCA focuses on the covariance matrix of visual and text features from different samples belonging to the same class label in order to consider the within-class divergence in Eq. (11). Furthermore, Eqs. (10) and (11) also include the weight matrix $\boldsymbol{W}$. Thus, Co-sMVCCA can deal with the ordinal scale between levels and the within-class divergence. Therefore, it is expected that the definition provides the projection matrix of effective feature transformation.

Finally, we solve the following generalized eigenvalue problem:

$$\overline{\boldsymbol{C}}\boldsymbol{P} = \lambda(\underline{\boldsymbol{C}} + \gamma\boldsymbol{I})\boldsymbol{P}, \tag{12}$$

where γ is a regularization parameter. The optimal projection $\hat{\boldsymbol{P}}_k \in \mathbb{R}^{d_k \times d_p}$ for the feature transformation is obtained by solving the above problem. The matrix $\hat{\boldsymbol{P}}_k$ is constructed by using the eigenvectors of the d_p-largest eigenvalues. Then we can calculate the projected features as follows:

$$\boldsymbol{X}_v = \hat{\boldsymbol{P}}_v^\top \boldsymbol{V} \in \mathbb{R}^{d_p \times N}, \tag{13}$$

$$\boldsymbol{X}_t = \hat{\boldsymbol{P}}_t^\top \boldsymbol{T} \in \mathbb{R}^{d_p \times N}, \tag{14}$$

where $\boldsymbol{X}_v = [\boldsymbol{x}_v^1, \boldsymbol{x}_v^2, ..., \boldsymbol{x}_v^N]$ and $\boldsymbol{X}_t = [\boldsymbol{x}_t^1, \boldsymbol{x}_t^2, ..., \boldsymbol{x}_t^N]$. Then we can obtain the projected features $\boldsymbol{x}^n = [(\boldsymbol{x}_v^n)^\top, (\boldsymbol{x}_t^n)^\top]^\top$ of the n th sample. The calculated features are used for training of a classifier. Consequently, Co-sMVCCA enables estimation of the effective projection considering both the ordinal scale and the within-class divergence.

2.3 Classification via Extreme Learning Machine

Training of an Extreme Learning Machine (ELM) is explained in this subsection [16]. An ELM is one of the feedforward neural networks and can be trained from a small number of training images. Note that the architecture of an ELM is similar to that of

Algorithm 1 Correlation maximization via Co-sMVCCA.

Require: $\boldsymbol{V} \in \mathbb{R}^{d_v \times N}$, $\boldsymbol{T} \in \mathbb{R}^{d_t \times N}$ and $\boldsymbol{L} \in \mathbb{R}^{C \times N}$, mean-normalized feature matrices.

Ensure: $\hat{\boldsymbol{P}}_k \in \mathbb{R}^{d_k \times d_P}$, an optimal projection matrix of Co-sMVCCA.

$\boldsymbol{W}$ is obtained as Eq. (8).
$\boldsymbol{C}_{VL}$ and $\boldsymbol{C}_{TL}$ are obtained as Eqs. (6) and (7).
$\boldsymbol{C}^s_{VL}$ and $\boldsymbol{C}^t_{TL}$ are obtained as Eqs. (10) and (11).
$\boldsymbol{C}_{VT}$ is obtained as Eq. (9).
$\overline{\boldsymbol{C}}$ and $\underline{\boldsymbol{C}}$ are obtained as Eqs. (4) and (5).
$\hat{\boldsymbol{P}}_k$ is obtained by solving Eq. (12).
return $\hat{\boldsymbol{P}}_k$.

a random vector functional-link net [29, 33]. First, we apply a sigmoid function $\boldsymbol{G}$ as an activation function to the projected features $\boldsymbol{x}^n$ as follows:

$$z(\boldsymbol{x}^n) = [G(\boldsymbol{a}_1^\top \boldsymbol{x}^n + b_1), G(\boldsymbol{a}_2^\top \boldsymbol{x}^n + b_2), \cdots, G(\boldsymbol{a}_U^\top \boldsymbol{x}^n + b_U)]^\top, \tag{15}$$

where U is the number of nodes of a hidden layer. $\boldsymbol{a}_u$ and b_u $(u = 1, 2, \cdots, U)$ of the activation function G are parameters that are randomly determined on the basis of a uniform distribution. Furthermore, by minimizing the least square error between the hidden layer's outputs $\boldsymbol{\beta Z}$ and the class label matrix $\boldsymbol{L}$, the ELM estimates a weight matrix as follows:

$$\boldsymbol{\beta} = \boldsymbol{L}\boldsymbol{Z}^\dagger, \tag{16}$$

where $\boldsymbol{Z} = [z(\boldsymbol{x}^1), z(\boldsymbol{x}^2), \cdots, z(\boldsymbol{x}^N)]$ is the hidden layer's output matrix, and $\boldsymbol{Z}^\dagger$ means the Moore-Penrose generalized inverse of $\boldsymbol{Z}$. Algorithms 1 and 2 show the algorithmic steps of the proposed method. As shown in these tables, it is confirmed that the proposed method can be calculated at low computational cost, and our approach is therefore suitable for real data analysis, for which preparation of a large number of images is difficult.

In the test phase, given a test vector $\boldsymbol{x}$, we obtain the output vector $\boldsymbol{y} = \boldsymbol{\beta} z(\boldsymbol{x})$. From the above, since parameters of the activation function are determined via random values and the weight matrix $\boldsymbol{\beta}$ is determined uniquely due to the least square estimation, the ELM can be effectively trained from a small amount of training data. The proposed method can effectively transform original features to high discriminant features instead of constructing multiple hidden layers adopted in general deep learning methods by setting the canonical correlation-based projection matrices into the hidden layer's weights. Thus, training can be performed from a small number of training images. Therefore, our novel neural network including correlation maximization via Co-sMVCCA can extract more discriminant features and accurately classify deterioration levels.

Algorithm 2 Construction of the ELM classifier.

Require: $\boldsymbol{V} \in \mathbb{R}^{d_v \times N}$, $\boldsymbol{T} \in \mathbb{R}^{d_t \times N}$, $\hat{\boldsymbol{P}}_v \in \mathbb{R}^{d_v \times d_P}$ and $\hat{\boldsymbol{P}}_t \in \mathbb{R}^{d_t \times d_P}$.
Ensure: $\boldsymbol{\beta} \in \mathbb{R}^{C \times U}$, a weight matrix of ELM.
$\boldsymbol{a}$ and b are randomly determined for activation function.
$\boldsymbol{X}_v$ and $\boldsymbol{X}_t$ are obtained as Eqs. (13) and (14).
$\boldsymbol{Z}$ which is output of a hidden layer is obtained as Eq. (15).
$\boldsymbol{\beta}$ is obtained as Eq. (16).
return $\boldsymbol{\beta}$.

3 Experimental Results

The effectiveness of the proposed method is verified in this section. The experimental conditions are explained in subsection 3.1 and evaluation of the performance of our method is described in subsection 3.2.

3.1 Experimental Conditions

In the proposed method, we used a dataset provided by Tokyo Electric Power Company Research Institute. This dataset includes distress images taken by inspectors and text data recorded by them during actual maintenance inspections. There are three deterioration levels in distress images, classes A, B and C, and each image belongs to one class. Note that there is an order between classes, e.g., class C means dangerous and class A means safe. The number of images in the dataset is very small. The numbers of images belonging to classes A, B and C are 589, 775 and 391, respectively. We adopted 10-fold cross validation as the verification process. In each cross validation, we divided all of the data into test data and the remaining data in a ratio of 1:9, and we also divided the remaining data into validation data and training data in a ratio of 1:4.

Moreover, we used some CNN models for verifying the robustness of our method. We used Inception-ResNet-v2 [35], DenseNet-201 [15] and Xception [5] implemented in Keras, and we extracted the outputs of the middle layers of those models as transfer learning. In addition, in order to verify the effectiveness of our method, we used nine comparative methods shown in Tables 2-4. The details are shown below. In this experiment, we first used comparative methods constructed by using only text features or only visual features. "Only text features" means that we calculate text features $\boldsymbol{t}^n$ and train the ELM by inputting the obtained text features. On the other hand, "only visual features" means that we calculate visual features $\boldsymbol{v}^n$ and train ELM by inputting the obtained visual features. Furthermore, since it has been reported in [8, 9] that fine-tuning was often an effective approach when the number of training images was small, we used fine-tuned CNNs as comparative methods. In this experiment, CNNs were pre-trained by using ImageNet and were retrained by using our dataset. The above three comparative methods focus on uni-modal information. On the other hand, the other six comparative methods focus on multi-modal information. As a baseline feature transformation approach, we adopted general canonical correlation analysis

(CCA) [13]. CCA estimates projection matrices maximizing the canonical correlation between two sets of features. Furthermore, we used other CCA-based approaches maximizing the correlation between multi-modal features with consideration of geometrical information. Specifically, graph-regularized multiset canonical correlations (GrMCCs) [44] introduce a locality intra-view structure into the objective function of multiset CCA (MCCA), which can deal with multi-view information. Laplacian multiset canonical correlations (LapMCCs) [43] introduce a locality inter-view structure into the objective function of MCCA. Furthermore, we adopted linear discriminant multi-set canonical correlation analysis (LDMCCA) [30] as a comparative method. LDMCCA contains class information of the training data and represents the fused features more efficiently and discriminatively in some dimensions. Deep CCA [3], which includes deep learning-based feature learning, was also used as a comparative method. Finally, we used supervised multi-view CCA (sMVCCA) [22] and ordinally sMVCCA (OsMVCCA) [28], which is our previous method. sMVCCA can deal with multi-modal features and consider class label information by using class label features as one modality. OsMVCCA is an extended version of sMVCCA and introduces a term balancing the ordinal scale into the objective function of sMVCCA. We compared results obtained by the proposed method with the results obtained by using the nine comparative methods to verify the effectiveness of the proposed method. Note that in the CCA series of comparative methods, the middle layer maximizing the canonical correlation was used for constructing the multi-modal neural network.

In the experiments, the number of hidden nodes U was determined in such a way that our method achieved the best performance for the validation dataset. The searching range of U was $\{100, 200, ..., 1000\}$. The parameters ϵ were experimentally set to 0.01. β values used in Co-sMVCCA dealing with Inception-ResNet-v2, DenseNet-201 and Xception were 0.3, 0.3 and 0.5, respectively. By using Recall, Precision and F-measure, we evaluated the performance of our method.

3.2 Performance Evaluation

Tables 2-4 show the classification results based on Inception-ResNet-v2, DenseNet-201 and Xception, respectively. Focusing on the average of classification results, since several CCA-based methods dealing with multi-modal features including Co-sMVCCA in our method outperform uni-modal approaches, fine-tuning and "only visual features" and "only text features", it was confirmed that multi-modal approaches are effective. Since the proposed network is superior to fine-tuning, not CNN training a large number of hidden layers but a shallow neural network transforming to discriminant features via Co-sMVCCA is effective for constructing classifiers from a small number of training images. Thus, the effectiveness of our first contribution is confirmed.

Furthermore, our neural network including Co-sMVCCA is superior to other CCA-based methods, LapMCCs, which is one of the state-of-the-art methods, and GrMCCs, which is a standard method that can be used for several applications [18], and Deep CCA [3] performing non-linearity of a feature space. Thus, it is verified that the ordinal scale and the within-class divergence are more effective than considering geometrical structures and non-linearity. Moreover, Co-sMVCCA outperforms LDMCCA combining linear discriminant analysis and MCCA. Also, by comparing Co-sMVCCA with sMVCCA, these results indicate that introduction of the ordinal scale and within-class divergence into the objective function of sMVCCA is effective. Furthermore, by comparing Co-sMVCCA with OsMVCCA, the use of category-based covariance matrices is useful for performance improvement. Thus, the effectiveness of our second contribution is confirmed.

In this experiment, "only visual features", CCA, and sMVCCA are the benchmarking methods. Since "only visual features" are calculated from pre-trained models that are often used in real data analysis, the results show the baseline performance as uni-modal analysis. As shown in Tables 2, 3 and 4, it is confirmed that multi-modal analysis exceeds the performance limitation of uni-modal analysis since these methods with both visual and text features outperform the method with "only visual features". Next, CCA, which is a multi-modal approach, is often used as a baseline method. In some CNN models, it is confirmed that the accuracy of CCA exceeds that of GrMCCs and LapMCCs considering the geometrical structure of features. Generally, GrMCCs and LapMCCs can easily capture the structure in the case of using images such as those used in general object recognition since objects included in different classes are obviously different. However, the images of transmission towers used in the deterioration level estimation differ significantly from such general images; that is, there is only a slight difference between visual characteristics in images belonging to different deterioration levels. Thus, it is considered that it is difficult to reflect these structures via GrMCCs and LapMCCs. On the other hand, the direct use of class information has been reported to be effective in such real data analysis [22, 28], and sMVCCA is a benchmarking method. Since sMVCCA outperforms GrMCCs and LapMCCs, which consider the geometrical structure, and Deep CCA, which deals with the nonlinear structure, it is indicated that the use of class information as one view is effective. Although Deep CCA is a strong method among recent CCA-based approaches due to the deep learning-based approaches, it requires more training data than does sMVCCA in

order to sufficiently train its model. On the other hand, since sMVCCA can directly use class information as one modality, it can represent deterioration levels without a large amount of training data. Therefore, since sMVCCA is valid for real data analysis, we turned our attention to the derivation of an extended version of sMVCCA.

Confusion matrices of Co-sMVCCA, OsMVCCA and sMVCCA by using Inception-ResNetV2 are shown in Fig. 3. As shown in Fig. 3 (b) and (c), although the ratio of images that belong to class A predicted as class A via OsMVCCA is equal to that via sMVCCA, the misclassification ratio of images that belong to class A predicted as class C via OsMVCCA is lower than that via sMVCCA. That is, by introducing the ordinal scale, it became easy to estimate to a level near the target level. On the other hand, since the misclassification ratio of images that belong to class C predicted as class A via OsMVCCA is high, it is confirmed that consideration of only the ordinal scale is not sufficient to represent the deterioration levels. Focusing on Co-sMVCCA as shown in Fig. 3 (a), the misclassification ratio is lower than that of the other methods. It is considered that the decrease of the misclassification ratio of images that belong to class C predicted as class B should be focused on. Thus, we focused on examples of distress images belonging to class C that were correctly classified by our method but were misclassified as class B by OsMVCCA. The number of those images is 22, but those images do not necessarily represent distresses that occurred in different transmission towers; that is, there are some distresses that occurred in the same transmission tower. Examples of those images are shown in Fig. 4. In addition, as shown in Table 5, we listed the methods that misclassified images in Fig. 4. Note that we adopted images that were misclassified by OsMVCCA since those images are used for validation of the differences between Co-sMVCCA and OsMVCCA. As shown in Table 5, Co-sMVCCA was the only method that correctly classified all images in Fig. 4. Since distress images of the same transmission tower generally have the same text data, it is considered that there is a correlation between those images and text data of the same tower. Thus, when the images are of the same towers and have the same levels, it is expected that calculation of the covariance matrix between heterogeneous features of different samples is effective for deterioration level estimation.

In order to discuss the representation ability of the transformed features, visualization results of the features based on tSNE are shown in Fig. 5. Figure 5 shows the visualization results of “only visual features”, CCA, sMVCCA, OsMVCCA and Co-sMVCCA when using InceptionResNet-v2 as the CNN model. The red, green and blue points in this figure represent features belonging to classes “A”, “B” and “C”, respectively. As shown in Fig. 5 (a), it was revealed that “only visual features” obtained from the pre-trained model do not represent the deterioration levels. Second, although CCA-based transformed features belonging to the same class are arranged nearby, they are scattered in the space, and they do not have high discriminative ability. On the other hand, sMVCCA shows that the transformed features belonging to the same class are grouped into a few clusters, and we confirm that it is more expressive than “only visual features” and basic CCA. Compared to sMVCCA, OsMVCCA-based features are efficiently grouped together. However, as shown in the resutls for OsMVCCA, red, blue and green features become admixed. That is, it is difficult to reflect the class information by both the class information as one view and the ordinal scale. Then, in Co-sMVCCA, by introducing class divergence into the computation process of the

covariance matrix, we succeeded in reducing the mixing of three classes as shown in Fig. 5 (e). In addition, in Fig. 5 (e), “classes A and B” and “classes B and C” are arranged nearby, but, interestingly, features of “classes A and C” are separated from each other. It is thought that this results from consideration of the category-based ordinal approach. From the above, Co-sMVCCA can strongly reflect class information and the ordinal scale, and it achieves better feature transformation than that achieved by various CCA-based methods.

Although previous CCA-based methods including OsMVCCA cannot calculate the covariance matrix between heterogeneous features of different samples, Co-sMVCCA can consider within-class divergence by dealing with the covariance matrix as shown in Eqs. (9) and (11), and this is the contribution of this paper. Therefore, the effectiveness of Co-sMVCCA including the category-based approach is verified.

Table 2 Recall, Precision and F-measure values of the proposed method and comparative methods using Inception-ResNetV2 [35].

Method	Class A			Class B			Class C			Average		
	R	P	F	R	P	F	R	P	F	R	P	F
Only text features	0.614	0.651	0.629	0.728	0.733	0.730	0.693	0.663	0.676	0.678	0.682	0.679
Only visual features [35]	0.626	0.605	0.612	0.567	0.596	0.580	0.644	0.636	0.636	0.612	0.612	0.610
Fine-tuning [35]	**0.845**	0.667	0.742	0.686	0.638	0.657	0.557	0.850	0.665	0.696	0.718	0.688
CCA [13]	0.638	0.670	0.653	0.696	0.678	0.686	0.692	0.690	0.690	0.675	0.679	0.676
GrMCCs [44]	0.638	0.690	0.661	0.711	0.702	0.705	0.771	0.740	0.754	0.706	0.711	0.707
LapMCCs [43]	0.706	0.738	0.720	0.670	0.683	0.676	0.748	0.711	0.728	0.708	0.711	0.708
LDMCCA [30]	0.685	0.725	0.704	0.735	0.704	0.719	0.735	0.739	0.736	0.718	0.723	0.719
Deep CCA [3]	0.735	0.753	0.743	0.696	0.706	0.700	0.744	0.721	0.732	0.725	0.727	0.725
sMVCCA [22]	0.811	0.846	0.826	0.811	0.810	0.809	0.866	0.844	0.854	0.829	0.833	0.830
OsMVCCA [28]	0.811	0.841	0.824	0.818	0.817	0.816	0.876	0.856	0.865	0.835	0.838	0.835
Co-sMVCCA (ours)	0.818	**0.857**	**0.836**	**0.827**	**0.832**	**0.829**	**0.893**	**0.858**	**0.874**	**0.846**	**0.849**	**0.846**

Table 3 Recall, Precision and F-measure values of the proposed method and comparative methods using DenseNet-201 [15].

Method	Class A			Class B			Class C			Average		
	R	P	F	R	P	F	R	P	F	R	P	F
Only text features	0.556	0.617	0.580	0.672	0.704	0.685	0.673	0.607	0.633	0.634	0.643	0.633
Only visual features [15]	0.646	0.647	0.646	0.677	0.681	0.678	0.725	0.724	0.724	0.683	0.684	0.683
Fine-tuning [15]	0.821	0.721	0.766	0.804	0.653	0.719	0.565	**0.924**	0.697	0.730	0.766	0.727
CCA [13]	0.709	0.715	0.712	0.716	0.753	0.732	0.791	0.750	0.768	0.739	0.739	0.737
GrMCCs [44]	0.643	0.696	0.666	0.671	0.683	0.675	0.767	0.712	0.738	0.694	0.697	0.693
LapMCCs [43]	0.674	0.675	0.673	0.673	0.698	0.684	0.758	0.729	0.743	0.702	0.701	0.700
LDMCCA [30]	0.674	0.691	0.681	0.722	0.722	0.722	0.755	0.743	0.748	0.717	0.719	0.71766
Deep CCA [3]	0.782	0.753	0.767	0.723	0.743	0.732	0.802	0.809	0.804	0.769	0.768	0.768
sMVCCA [22]	0.835	0.836	0.835	0.827	0.827	0.826	0.880	0.884	0.881	0.847	0.849	0.847
OsMVCCA [28]	**0.859**	0.850	0.853	0.854	0.854	0.853	0.887	0.900	0.893	0.866	0.868	0.866
Co-sMVCCA (ours)	0.852	**0.868**	**0.859**	**0.867**	**0.862**	**0.864**	**0.909**	0.905	**0.906**	**0.876**	**0.878**	**0.876**

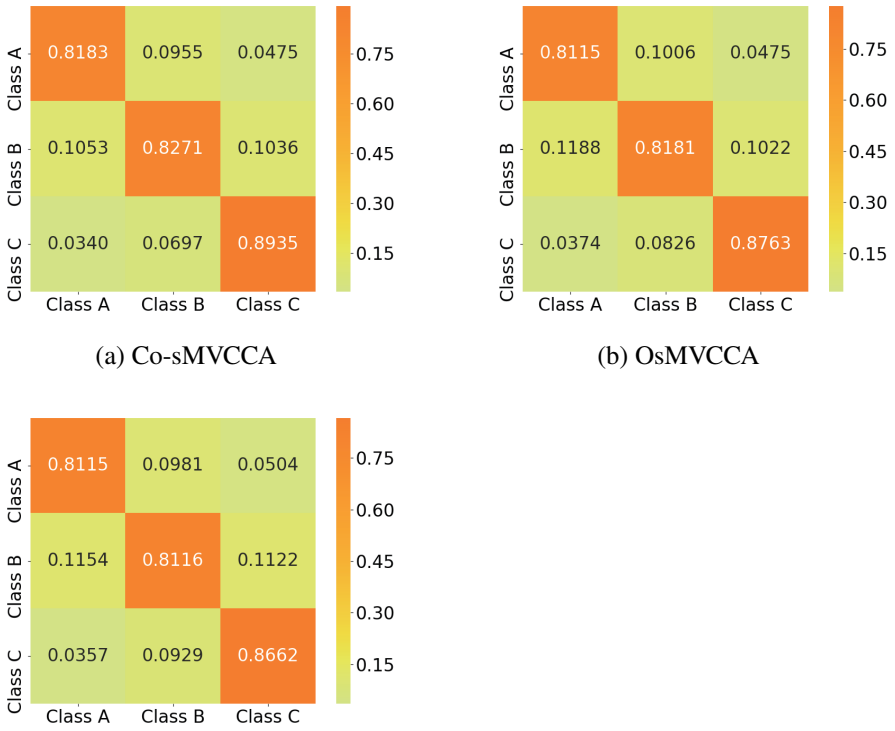
Table 4 Recall, Precision and F-measure values of the proposed method and comparative methods using Xception [5].

Method	Class A			Class B			Class C			Average		
	R	P	F	R	P	F	R	P	F	R	P	F
Only text features	0.572	0.625	0.587	0.720	0.714	0.716	0.670	0.646	0.654	0.654	0.662	0.652
Only visual features [5]	0.619	0.598	0.608	0.594	0.621	0.607	0.664	0.656	0.660	0.626	0.625	0.625
Fine-tuning [5]	**0.911**	0.637	0.746	0.640	0.689	0.658	0.589	0.856	0.695	0.713	0.727	0.700
CCA [13]	0.648	0.696	0.670	0.695	0.682	0.687	0.714	0.693	0.702	0.686	0.690	0.686
GrMCCs [44]	0.650	0.661	0.655	0.677	0.682	0.678	0.703	0.691	0.696	0.677	0.678	0.676
LapMCCs [43]	0.714	0.721	0.717	0.689	0.708	0.697	0.779	0.754	0.765	0.728	0.728	0.726
LDMCCA [30]	0.689	0.680	0.682	0.698	0.699	0.697	0.667	0.678	0.671	0.684	0.685	0.683
Deep CCA [3]	0.757	0.716	0.735	0.710	0.732	0.720	0.762	0.777	0.768	0.743	0.742	0.741
sMVCCA [22]	0.843	0.822	0.832	0.801	0.823	0.811	0.845	0.843	0.843	0.830	0.829	0.829
OsMVCCA [28]	0.835	0.846	0.839	0.824	0.831	0.826	0.863	0.850	0.856	0.841	0.843	0.840
Co-sMVCCA (ours)	0.865	**0.849**	**0.856**	**0.828**	**0.845**	**0.835**	**0.866**	**0.866**	**0.865**	**0.853**	**0.853**	**0.852**

4 Conclusion

We have proposed a deterioration level estimation method via a novel neural network maximizing category-based ordinally supervised multi-view canonical correlation. The proposed method has two contributions. One is construction of a neural network-based classifier that can be trained from a small number of training images by introducing a canonical correlation maximization approach into a middle layer of the network. The other contribution is derivation of a novel CCA-based method, Co-sMVCCA, that can consider both the ordinal scale between different classes and the within-class divergence. Consequently, the proposed neural network including Co-sMVCCA realizes accurate deterioration level estimation.

When the number of images is small, it is difficult to train CNNs, and other approaches are needed for calculating visual features with high representation ability. Thus, the calculation of visual features is limited to the use of pre-trained models or fine-tuning. That is, it is not possible to directly calculate the features that are effective for class discrimination from the images. Therefore, the strong point of the proposed method is that it enables learning from a small amount of training data by considering the correlation between multi-modal features even if the representation ability of the calculated features is low. Although this approach has been applied to the image data of transmission towers, this approach can solve the problem of a small amount of data in the case of real data, and it is therefore considered that it has high applicability to various fields. Furthermore, when focusing on multivariate analysis, our proposed Co-sMVCCA is the first CCA-based multi-modal approach that can consider label features, the ordinal scale, and class divergence, and we believe that it will make a significant contribution to the field of multivariate analysis.



(a) Co-sMVCCA

(b) OsMVCCA

(c) sMVCCA

Fig. 3 Confusion matrices of Co-sMVCCA, OsMVCCA and sMVCCA by using Inception-ResNet-V2. The value in the each cell represents the ratio of classification results. The vertical axis is the truth label and the horizontal axis is the predicted label.

5 Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Numbers JP20K19856 and JP17H01744. In this research, we utilized the data that were provided by Tokyo Electric Power Company Research Institute.

References

1. Alghamdi, A., Hammad, M., Ugail, H., Abdel-Raheem, A., Muhammad, K., Khalifa, H.S., Abd El-Latif, A.A.: Detection of myocardial infarction based on novel deep transfer learning methods for urban healthcare in smart cities. *Multimedia Tools and Applications* pp. 1–22 (2020)
2. Allouch, A., Koubâa, A., Abbes, T., Ammar, A.: Roadsense: Smartphone application to estimate road conditions using accelerometer and gyroscope. *IEEE Sensors Journal* **17**(13), 4231–4238 (2017)
3. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: Proc. International Conference on Machine Learning, pp. 1247–1255 (2013)

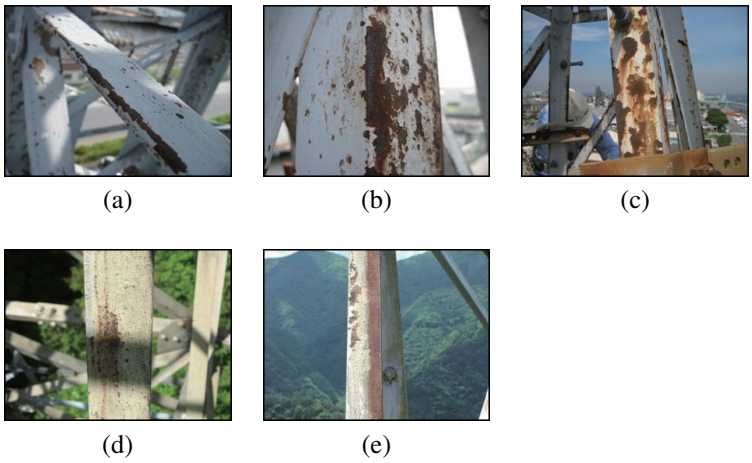
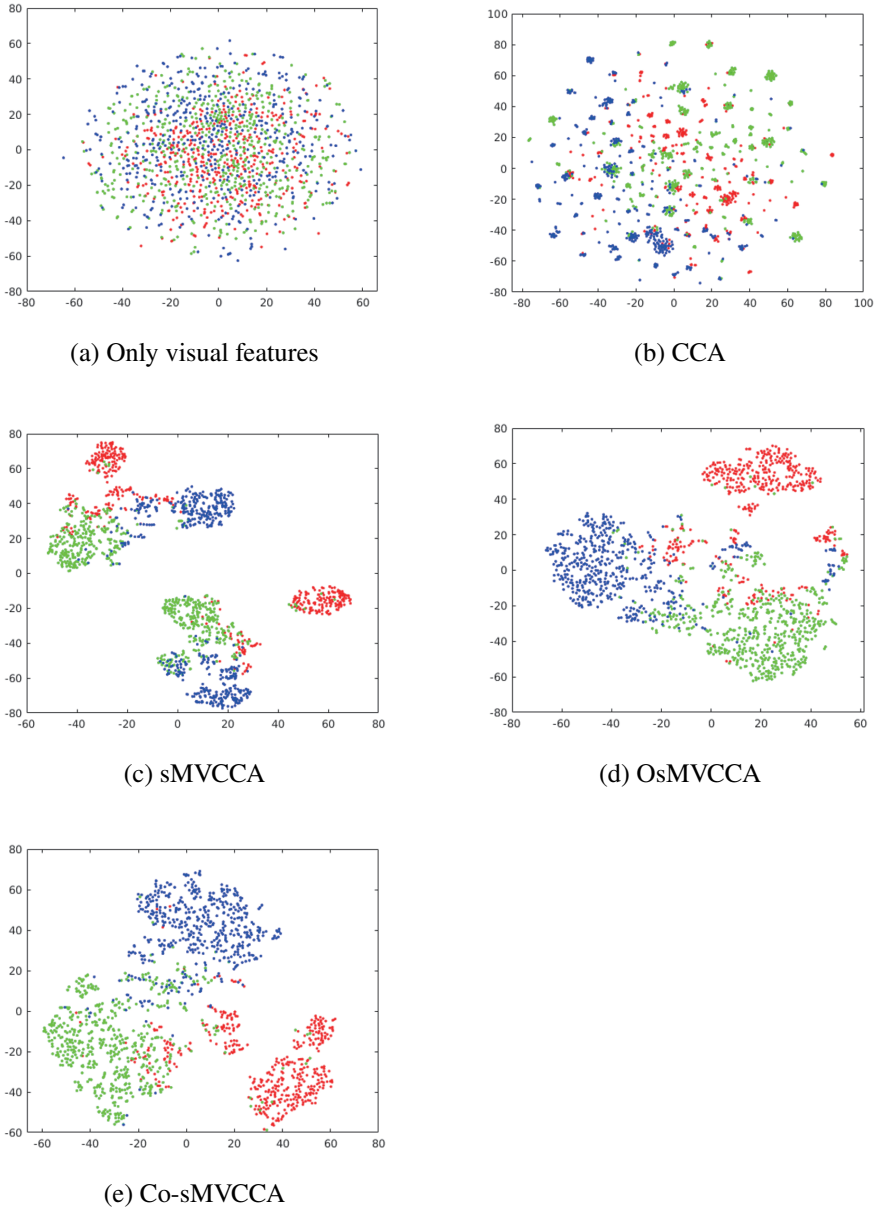


Fig. 4 Examples of distress images that were correctly classified by our method but were misclassified by the other methods. Note that (a), (b) and (c) are images of the same transmission tower and (d) and (e) are also images of the same transmission tower.

Table 5 Methods that misclassified images in Fig. 4 are marked with ✓.

Method	(a)	(b)	(c)	(d)	(e)
Only text features	✓	✓	✓	-	-
Only visual features [5]	✓	✓	-	-	✓
Fine-tuning [5]	-	✓	✓	-	-
CCA [13]	✓	-	-	✓	-
GrMCCs [44]	-	-	-	-	✓
LapMCCs [43]	✓	✓	✓	-	-
LDMCCA [30]	✓	-	-	-	-
Deep CCA [3]	-	-	✓	✓	-
sMVCCA [22]	✓	-	✓	✓	✓
OsMVCCA [28]	✓	✓	✓	✓	✓
Co-sMVCCA	-	-	-	-	-

4. Cha, Y.J., Choi, W., Büyüköztürk, O.: Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering* **32**(5), 361–378 (2017)
5. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. arXiv preprint pp. 1610–02357 (2017)
6. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: Proc. IEEE International Conference on Machine Learning, vol. 32, pp. 647–655 (2014)
7. Fan, D.P., Lin, Z., Zhao, J.X., Liu, Y., Zhang, Z., Hou, Q., Zhu, M., Cheng, M.M.: Rethinking rgb-d salient object detection: Models, datasets, and large-scale benchmarks. *arXiv preprint arXiv:1907.06781* (2019)
8. Gao, Y., Mosalam, K.M.: Deep transfer learning for image-based structural damage recognition. Computer-Aided Civil and Infrastructure Engineering **33**(9), 748–768 (2018)
9. Gopalakrishnan, K., Gholami, H., Vidyadharan, A., Choudhary, A., Agrawal, A.: Crack damage detection in unmanned aerial vehicle images of civil infrastructure using pre-trained deep learning



(a) Only visual features

(b) CCA

(c) sMVCCA

(d) OsMVCCA

(e) Co-sMVCCA

Fig. 5 Visualization results of transformed features when using Inception-ResNet-v2 as the CNN model. This visualization is performed via tSNE. (a), (b), (c), (d) and (e) show the results for "only visual features", CCA, sMVCCA, OsMVCCA and Co-sMVCCA, respectively. The red, green and blue points in this figure represent features belonging to classes "A", "B" and "C", respectively.

10. Gupta, S., Agrawal, A., Gopalakrishnan, K., Narayanan, P.: Deep learning with limited numerical precision. In: Proc. International Conference on Machine Learning, pp. 1737–1746 (2015)
11. H. Hamada *et al.*: Tokyo Electronic Power Company Holdings, Inc. (*private communication*)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3/4), 321–377 (1936)
14. Hsu, T.M.H., Weng, W.H., Boag, W., McDermott, M., Szolovits, P.: Unsupervised multimodal representation learning across medical images and reports. arXiv preprint arXiv:1811.08615 (2018)
15. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
16. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: A new learning scheme of feedforward neural networks. In: Proc. IEEE International Joint Conference on Neural Networks, vol. 2, pp. 985–990 (2004)
17. Im, J., Fujii, H., Yamashita, A., Asama, H.: Multi-modal diagnostic method for detection of concrete crack direction using light-section method and hammering test. In: *Proc. International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pp. 922–927 (2017)
18. Ji, H.K., Sun, Q.S., Yuan, Y.H., Ji, Z.X.: C2dmcp: View-consistent collaborative discriminative multiset correlation projection for data representation. Journal of Visual Communication and Image Representation **40**, 393–405 (2016)
19. Jonsson, P., Casselgren, J., Thörnberg, B.: Road surface status classification using spectral analysis of nir camera images. *IEEE Sensors Journal* **15**(3), 1641–1656 (2014)
20. Kamilaris, A., Prenafeta-Boldú, F.X.: Deep learning in agriculture: A survey. Computers and Electronics in Agriculture **147**, 70–90 (2018)
21. Kasahara, J.Y.L., Fujii, H., Yamashita, A., Asama, H.: Fuzzy clustering of spatially relevant acoustic data for defect detection. *IEEE Robotics and Automation Letters* **3**(3), 2616–2623 (2018)
22. Lee, G., Singanamalli, A., Wang, H., Feldman, M.D., Master, S.R., Shih, N.N., Spangler, E., Rebbeck, T., Tomaszewski, J.E., Madabhushi, A.: Supervised multi-view canonical correlation analysis (smvcca): integrating histologic and proteomic features for predicting recurrent prostate cancer. IEEE Trans. Medical Imaging **34**(1), 284–297 (2015)
23. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
24. Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., Omata, H.: Road damage detection and classification using deep neural networks with smartphone images. Computer-Aided Civil and Infrastructure Engineering **33**(12), 1127–1141 (2018)
25. Maeda, K., Takahashi, S., Ogawa, T., Haseyama, M.: Distress classification of road structures via adaptive Bayesian network model selection. Journal of Computing in Civil Engineering **31**(5), 04017044 (2017)
26. Maeda, K., Takahashi, S., Ogawa, T., Haseyama, M.: Estimation of deterioration levels of transmission towers via deep learning maximizing canonical correlation between heterogeneous features. IEEE Journal of Selected Topics in Signal Processing **12**(4), 633–644 (2018)
27. Maeda, K., Takahashi, S., Ogawa, T., Haseyama, M.: Convolutional sparse coding-based deep random vector functional link network for distress classification of road structures. Computer-Aided Civil and Infrastructure Engineering (2019)
28. Maeda, K., Takahashi, S., Ogawa, T., Haseyama, M.: Neural network maximizing ordinally supervised multi-view canonical correlation for deterioration level estimation. In: Proc. IEEE International Conference on Image Processing, pp. 919–923 (2019)
29. Pao, Y.H., Park, G.H., Sobajic, D.J.: Learning and generalization characteristics of the random vector functional-link net. Neurocomputing **6**(2), 163–180 (1994)
30. Peng, J., Li, Q., El-Latif, A.A.A., Niu, X.: Linear discriminant multi-set canonical correlations analysis (ldmcca): an efficient approach for feature fusion of finger biometrics. *Multimedia Tools and Applications* **74**(13), 4469–4486 (2015)
31. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: *Proc. Advances in neural information processing systems*, pp. 91–99 (2015)
33. Schmidt, W.F., Kraaijveld, M.A., Duin, R.P., et al.: Feed forward neural networks with random weights. In: International Conference on Pattern Recognition, pp. 1–1 (1992)

34. Sun, T.K., Chen, S.C., Jin, Z., Yang, J.Y.: Kernelized discriminative canonical correlation analysis. In: Proc. International Conference on Wavelet Analysis and Pattern Recognition, vol. 3, pp. 1283–1287 (2007)
35. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proc. AAAI Conference on Artificial Intelligence, pp. 4278–4284 (2017)
36. Tsutsumi, F., Murata, H., Onoda, T., Oguri, O., Tanaka, H.: Automatic corrosion estimation using galvanized steel images on power transmission towers. In: *Proc. Transmission & Distribution Conference & Exposition: Asia and Pacific*, pp. 1–4 (2009)
37. Wang, H., Yang, G., Li, E., Tian, Y., Zhao, M., Liang, Z.: High-voltage power transmission tower detection based on faster r-cnn and yolo-v3. In: *Proc. Chinese Control Conference (CCC)*, pp. 8750–8755 (2019)
38. Woo, S., Chu, I., Youn, B., Kim, K.: Development of the corrosion deterioration inspection tool for transmission tower members. KEPCO Journal on Electric Power and Energy **2**(2), 293–298 (2016)
39. Yan, B., Goto, S., Miyamoto, A., Zhao, H.: Imaging-based rating for corrosion states of weathering steel using wavelet transform and pso-svm techniques. Journal of Computing in Civil Engineering **28**(3), 04014008 (2013)
40. Yeh, Y.R., Huang, C.H., Wang, Y.C.F.: Heterogeneous domain adaptation and classification by exploiting the correlation subspace. IEEE Trans. Image Processing **23**(5), 2009–2018 (2014)
41. Yu, Y., Beuret, S., Zeng, D., Oyama, K.: Deep learning of human perception in audio event classification. In: Proc. IEEE International Symposium on Multimedia, pp. 188–189 (2018)
42. Yu, Y., Tang, S., Aizawa, K., Aizawa, A.: Category-based deep cca for fine-grained venue discovery from multimodal data. Proc. IEEE transactions on neural networks and learning systems **30**(4), 1250–1258 (2018)
43. Yuan, Y.H., Li, Y., Shen, X.B., Sun, Q.S., Yang, J.L.: Laplacian multiset canonical correlations for multiview feature extraction and image recognition. Multimedia Tools and Applications **76**(1), 731–755 (2017)
44. Yuan, Y.H., Sun, Q.S.: Graph regularized multiset canonical correlations with applications to joint feature extraction. Pattern Recognition **47**(12), 3907–3919 (2014)
45. Zhang, Q., Yang, L.T., Chen, Z., Li, P.: A survey on deep learning for big data. Information Fusion **42**, 146–157 (2018)
46. Zhao, J.X., Cao, Y., Fan, D.P., Cheng, M.M., Li, X.Y., Zhang, L.: Contrast prior and fluid pyramid integration for rgbd salient object detection. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3927–3936 (2019)