PAPER Special Section on Picture Coding and Image Media Processing

Visualizing Web Images Using Fisher Discriminant Locality Preserving Canonical Correlation Analysis

Kohei TATENO[†a)], ***Nonmember***, **Takahiro OGAWA**[†b)], ***and*** **Miki HASEYAMA**[†c)], ***Members***

SUMMARY A novel dimensionality reduction method, Fisher Discriminant Locality Preserving Canonical Correlation Analysis (FDLP-CCA), for visualizing Web images is presented in this paper. FDLP-CCA can integrate two modalities and discriminate target items in terms of their semantics by considering unique characteristics of the two modalities. In this paper, we focus on Web images with text uploaded on Social Networking Services for these two modalities. Specifically, text features have high discriminate power in terms of semantics. On the other hand, visual features of images give their perceptual relationships. In order to consider both of the above unique characteristics of these two modalities, FDLP-CCA estimates the correlation between the text and visual features with consideration of the cluster structure based on the text features and the local structures based on the visual features. Thus, FDLP-CCA can integrate the different modalities and provide separated manifolds to organize enhanced compactness within each natural cluster.

key words: *dimensionality reduction, visualization, Fisher discriminant analysis, canonical correlation analysis, locality preserving approach*

1. Introduction

With the explosive growth of social media, a massive volume of multimedia data such as images and videos are created and shared online everyday. A representative example is Flickr*, which hosted over 10 billion images in 2015 [1]. Since it is difficult for users to search for desired data from the enormous volume of data, techniques that enable efficient exploration of multimedia data are needed [2].

Browsing and exploring data require methods for visualizing items to make clear both the content of individual items and any relationships between these items. One approach is to map the items into a low-dimensional (2-D or 3-D) space based on data similarities [3]. In this paper, "items" denote original data and "data points" denote the results of their mapping into the low-dimensional space. Users can perceive a set of items that has the same semantic relationships if their data points are presented close to each other in the visualization results.

In most existing visualization methods, only one type of multimedia data, *e.g.*, images or videos, is considered [4], [5]. However, target multimedia data often contain multiple data descriptions (so-called modalities). For example, in Social Networking Services (SNSs) such as Flickr and Facebook**, uploaded multimedia data often contain different modalities, *e.g.*, images, sounds, tags, comments and geo-information. Thus, visualization methods can utilize several modalities to provide more semantic relationships of items.

Dimensionality reduction is widely used for data visualization [6], [7]. Data visualizations lay out items so similar items appear close to one another while very different items will be further apart. These differ in how they perform dimensionality reduction to map the distribution of items from the high-dimensional space to a low-dimensional space. In recent years, many dimensionality reduction methods have been proposed [8]–[10]. They are very popular due to their relative simplicity and effectiveness. However, when these methods are applied to multimedia data that contain several modalities, most of them use a high-dimensional feature vector obtained by concatenating multiple features extracted from these modalities and cannot effectively consider the characteristics of each modality. It has been reported in [11], [12] that a scheme that can consider the characteristics of each modality provides better performance for multimedia analysis than does a scheme utilizing only one modality. In order to provide better visualization for multimedia data, dimensionality reduction methods need to integrate several kinds of features.

The challenge in multimodal dimensionality reduction is to obtain a more informative projection by considering the complementarities and redundancies of all available modalities. Although some dimensionality reduction methods that can integrate different types of features have been proposed [13]–[15], they cannot consider unique characteristics of each modality.

A new dimensionality reduction method that can consider the unique characteristics of each modality is presented in this paper. In this paper, we focus on multimedia data represented by two typical modalities, text and image, that are commonly used in almost all SNSs. We regard the multimedia data as items and project them into a low-dimensional space by using the dimensionality reduction method. We can use text features to group images into semantic clusters since the text features have high discriminative power in terms of semantics of images [16]. Therefore, we use the

Manuscript received December 15, 2016.
Manuscript revised March 23, 2017.
Manuscript publicized June 14, 2017.

[†]The authors are with Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060–0814 Japan.
a) E-mail: tateno@lmd.ist.hokudai.ac.jp
b) E-mail: ogawa@lmd.ist.hokudai.ac.jp
c) E-mail: miki@ist.hokudai.ac.jp

DOI: 10.1587/transinf.2016PCP0005

*https://www.flickr.com/
**https://www.facebook.com/

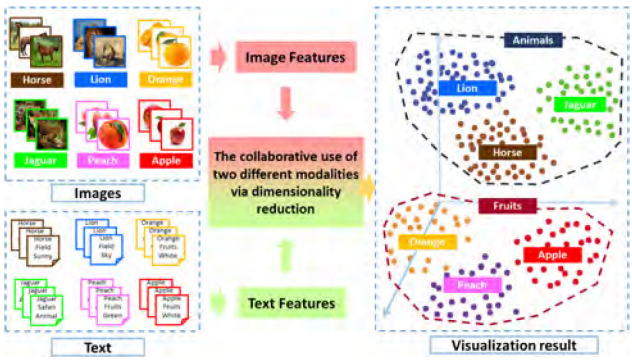


Fig. 1 Unique characteristics of visualization realized by the proposed dimensionality reduction method.

text features to group images into semantic clusters. On the other hand, in order to perform visualization that is suitable for visual perception of humans, visual information of images is also taken into consideration since the visual information gives perceptual relationships between items for humans [17]. Therefore, we can additionally use the visual features of images to represent the visual relationships among items in a low-dimensional space.

Our dimensionality reduction method is modeled for the goal of grouping items containing the same semantics in the low-dimensional space. Specifically, our method is a hybrid version of two dimensionality reduction methods, Locality Preserving Canonical Correlation Analysis (LPCCA) [18] and Fisher Discriminant Analysis (FDA) [19], and we thus call our hybrid method "Fisher Discriminant Locality Preserving Canonical Correlation Analysis" (FDLP-CCA). We formulate the optimization problem of FDLP-CCA by combining those of LPCCA and FDA. This procedure allows us to maintain the computational efficiency and reliability of LPCCA and FDA. LPCCA is a multivariate analysis method that extracts latent features based on the correlation between different features with consideration of the local structures of data. Thus, our method adopts LPCCA, which focuses on local structures based on visual features, and this contributes to the representation of visual relationships and the integration of the two modalities. Since FDA is a supervised dimensionality reduction method for discriminating different classes, our method can preserve the cluster structure based on the text features. In this way, FDLP-CCA can integrate multiple features and discriminate items in terms of semantics. Therefore, visualization via FDLP-CCA can group items containing the same semantics in the low-dimensional space.

Figure 1 shows the goal of the visualization via FDLP-CCA. Specifically, successful visualization results need to represent the semantics in the datasets. However, one single image contains various semantics at multiple semantic levels [20]. We therefore evaluate the visualization in terms of hierarchical image semantics with the same way in [21]. As shown in examples in Fig. 1, visualization representing each of the semantics (*horse, jaguar, lion, apple, orange, peach*) and similar semantics categories (*animals or fruits*), is realized, *i.e.*, the visualization results reflect the semantic hierarchy.

2. Related Work

In this section, we provide a brief review of dimensionality reduction methods. Dimensionality reduction generally converts high-dimensional data into low-dimensional data with preservation of their intrinsic structures. Traditional methods such as Principle Components Analysis (PCA) [22] and Multidimensional Scaling (MDS) [23] are linear techniques. For high-dimensional data that lies on a non-linear manifold, it is usually more important to keep similar data points close together, which is usually not possible with linear mapping. Modern dimensionality reduction methods use non-linear projections to project the data into a low-dimensional space. Some methods, such as Stochastic Neighbor Embedding (SNE) [24], t-distributed Stochastic Neighbor Embedding (t-SNE) [10] and Barnes-Hut SNE (BH-SNE) [25], attempt to match probability distributions induced by pairwise data dissimilarities in the high-dimensional space. Other methods use local linear relationships to measure the local structure, as in Local Linear Embedding (LLE) [26] and Locality Preserving Projection (LPP) [9]. The above dimensionality reduction methods are often used in data visualization [4]. However, it has been reported in [10] that despite the strong performance of these methods for artificial datasets, they are often not successful for visualization of multimedia datasets.

As mentioned in the previous section, recent multimedia data contain several modalities. It has been reported in [27] and [28] that a scheme that utilizes several kinds of features provides better performance for multimedia analysis than does a scheme that utilizes only one feature. However, since the above dimensionality reduction methods can utilize only one kind of feature, it is difficult for these methods to use several kinds of modalities effectively when they are applied to multimedia data. In [13] and [14], dimensionality reduction methods that can integrate multiple features have been proposed. Lin et al. proposed Multiple Kernel Learning for Dimensionality Reduction (MKL-DR) [13]. MKL-DR introduces multiple kernel learning into the training process of dimensionality reduction methods. It works with multiple base kernels and fuses the descriptors in the domain of kernel matrices. In [14], Yun et al. proposed Multi-set Canonical Correlations using Globality-preserving Projections (MCC-GP), which can perform joint dimensionality reduction for high-dimensional data. MCC-GP represents the correlations of any pair of feature sets in the low-dimensional space. Although these methods integrate different kinds of features for realizing dimensionality reduction, they cannot consider the unique characteristics of each modality. In [27], it has been reported that considering the characteristics of target modalities is necessary for successful multimodal data analysis.

3. Dimensionality Reduction via Fisher Discriminant Locality Preserving Canonical Correlation Analysis

In this section, we present a novel dimensionality reduction method based on FDLP-CCA. By considering the unique characteristics of the text and visual features, FDLP-CCA can not only integrate multiple features maximizing their correlation but also discriminate items as semantics. Specifically, this dimensionality reduction is based on a hybrid version of two dimensionality reduction methods: LPCCA [18] and FDA [19].

This section is organized as follows. We first explain the unique characteristics of each modality that are useful for dimensionality reduction and show the goal of FDLP-CCA in 3.1. Next, we show the model formulation of FDLP-CCA in 3.2. Finally, detailed explanations of implementations of FDLP-CCA are shown in 3.3.

3.1 Unique Characteristics of Each Modality and Goal of FDLP-CCA

In SNSs, when uploading images, users often attach text to the images. In this paper, since we use Web images with text, they have two modalities (*i.e.,* text and image). As described in the previous section, dimensionality reduction should be performed with consideration of the characteristics of their modalities.

Since text that is manually assigned represents the semantics of images, text features have high discriminative power in terms of semantics [16]. Therefore, we use the text features to group images into semantic clusters. On the other hand, visual information of images should also be taken into consideration to improve the projection since the visual information gives perceptual relationships between images for humans. Specifically, it has been reported in [17] that semantic is the knowledge of human perception and that 80% of human cognition comes from visual information. It is reasonable that visual information generate the knowledge about semantic relationships. Therefore, simple use of text is not a reliable and reasonable solution, and visual information of images should also be taken into consideration to improve the representation of semantic relationships.

Note that the relationships among visual features are not generally linear. Therefore, dimensionality reduction should preserve the locality of visual features for representing such a non-linearity. In addition, when using the two modalities, we need to project items from the original two feature spaces to a lower-dimensional space. In order to combine text and visual features, our dimensionality reduction has to estimate the correlation between the text and visual features. By considering the above characteristics, FDLP-CCA can integrate multiple features and discriminate items in terms of semantics.

3.2 Model Formulation of FDLP-CCA

We explain the model formulation of FDLP-CCA in this subsection. Given items I_n ($n = 1, 2, \ldots, N$; N being the total number of items), two kinds of feature vectors $\boldsymbol{x}_{\mathrm{m}_1,n} \in \mathbb{R}^{d_{\mathrm{m}_1}}$ and $\boldsymbol{x}_{\mathrm{m}_2,n} \in \mathbb{R}^{d_{\mathrm{m}_2}}$ are extracted from a pair of the two modalities $(\mathrm{m}_1, \mathrm{m}_2)$ of item I_n, respectively, where d_{m_l} and d_{m_2} represent the dimensions of $\boldsymbol{x}_{\mathrm{m}_1,n}$ and $\boldsymbol{x}_{\mathrm{m}_2,n}$, respectively. In addition, each item belongs to a class, and the class to which it belongs is denoted as $l_n (\in \{1, 2, \ldots, K\}$; K being the total number of classes). Furthermore, we define $\boldsymbol{X}_{\mathrm{m}_1} = [\boldsymbol{x}_{\mathrm{m}_1,1}, \ldots, \boldsymbol{x}_{\mathrm{m}_1,N}]$, $\boldsymbol{X}_{\mathrm{m}_2} = [\boldsymbol{x}_{\mathrm{m}_2,1}, \ldots, \boldsymbol{x}_{\mathrm{m}_2,N}]$ and assume that these matrices are centered for convenience. FDLP-CCA seeks a set of projections $\boldsymbol{w}_{\mathrm{m}_1} \in \mathbb{R}^{d_{\mathrm{m}_1}}$ and $\boldsymbol{w}_{\mathrm{m}_2} \in \mathbb{R}^{d_{\mathrm{m}_2}}$, and its details are shown below.

The basic idea of FDLP-CCA is to integrate the two features and discriminate different semantics in a low-dimensional space. Thus, we try to obtain latent features by considering the correlation between the two kinds of features and discriminating items that have different semantics with consideration of the non-linear structures. Then we focus on LPCCA and FDA to formulate the optimization problem of FDLP-CCA. LPCCA is a locally linear multivariate analysis method and has the effect of globally non-linear dimensionality reduction. By this method, local structure information is preserved, and the correlation between the two features is also obtained. On the other hand, FDA is a supervised dimensionality reduction method. FDA can project the original high-dimensional data onto the low-dimensional space, where all classes are separated well by maximizing the ratio of between-class scatter matrix to within-class scatter matrix. We formulate the objective function of FDLP-CCA by combining the objective functions of FDA and LPCCA. This allows us to maintain the computational efficiency and reliability of FDA and LPCCA. Therefore, FDLP-CCA has the advantages of the above two methods and can be regarded as a hybrid version of LPCCA and FDA.

When considering LPCCA, the following optimization problem is provided:

$$\begin{aligned} \arg\max_{\boldsymbol{w}_{\mathrm{m}_1}\boldsymbol{w}_{\mathrm{m}_2}} \quad & \boldsymbol{w}_{\mathrm{m}_1}^{\top}\boldsymbol{X}_{\mathrm{m}_1}\boldsymbol{G}_{\mathrm{m}_1,\mathrm{m}_2}\boldsymbol{X}_{\mathrm{m}_2}{}^{\top}\boldsymbol{w}_{\mathrm{m}_2}, \\ \text{subject to} \quad & \boldsymbol{w}_{\mathrm{m}_1}^{\top}\boldsymbol{X}_{\mathrm{m}_1}\boldsymbol{G}_{\mathrm{m}_1,\mathrm{m}_1}\boldsymbol{X}_{\mathrm{m}_1}{}^{\top}\boldsymbol{w}_{\mathrm{m}_1} = 1, \\ & \boldsymbol{w}_{\mathrm{m}_2}^{\top}\boldsymbol{X}_{\mathrm{m}_2}\boldsymbol{G}_{\mathrm{m}_2,\mathrm{m}_2}\boldsymbol{X}_{\mathrm{m}_2}{}^{\top}\boldsymbol{w}_{\mathrm{m}_2} = 1, \end{aligned} \tag{1}$$

where $\boldsymbol{G}_{\mathrm{m}_1,\mathrm{m}_2} = \boldsymbol{D}_{\mathrm{m}_1,\mathrm{m}_2} - \boldsymbol{S}_{\mathrm{m}_1} \circ \boldsymbol{S}_{\mathrm{m}_2}$ is called the Laplacian matrix. Furthermore, $\boldsymbol{S}_{\mathrm{m}_1} \circ \boldsymbol{S}_{\mathrm{m}_2}$ is the Hadamard product of the similarity matrices $\boldsymbol{S}_{\mathrm{m}_1}$ and $\boldsymbol{S}_{\mathrm{m}_2}$, $\boldsymbol{D}_{\mathrm{m}_1,\mathrm{m}_2}$ is a diagonal matrix of size $N \times N$, and its i th diagonal entry equals the sum of the entries in the i th row of the matrix $\boldsymbol{S}_{\mathrm{m}_1} \circ \boldsymbol{S}_{\mathrm{m}_2}$. Given the affinity $A_{\mathrm{m}}(i, j) (\mathrm{m} \in \{\mathrm{m}_1, \mathrm{m}_2\})$ between $\boldsymbol{x}_{\mathrm{m},i}$ and $\boldsymbol{x}_{\mathrm{m},j}$ as

$$A_{\mathrm{m}}(i, j) = \exp\left(-\frac{\|\boldsymbol{x}_{\mathrm{m},i} - \boldsymbol{x}_{\mathrm{m},j}\|^2}{t_{\boldsymbol{x}^{\mathrm{m}}}}\right), \tag{2}$$

the similarity matrices $\boldsymbol{S}_{\mathrm{m}}(\mathrm{m} \in \{\mathrm{m}_1, \mathrm{m}_2\})$ are obtained as

$$\boldsymbol{S}_{\mathrm{m}}(i,j) = \begin{cases} A_{\mathrm{m}}(i,j), & \text{if } I_j \in \mathrm{LN}(I_i) \text{ or } I_i \in \mathrm{LN}(I_j) \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

Let LN (I_i) be an item set that comprises the local neighbors of item I_i, and the parameter $t_{\boldsymbol{x}_{\mathrm{m}}}$ be generally taken as the mean square distance $\sum_{i=1}^{N}\sum_{j=1}^{N} 2\|\boldsymbol{x}_{\mathrm{m},i} - \boldsymbol{x}_{\mathrm{m},j}\|^2/(N(N-1))$. Note that $\boldsymbol{G}_{\mathrm{m}_1,\mathrm{m}_1}$, $\boldsymbol{G}_{\mathrm{m}_2,\mathrm{m}_2}$ in Eq. (1) and $\boldsymbol{G}_{\mathrm{m}_2,\mathrm{m}_1}$ can be computed in the same manner as $\boldsymbol{G}_{\mathrm{m}_1,\mathrm{m}_2}$. The solutions of Eq. (1) are obtained by solving the following generalized eigenvalue decomposition problem:

$$\overline{\boldsymbol{C}}^{\mathrm{LPCCA}} \begin{pmatrix} \boldsymbol{w}_{\mathrm{m}_1} \\ \boldsymbol{w}_{\mathrm{m}_2} \end{pmatrix} = \lambda \underline{\boldsymbol{C}}^{\mathrm{LPCCA}} \begin{pmatrix} \boldsymbol{w}_{\mathrm{m}_1} \\ \boldsymbol{w}_{\mathrm{m}_2} \end{pmatrix}, \tag{4}$$

where $\overline{\boldsymbol{C}}^{\mathrm{LPCCA}}$ and $\underline{\boldsymbol{C}}^{\mathrm{LPCCA}}$ are defined as

$$\overline{\boldsymbol{C}}^{\mathrm{LPCCA}} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{X}_{\mathrm{m}_1}\boldsymbol{G}_{\mathrm{m}_1,\mathrm{m}_2}\boldsymbol{X}_{\mathrm{m}_2}{}^{\top} \\ \boldsymbol{X}_{\mathrm{m}_2}\boldsymbol{G}_{\mathrm{m}_2,\mathrm{m}_1}\boldsymbol{X}_{\mathrm{m}_1}{}^{\top} & \boldsymbol{0} \end{pmatrix}, \tag{5}$$

$$\underline{\boldsymbol{C}}^{\mathrm{LPCCA}} = \begin{pmatrix} \boldsymbol{X}_{\mathrm{m}_1}\boldsymbol{G}_{\mathrm{m}_1,\mathrm{m}_1}\boldsymbol{X}_{\mathrm{m}_1}{}^{\top} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_{\mathrm{m}_2}\boldsymbol{G}_{\mathrm{m}_2,\mathrm{m}_2}\boldsymbol{X}_{\mathrm{m}_2}{}^{\top} \end{pmatrix}. \tag{6}$$

On the other hand, when considering FDA, the following optimization problem is provided:

$$\arg\max_{\boldsymbol{w}} \ \boldsymbol{w}^{\top}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{w} \ \text{ subject to } \ \boldsymbol{w}^{\top}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{w} = 1, \tag{7}$$

where $\boldsymbol{S}_{\mathrm{B}}$ and $\boldsymbol{S}_{\mathrm{W}}$ are the between-class scatter matrix and the within-class scatter matrix, respectively, and are defined as

$$\boldsymbol{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\top}, \tag{8}$$

$$\boldsymbol{S}_{\mathrm{W}} = \sum_{k=1}^{K} \sum_{n:l_n=k} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^{\top}, \tag{9}$$

where $\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n:l_n=k}\boldsymbol{x}_n$, $\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n$, $\boldsymbol{x}_n = [\boldsymbol{x}_{\mathrm{m}_1,n}{}^{\top}, \boldsymbol{x}_{\mathrm{m}_2,n}{}^{\top}]^{\top}$, $\boldsymbol{w} = [\boldsymbol{w}_{\mathrm{m}_1}^{\top}, \boldsymbol{w}_{\mathrm{m}_2}^{\top}]^{\top}$, and N_k is the total number of items belonging to class k. The solutions of Eq. (7) are obtained by solving the following generalized eigenvalue decomposition problem:

$$\boldsymbol{S}_{\mathrm{B}}\boldsymbol{w} = \lambda\boldsymbol{S}_{\mathrm{W}}\boldsymbol{w}. \tag{10}$$

As mentioned above, FDLP-CCA can be regarded as a hybrid version of LPCCA and FDA. Therefore, the combined optimization problem is defined. We derive the following Lagrange multiplier approach in FDLP-CCA:

$$\begin{aligned} L = \alpha\Big\{ & \boldsymbol{w}_{\mathrm{m}_1}^{\top}\boldsymbol{X}_{\mathrm{m}_1}\boldsymbol{G}_{\mathrm{m}_1,\mathrm{m}_2}\boldsymbol{X}_{\mathrm{m}_2}{}^{\top}\boldsymbol{w}_{\mathrm{m}_2} \\ & - \frac{\lambda}{2}(\boldsymbol{w}_{\mathrm{m}_1}^{\top}\boldsymbol{X}_{\mathrm{m}_1}\boldsymbol{G}_{\mathrm{m}_1,\mathrm{m}_1}\boldsymbol{X}_{\mathrm{m}_1}{}^{\top}\boldsymbol{w}_{\mathrm{m}_1} - 1) \\ & - \frac{\lambda}{2}(\boldsymbol{w}_{\mathrm{m}_2}^{\top}\boldsymbol{X}_{\mathrm{m}_2}\boldsymbol{G}_{\mathrm{m}_2,\mathrm{m}_2}\boldsymbol{X}_{\mathrm{m}_2}{}^{\top}\boldsymbol{w}_{\mathrm{m}_2} - 1)\Big\} \\ & + (1-\alpha)\Big\{\boldsymbol{w}^{\top}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{w} - \lambda(\boldsymbol{w}^{\top}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{w} - 1)\Big\}, \end{aligned} \tag{11}$$

where $\alpha(\in [0,1])$ is a trade-off parameter. In order to obtain the optimal vectors $\boldsymbol{w}_{\mathrm{m}_1}$ and $\boldsymbol{w}_{\mathrm{m}_2}$ from Eq. (11), we calculate $\frac{\partial L}{\partial \boldsymbol{w}_{\mathrm{m}_1}} = 0$ and $\frac{\partial L}{\partial \boldsymbol{w}_{\mathrm{m}_2}} = 0$, and then the following generalized eigenvalue problem can be derived:

$$\overline{\boldsymbol{C}}_{\mathrm{P}}\boldsymbol{w} = \lambda\underline{\boldsymbol{C}}_{\mathrm{P}}\boldsymbol{w}, \tag{12}$$

where $\overline{\boldsymbol{C}}_P$ and $\underline{\boldsymbol{C}}_P$ are defined as

$$\overline{\boldsymbol{C}}_P = \alpha\overline{\boldsymbol{C}}^{\mathrm{LPCCA}} + (1-\alpha)\boldsymbol{S}_{\mathrm{B}}, \tag{13}$$

$$\underline{\boldsymbol{C}}_P = \alpha\underline{\boldsymbol{C}}^{\mathrm{LPCCA}} + (1-\alpha)\boldsymbol{S}_{\mathrm{W}}. \tag{14}$$

By solving Eq. (12), the optimal vectors $\boldsymbol{w}_{\mathrm{m}_1}$ and $\boldsymbol{w}_{\mathrm{m}_2}$ are obtained. The solutions of this dimensionality reduction can be computed in the same way as FDA or LPCCA.

Compared to LPCCA, FDLP-CCA has the following advantage: the projection enables discrimination based on classes. On the other hand, compared to FDA, FDLP-CCA has the following advantage: the projection considers the correlation between the two kinds of feature vectors and the local structure of items. In this way, this dimensionality reduction can integrate the two kinds of features and discriminate items based on the classes. By applying FDLP-CCA to multimedia data that include image and text data, the visualization that groups items into similar semantics in the low-dimensional space becomes feasible. The details are shown in the following subsection.

3.3 Implementation of FDLP-CCA for Dimensionality Reduction

In this subsection, we explain the implementation of FDLP-CCA for dimensionality reduction. First, we explain the details of the items and modalities. In this paper, we use Web images with text that have two modalities (*i.e.*, text and image) as the items and extract text features and visual features from the text and images, respectively. We calculate the text feature vector $\boldsymbol{x}_{\mathrm{t},n}$ and the visual feature vector $\boldsymbol{x}_{\mathrm{v},n}$. Then $\boldsymbol{x}_{\mathrm{t},n}$ and $\boldsymbol{x}_{\mathrm{v},n}$ correspond to $\boldsymbol{x}_{\mathrm{m}_1,n}$ and $\boldsymbol{x}_{\mathrm{m}_2,n}$ in the previous subsection, respectively.

Next, we explain the way to obtain the class label that is used in FDLP-CCA. As described in 3.1, text often has a cluster structure that represents semantics [16]. Therefore, clustering based on the text features can divide items into clusters. In order to attach each image to class l_n, we simply apply k-means clustering to the text features. We also explain the similarity of items and define the similarity matrices of Eq. (3). As described above, detailed similarity relationships can be represented by visual similarity. Therefore, we define the item set LN (I_i) that comprises the local neighbors of item I_i in terms of visual feature vectors $\boldsymbol{x}_{\mathrm{v},i}$ based on their Euclidean distances.

Finally, we show the algorithm of FDLP-CCA for dimensionality reduction in **Algorithm 1**.

Algorithm 1 : Algorithm of Fisher Discriminant Canonical Correlation Analysis for Dimensionality Reduction.

Input: Text and visual feature vectors $\boldsymbol{x}_{\mathrm{t},n}$ and $\boldsymbol{x}_{\mathrm{v},n} (n = 1, 2, \cdots, N)$.
Output: Low-dimensional representations $\boldsymbol{Y} \in \mathbb{R}^{d \times N}$ that are coordinates in d-dimensional space.

1: Apply k-means clustering to $\boldsymbol{x}_n$ and obtain the class labels $l_n (\in \{1, 2, \ldots, K\})$.
2: Compute the between-class scatter matrix ($\boldsymbol{S}_{\mathrm{B}}$), within-class scatter matrix ($\boldsymbol{S}_{\mathrm{W}}$) and the matrices ($\overline{\boldsymbol{C}}^{\mathrm{LPCCA}}$ and $\underline{\boldsymbol{C}}^{\mathrm{LPCCA}}$).
3: Obtain the matrices $\overline{\boldsymbol{C}}_P$ and $\underline{\boldsymbol{C}}_P$ from Eqs. (13) and (14).
4: Solve the eigenvalue problem $\overline{\boldsymbol{C}}_{\mathrm{P}}\boldsymbol{w} = \lambda \underline{\boldsymbol{C}}_{\mathrm{P}}\boldsymbol{w}$ in Eq. (12).
5: Sort eigenvalues with descending order and obtain the top d largest positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.
6: Compute the eigenvectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_d$ corresponding to the top d largest positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, respectively.
7: Compute the representation from $\boldsymbol{Y} = \boldsymbol{W}^\top \boldsymbol{X}$ using $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_d]$, where $\boldsymbol{X} = [\boldsymbol{X}_{\mathrm{v}}{}^\top, \boldsymbol{X}_{\mathrm{t}}{}^\top]^\top$.
8: Return $\boldsymbol{Y}$.

4. Experimental Results

In this section, we show experimental results obtained by using real data to validate our method. The datasets contain images with text collected from Flickr. We explain experimental settings in Sect. 4.1 and then show the results of the dimensionality reduction in Sect. 4.2. Moreover, in Sect. 4.3, we also investigate the impact of different parameters in the our method.

4.1 Experimental Settings

In this subsection, we explain the goal of our experiments, the datasets, the comparison methods and the evaluation measures.

4.1.1 Goal of Our Experiments

The goal of our experiments is to show the effectiveness of visualization via the proposed dimensionality reduction method. It is a challenging problem for dimensionality reduction to realize visualization that represents semantic relationships. Since a single image contains various semantics at multiple semantic levels [29], we evaluate the results in terms of hierarchical image semantics (*i.e.*, "concept", "category" and "higher category"). Specifically, the concept corresponds to object level, and the category represents the relationship between similar concepts like hypernymy. The higher category represents the relationships between similar categories. Figure 2 shows an example of hierarchical image semantics. The concepts in the same category do not always co-occur in an image but instead are correlated compared to the concepts of other categories.

4.1.2 Datasets

Our experiments were conducted by using four datasets (Datasets 1, 2, 3 and 4) crawled on the Web and the public NUS-WIDE [30] dataset. Datasets 1, 2, 3 and 4 consist of images that were collected from Flickr. Specifically,

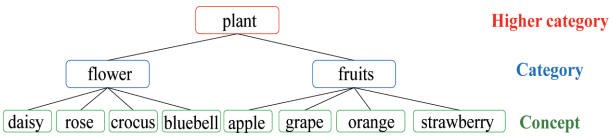


Fig. 2 Example of hierarchical image semantics.

Table 1 Query keywords used for obtaining each dataset.

Dataset	Query keywords
Dataset 1	cow, horse, jaguar, lion, panda, rabbit, wolf, apple, orange, peach, pineapple, strawberry
Dataset 2	bluebell, crocus, daisy, rose, sunflower, airplane, bicycle, bus, helicopter, motorcycle, train
Dataset 3	tables, chair, rack, shelf, sofa, dolphin, shark, starfish, jellyfish, fish, turtle
Dataset 4	cow, horse, jaguar, lion, panda, rabbit, wolf, apple, orange, peach, pineapple, strawberry, bluebell, crocus, daisy, rose, sunflower, airplane, bicycle, bus, helicopter, motorcycle, train, tables, chair, rack, shelf, sofa, dolphin, shark, starfish, jellyfish, fish, turtle

we collected images with text from the keyword search results. Note that we used tags provided by Flickr as text in this experiment. Table 1 shows the query keywords in each dataset. The number of images in datasets 1, 2 and 3 is 1000 per keyword, and that in dataset 4 is 300 per keyword. In these datasets, we extracted text features and visual features from the text and image, respectively. We calculated inherent topics in the text by using probabilistic latent semantic analysis (PLSA) [31] as text features. Next, we used hue saturation (HS) histogram, scale invariant feature transform (SIFT) [32] and histogram of oriented gradients (HOG) [33] to describe the visual features. Note that we obtained SIFT features by applying the BoF approach [34] to 128-dimensional SIFT descriptors. We used 4276-dimensional visual features (HS: 360, SIFT: 1000, HOG: 2916) and 150-dimensional text features in datasets 1-4. NUS-WIDE [30] is a popular social image benchmark dataset. The images are manually categorized into 81 classes and represented as low-level features such as SIFT, color histogram, wavelet texture and bags of text tags. In this experiment, we selected a subset of NUS-WIDE in order to reduce computational quantity. We selected 30 classes, which included more than 500 images, in alphabetical order and randomly used 500 images per class. Specifically, we used "airport, animal, beach, birds, bridge, buildings, cat, clouds, dog, fire, fish, flowers, food, garden, grass, horses, house, military, ocean, police, protest, reflection, sign, sky, snow, sports, street, sun, sunset and temple". We used 933-dimensional visual features (SIFT: 500, Color histogram: 64, Color correlogram: 144, Block-wide color moments: 225) and 954-dimensional text features in the NUS-WIDE dataset.

As mentioned above, each image contains various meanings at multiple semantic levels, and we thus attached evaluation labels based on hierarchical image semantics to each image. The hierarchical image semantics of each query keyword are shown in Table 2. Datasets 1-3 consist of images that belong to either of two categories, and dataset 4 consists of images that belong to one of three higher cat-

Table 2 Query keywords used for obtaining each dataset and the hierarchical image semantics.

Query keywords (= Concept)	Category	Higher Category
cow, horse, jaguar, lion, panda, rabbit, wolf,	land animals	animals
dolphin, shark, starfish, jellyfish, fish, turtle	sea animals	
apple, orange, peach, pineapple, strawberry	fruits	plants
bluebell, crocus, daisy, rose, sunflower,	flower	
airplane, bicycle, bus, helicopter, motorcycle, train	vehicle	artifacts
tables, chair, rack, shelf, sofa,	furniture	

egories (animals, plants and artifacts). Therefore, when we used datasets 1, 2 and 3, we evaluated the dimensionality reduction methods based on the concept and category levels. When we used dataset 4, we performed evaluation based on the concept, category and higher category levels. On the other hand, when we used NUS-WIDE, we performed evaluation based on the concept level that represents 30 classes.

4.1.3 Comparison Methods

We compared our proposed method (*i.e.*, FDLP-CCA) with seven other dimensionality reduction methods as comparison methods: LPCCA [18], FDA [19], MDS [23], Isomap [8], t-SNE [10], BH-SNE [25] and m-SNE [15]. Since FDLP-CCA is a hybrid version of LPCCA and FDA, LPCCA and FDA correspond to FDLP-CCA of $\alpha = 1$ and $\alpha = 0$, respectively. Therefore, we used LPCCA and FDA as comparative methods. MDS is a traditional linear dimensionality reduction method and is often used for visualization. Isomap and t-SNE are benchmarking non-linear dimensionality reduction methods for visualization. BH-SNE is a state-of-the-art dimensionality reduction method for visualization. We perform the visualization via the above methods by concatenating the text feature vector and the visual feature vector. On the other hand, m-SNE is a state-of-the-art multimodal dimensionality reduction method that can integrate heterogeneous features. By applying these methods and comparing their results with the results of our method, we verified the validity of our contributions.

4.1.4 Evaluation Measures

In order to compare different dimensionality reduction methods, Pseudo F [35] and k-NN classification accuracy that quantified the suitability of a particular item placement were used. The Pseudo F statistic describes the ratio of the mean sum of squares between classes to the mean sum of squares within a class. A large value of Pseudo F indicates separated classes. The k-NN classification accuracy is calculated by performing k-NN classification in the low-dimensional space. For k-NN evaluation, the results are obtained from all items and their 200 neighbors. A large value of k-NN classification accuracy indicates that the neighbors tend to have the same class.

4.2 Results of Dimensionality Reduction

We show the experimental results of visualization via dimensionality reduction. We first visualized images and subjectively evaluated the performance of each dimensionality reduction method. Then we quantitatively evaluated their performances based on the evaluation measures shown in the previous subsection. Our paper aims for the visualization, so we focus on the number of dimensions $d = 2$ and 3. In this experiment, we experimentally determined α. Specifically, the value of α was changed from 0 to 1 increasing by 0.05, and we selected the results when the k-nn classification accuracy of the concept level became the highest. In the case of $d = 2$, for each dataset, the optimal values of α are 0.90, 0.75, 0.70, 0.80 and 0.75, respectively. In the case of $d = 3$, for each dataset, the optimal values of α are 0.85, 1.0, 0.85, 0.80 and 0.80, respectively. Furthermore, we experimentally determined K of k-means. Specifically, the value of K was changed from 5 to 15 increasing by 2, and we selected the results when the k-nn classification accuracy of the concept level became the highest. In the case of $d = 2$, for each dataset, the optimal values of K are 15, 7, 11, 9 and 15, respectively. In the case of $d = 3$, for each dataset, the optimal values of K are 13, 11, 9, 9 and 11, respectively.

First, we compare the results for datasets 1 and 2. Due to the limitation of space, we only show the 2D visualization results for datasets 1 and 2 via FDLP-CCA, Isomap as a benchmarking dimensionality reduction method, BH-SNE as a state-of-the-art dimensionality reduction method in Figs. 3 and 4. In these figures, we also show results by LPCCA and FDA. In Figs. 3 and 4, each point plus a color denotes the concept or the category. From the results, we can find the following advantages of FDLP-CCA.

1. Compared to the comparison methods, FDLP-CCA can deliver more separated manifolds based on each concept.
2. Compared to the comparison methods, FDLP-CCA can organize enhanced compactness within each category.
3. Compared to LPCCA and FDA, FDLP-CCA can perform the better visualization by considering the characteristics of both LPCCA and FDA.

As shown in Figs. 3 and 4, FDLP-CCA can represent hierarchical image semantics of each dataset.

Secondly, we compare our method and the other comparison methods based on the evaluation measures using datasets 1, 2, 3 and 4. Tables 3 and 5 respectively show Pseudo F and k-NN classification accuracy of the obtained results in $d = 2$. Tables 4 and 6 respectively show Pseudo F and k-NN classification accuracy of the obtained results in $d = 3$. For Pseudo F and k-NN classification accuracy, most of the results of FDLP-CCA are larger than those of the comparison methods in both concept level and category level for all datasets. The results of FDLP-CCA are also larger than those of the comparison methods in the higher category for dataset 4. From the above results, the hierar-

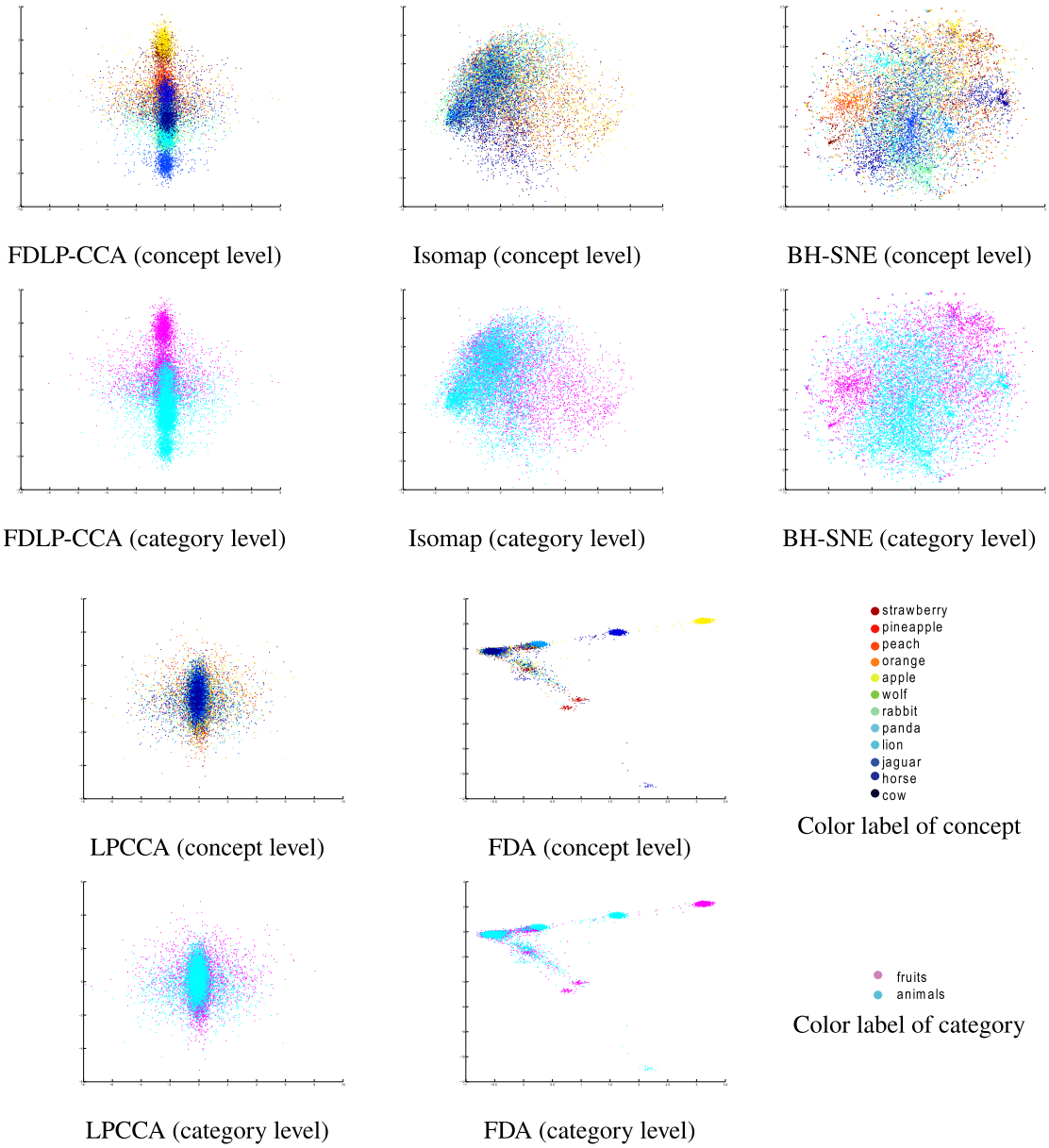


Fig. 3 Results of visualization via the dimensionality reduction methods using dataset 1.

chical image semantics are more clearly visible in FDLP-CCA. Since the performances of FDLP-CCA in $d = 2$ and 3 are better than the other dimensionality reduction methods, FDLP-CCA is the suitable dimensionality reduction method for visualization. As shown in this quantitative comparison, we can see that the proposed method performs better dimensionality reduction than do the comparison methods.

Furthermore, we show experimental results using only one of the two modalities. Specifically, we show the results via LPP [9], which is a locality preserving dimensionality reduction method, obtained by applying it to the visual feature vector or the text feature vector. Table 7 shows Pseudo F and k-NN classification accuracy. The results based on the text features are larger than the results based on the visual features. By applying it to one of the modalities and comparing their results, we confirmed that text features had high discriminative power in terms of semantics.

From the experiments, we found that FDLP-CCA can represent more semantic structures in real data than the comparative dimensionality reduction methods can. Since the comparative methods utilize a combination of different kinds of modalities for realizing dimensionality reduction, they cannot consider the unique characteristics of each modality and are often not successful in visualizing real multimedia data. On the other hand, FDLP-CCA represents the correlation between text and visual features, the local structures based on visual features and the cluster structure based on text features. As shown in Tables 3, 4, 5 and 6, we can see that the dimensionality reduction method that considers the correlation between two modalities (*i.e.*, LPCCA)

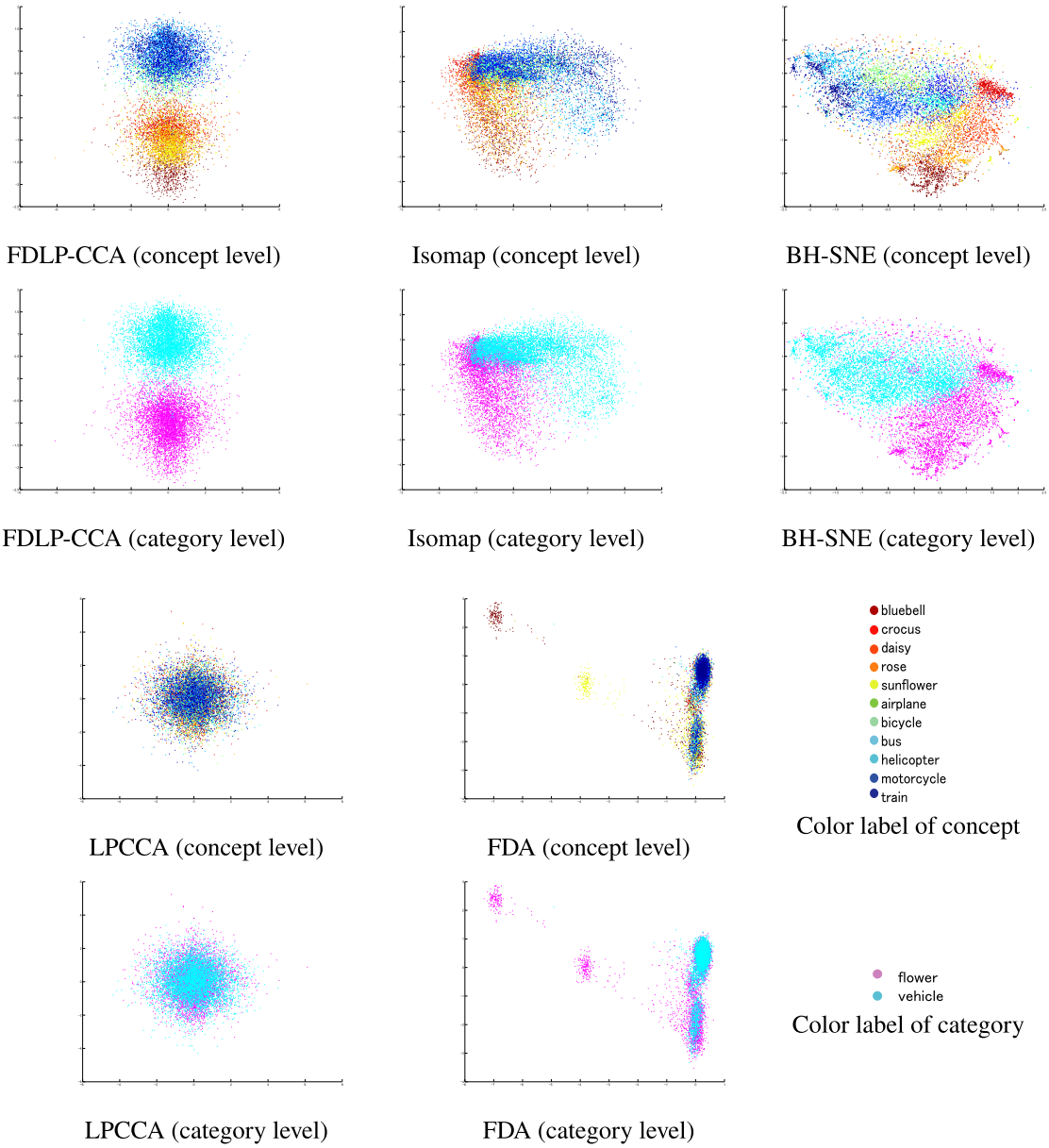


Fig. 4 Results of visualization via the dimensionality reduction methods using dataset 2.

has better performance than that of the other comparison methods. From the results, we can verify that the dimensionality reduction method which considers the correlation gives a good projection in which hierarchical image semantics are well separated. Furthermore, although the dimensionality reduction method that considers the text cluster structure (*i.e.,* FDA) fails to achieve a projection that represents hierarchical image semantics, dimensionality reduction that has the characteristics of LPCCA and FDA (*i.e.,* FDLP-CCA) has significantly better dimensionality reduction performance than that of FDA. Thus, by considering the text cluster structure, the performance of dimensionality reduction is improved. From the above results, we can verify that text cluster structure and the correlation between the text and visual features concerning the local structures are effective for visualization. Finally, we can mention here that FDLP-CCA has significantly better performance in terms of hierarchical image semantics discrimination than that of all other comparative methods.

4.3 Parameter Analysis

Finally, we report the effects of the parameter in FDLPCCA (*i.e.,* α and K). $\alpha(\in [0, 1])$ is the influential parameter which maintains the computational efficiency and reliability of FDA and LPCCA. In Fig. 5, we plot the k-NN classification accuracy of the proposed method's performance using dataset 1. The parameter α is changed from 0 to 1 increasing by 0.05. In Fig. 5, it can be seen that the performance peaks at large values of α around 0.8. Larger value of α in

Table 3 Performance comparison of results based on Pseudo F ($d = 2$).

	Hierarchical semantics	FDLP-CCA	LPCCA	FDA	MDS	Isomap	t-SNE	BH-SNE	m-SNE
Dataset 1	Concept	$\mathbf{9.905 \times 10^4}$	1.280×10^4	1.821×10^4	1.755×10^4	1.449×10^4	1.500×10^4	1.361×10^4	1.600×10^4
	Category	$\mathbf{3.390 \times 10^4}$	1.102×10^4	1.402×10^4	1.595×10^4	1.313×10^4	1.340×10^4	1.339×10^4	1.391×10^4
Dataset 2	Concept	$\mathbf{2.093 \times 10^4}$	1.104×10^4	2.252×10^4	2.694×10^4	2.281×10^4	2.528×10^4	2.760×10^4	2.770×10^4
	Category	$\mathbf{2.597 \times 10^4}$	1.074×10^4	1.599×10^4	1.931×10^4	1.730×10^4	1.836×10^4	2.046×10^4	2.156×10^4
Dataset 3	Concept	$\mathbf{3.219 \times 10^4}$	1.104×10^4	1.870×10^4	2.060×10^4	1.987×10^4	2.180×10^4	2.291×10^4	1.610×10^4
	Category	$\mathbf{2.521 \times 10^4}$	1.075×10^4	1.080×10^4	1.428×10^4	1.676×10^4	1.765×10^4	1.724×10^4	1.483×10^4
Dataset 4	Concept	$\mathbf{1.762 \times 10^4}$	1.031×10^4	1.492×10^4	1.373×10^4	1.396×10^4	1.684×10^4	1.433×10^4	1.673×10^4
	Category	$\mathbf{1.252 \times 10^4}$	1.017×10^4	1.036×10^4	1.316×10^4	1.215×10^4	1.330×10^4	1.291×10^4	1.338×10^4
	Higher Category	$\mathbf{1.255 \times 10^4}$	1.075×10^4	1.135×10^4	1.161×10^4	1.058×10^4	1.030×10^4	1.030×10^4	1.026×10^4
NUS-WIDE	Concept	$\mathbf{3.164 \times 10^4}$	1.590×10^4	1.509×10^4	1.552×10^4	1.558×10^4	1.500×10^4	1.582×10^4	1.624×10^4

Table 4 Performance comparison of results based on Pseudo F ($d = 3$).

	Hierarchical semantics	FDLP-CCA	LPCCA	FDA	MDS	Isomap	t-SNE	BH-SNE	m-SNE
Dataset 1	Concept	$\mathbf{3.543 \times 10^4}$	2.312×10^4	1.575×10^4	1.727×10^4	1.503×10^4	1.482×10^4	1.435×10^4	1.435×10^4
	Category	$\mathbf{1.912 \times 10^4}$	1.713×10^4	1.213×10^4	1.573×10^4	1.382×10^4	1.334×10^4	1.363×10^4	1.303×10^4
Dataset 2	Concept	$\mathbf{2.891 \times 10^4}$	$\mathbf{2.891 \times 10^4}$	1.228×10^4	1.311×10^4	1.417×10^4	1.095×10^4	1.701×10^4	1.400×10^4
	Category	$\mathbf{2.070 \times 10^4}$	$\mathbf{2.070 \times 10^4}$	1.126×10^4	1.137×10^4	1.240×10^4	1.078×10^4	1.593×10^4	1.355×10^4
Dataset 3	Concept	$\mathbf{2.624 \times 10^4}$	2.603×10^4	1.447×10^4	1.340×10^4	1.435×10^4	1.583×10^4	1.554×10^4	1.095×10^4
	Category	1.353×10^4	1.414×10^4	1.119×10^4	1.221×10^4	1.233×10^4	$\mathbf{1.445 \times 10^4}$	1.381×10^4	1.076×10^4
Dataset 4	Concept	$\mathbf{3.194 \times 10^4}$	2.162×10^4	1.245×10^4	1.269×10^4	1.187×10^4	1.394×10^4	1.416×10^4	1.415×10^4
	Category	$\mathbf{2.222 \times 10^4}$	$\mathbf{2.222 \times 10^4}$	1.285×10^4	1.310×10^4	1.278×10^4	1.245×10^4	1.263×10^4	1.264×10^4
	Higher Category	$\mathbf{1.386 \times 10^4}$	1.152×10^4	1.090×10^4	1.058×10^4	1.054×10^4	1.071×10^4	1.032×10^4	1.046×10^4
NUS-WIDE	Concept	$\mathbf{2.161 \times 10^4}$	$\mathbf{2.161 \times 10^4}$	1.524×10^4	1.531×10^4	1.572×10^4	1.622×10^4	1.626×10^4	1.615×10^4

Table 5 Performance comparison of results based on k-NN classification accuracy ($d = 2$).

	Hierarchical semantics	FDLP-CCA	LPCCA	FDA	MDS	Isomap	t-SNE	BH-SNE	m-SNE
Dataset 1	Concept	**0.3503**	0.2237	0.2335	0.1624	0.1228	0.1998	0.2236	0.2237
	Category	**0.7664**	0.5973	0.5799	0.6651	0.5777	0.6549	0.6725	0.6735
Dataset 2	Concept	0.2727	0.0095	0.2374	0.2763	0.2443	0.4031	**0.4532**	0.3930
	Category	**0.9712**	0.5112	0.6523	0.8343	0.7825	0.8360	0.8671	0.8527
Dataset 3	Concept	**0.3922**	0.1121	0.2846	0.2341	0.2400	0.3366	0.2438	0.3807
	Category	**0.9578**	0.9575	0.6143	0.6888	0.7422	0.7980	0.6982	0.8322
Dataset 4	Concept	0.0984	0.0322	0.0494	0.0618	0.0704	**0.1414**	0.1149	0.1273
	Category	0.3829	0.1760	0.3473	0.4599	0.4065	0.5384	0.4976	**0.5693**
	Higher Category	**0.6441**	0.3432	0.3872	0.4999	0.5065	0.5684	0.5976	0.5293
NUS-WIDE	Concept	**0.1058**	0.0420	0.0541	0.0376	0.0393	0.0333	0.0456	0.0560

Table 6 Performance comparison of results based on k-NN classification accuracy ($d = 3$).

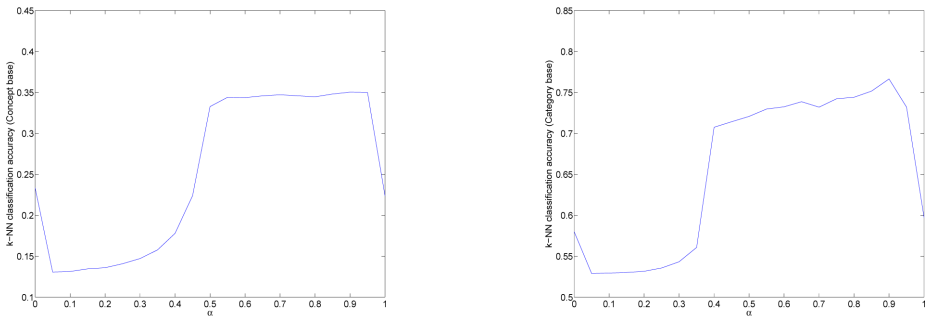
	Hierarchical semantics	FDLP-CCA	LPCCA	FDA	MDS	Isomap	t-SNE	BH-SNE	m-SNE
Dataset 1	Concept	**0.6828**	0.0981	0.1888	0.2003	0.1674	0.2642	0.2940	0.2948
	Category	**0.9541**	0.9193	0.6660	0.7033	0.6476	0.7002	0.7295	0.6543
Dataset 2	Concept	**0.5418**	**0.5418**	0.2243	0.2689	0.2801	0.0895	0.3332	0.2748
	Category	**0.9751**	**0.9751**	0.6831	0.8164	0.8071	0.5027	0.8153	0.7770
Dataset 3	Concept	**0.5778**	0.3622	0.2187	0.2137	0.2186	0.2461	0.2441	0.1905
	Category	**0.9734**	0.8136	0.6638	0.6746	0.6710	0.2461	0.2441	0.5040
Dataset 4	Concept	**0.1766**	0.1764	0.1078	0.1122	0.1098	0.1578	0.1548	0.1187
	Category	**0.4083**	0.2108	0.2517	0.3535	0.3361	0.3848	0.3808	0.3393
	Higher Category	**0.6941**	0.6686	0.4175	0.5142	0.4998	0.5326	0.5297	0.4935
NUS-WIDE	Concept	**0.0943**	0.0943	0.0436	0.0414	0.0396	0.0468	0.0468	0.0463

FDLPPCA means that the correlation between two different modalities is a primary factor of dimensionality reduction performances. Other datasets get the similar results. In addition, in the case of $d = 3$, we confirmed the similar results.

On the other hand, K is the number of clusters for applying k-means clustering to the text features. In Fig. 6, we plot the k-NN classification accuracy of the proposed method's performance using dataset 1. The number of K is changed from 5 to 15 increasing by 2. In Fig. 6, it can be seen that the performance of FDLP-CCA tends not to be sensitive to K. Other datasets get the similar results. In addition, in the case of $d = 3$, we confirmed the similar results.

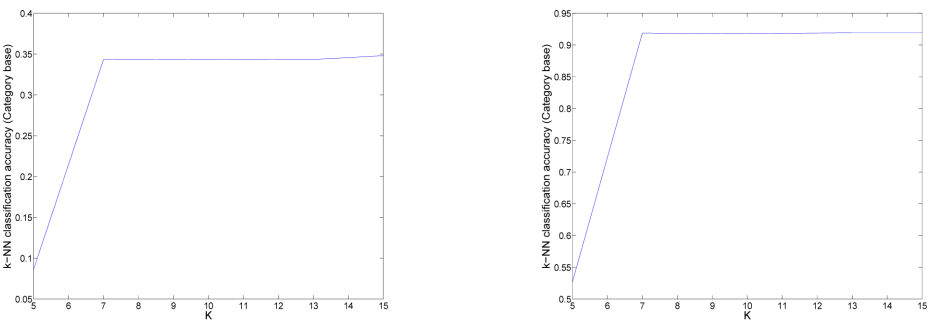
Table 7 Results of LPP using the visual features or the text features ($d = 2$).

	Hierarchical semantics	LPP (Visual)		LPP (Text)	
	Evaluation measure	Psuedo F	k-NN	Psuedo F	k-NN
Dataset 1	Concept	1.395×10^4	0.1207	1.254×10^4	0.7620
	Category	1.292×10^4	0.5747	1.013×10^4	0.8787
Dataset 2	Concept	1.416×10^4	0.1746	3.466×10^4	0.4816
	Category	1.163×10^4	0.6454	1.583×10^4	0.8939
Dataset 3	Concept	1.329×10^4	0.1326	1.612×10^4	0.7527
	Category	1.149×10^4	0.5358	1.186×10^4	0.9069
Dataset 4	Concept	1.209×10^4	0.0424	2.411×10^4	0.1788
	Category	1.045×10^4	0.1848	1.135×10^4	0.3412
	Higher Category	1.026×10^4	0.3508	1.171×10^4	0.4681
NUS-WIDE	Concept	1.569×10^4	0.0374	4.913×10^4	0.3247



Concept level Category level

Fig. 5 Results based on k-NN classification accuracy of FDLP-CCA using dataset 1 in α ($d = 2$).



Concept level Category level

Fig. 6 Results based on k-NN classification accuracy of FDLP-CCA using dataset 1 in K ($d = 2$).

5. Conclusions

In this paper, we have presented a dimensionality reduction method for data visualization. We use multimedia data that are represented by two typical modalities, text and image. The proposed dimensionality reduction method can integrate text and visual features of images and discriminate items in terms of the semantics by considering the unique characteristics of these features. Specifically, we consider the power of grouping items into semantic clusters based on the text features and the most perceptual relationships between items for humans based on the visual features. By considering the above characteristics, FDLP-CCA can integrate multiple features and discriminate them in terms of semantics. Specifically, our dimensionality reduction method is a hybrid version of two dimensionality reduction methods: LPCCA and FDA. Therefore, visualization via FDLP-CCA can group items containing the same semantics in a low-dimensional space.

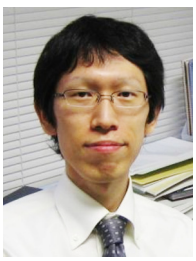
Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers JP17H01744, JP15K12023 and Grant-in-Aid for Scientific Research on Innovative Areas JP 24120001 from the MEXT.

References

[1] flickrBLOG, "Find every photo with flickrs new unified search experience." http://www.teu.ac.jp/media/ earth/FK/, Last accessed: 09/05/2016.
[2] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques, Elsevier, 2011.
[3] U. Fayyad, A. Wierse, and G. Grinstein, Information visualization in data mining and knowledge discovery, Morgan Kaufmann, 2002.
[4] K. Schoeffmann, D. Ahlström, and M.A. Hudelist, "3-d interfaces to improve the performance of visual known-item search," IEEE Trans. Multimedia, vol.16, no.7, pp.1942–1951, 2014.
[5] X. Han, C. Zhang, W. Lin, M. Xu, B. Sheng, and T. Mei, "Tree-based visualization and optimization for image collection," IEEE Trans. Cybern., vol.46, no.6, pp.1286–1300, 2016.
[6] R. Duda, P. Hart, and D. Stork, Pattern classification, John Wiley & Sons, 2012.
[7] K. Yves, Introduction to machine learning, Morgan Kaufmann, 2014.
[8] J.B. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science, vol.290, no.5500, pp.2319–2323, 2000.
[9] H. Xiaofei and N. Partha, "Locality preserving projections," Proceedings of Neural Information Processing Systems, pp.153–160, MIT Press, 2003.
[10] L. Maaten and G. Hinton, "Visualizing data using t-sne," Journal of Machine Learning Research, vol.9, pp.2579–2605, 2008.
[11] H. Ma, J. Zhu, M.R.-T. Lyu, and I. King, "Bridging the semantic gap between image contents and tags," IEEE Trans. Multimedia, vol.12, no.5, pp.462–473, 2010.
[12] L. Chu, Y. Zhang, G. Li, S. Wang, W. Zhang, and Q. Huang, "Effective multimodality fusion framework for cross-media topic detection," IEEE Trans. Circuits Syst. Video Technol., vol.26, no.3, pp.556–569, 2016.
[13] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," IEEE Trans. Pattern Anal. Mach. Intell., vol.33, no.6, pp.1147–1160, 2011.
[14] Y. Yuan and Q. Sun, "Multiset canonical correlations using globality preserving projections with applications to feature extraction and recognition," IEEE Trans. Neural Netw. Learn. Syst., vol.25, no.6, pp.1131–1146, 2014.
[15] B. Xie, Y. Mu, D. Tao, and K. Huang, "m-sne: Multiview stochastic neighbor embedding," IEEE Trans. Syst., Man, Cybern., Part B, vol.41, no.4, pp.1088–1096, 2011.
[16] L. Wu, R. Jin, and A.K. Jain, "Tag completion for image retrieval," IEEE Trans. Pattern Anal. Mach. Intell., vol.35, no.3, pp.716–727, 2013.
[17] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li, "Flickr distance," Proceedings of the 16th ACM International Conference on Multimedia, pp.31–40, ACM, 2008.
[18] T. Sun and S. Chen, "Locality preserving cca with applications to data visualization and pose estimation," Image and Vision Computing, vol.25, no.5, pp.531–543, 2007.
[19] R. Fisher, "The use of multiple measurements in taxonomic problems," Annals of eugenics, vol.7, no.2, pp.179–188, 1936.
[20] R. Sternberg, Cognitive psychology, Cengage Learning, 2008.
[21] Y. Hua, S. Wang, S. Liu, Q. Huang, and A. Cai, "Tina: Cross-modal correlation learning by adaptive hierarchical semantic aggregation," Proceedings of IEEE International Conference on Data Mining, pp.190–199, IEEE, 2014.
[22] I. Jolliffe, Principal component analysis, Wiley Online Library, 2014.
[23] W.S. Torgerson, "Multidimensional scaling: I. theory and method," Psychometrika, vol.17, no.4, pp.401–419, 1952.
[24] G. Hinton and S. Roweis, "Stochastic neighbor embedding," Proceedings of Advances in Neural Information Processing Systems, pp.833–840, 2002.
[25] L. Maaten, "Accelerating t-sne using tree-based algorithms," Journal of Machine Learning Research, vol.15, no.1, pp.3221–3245, 2014.
[26] S.T. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol.290, no.5500, pp.2323–2326, 2000.
[27] J. Fan, Y. Gao, H. Luo, D.A. Keim, and Z. Li, "A novel approach to enable semantic and visual image summarization for exploratory image search," Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, pp.358–365, ACM, 2008.
[28] Y.-H. Kuo, W.-H. Cheng, H.-T. Lin, and W.H. Hsu, "Unsupervised semantic feature discovery for image object retrieval and tag refinement," IEEE Trans. Multimedia, vol.14, no.4, pp.1079–1090, 2012.
[29] J. Deng, A. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp.785–792, IEEE, 2011.
[30] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '09, no.48, pp.1–9, ACM, 2009.
[31] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," Machine learning, vol.42, no.1, pp.177–196, 2001.
[32] D.G. Lowe, "Object recognition from local scale-invariant features," Proceedings of IEEE International Conference on Computer Vision, pp.1150–1157, IEEE, 1999.
[33] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.886–893, IEEE, 2005.
[34] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," Proceedings of ECCV International Workshop on Statistical Learning in Computer Vision, pp.1–22, Prague, 2004.
[35] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," Communications in Statistics-theory and Methods, vol.3, no.1, pp.1–27, 1974.

Kohei Tateno received the B.S. degree in electronics and information engineering from Hokkaido University, Sapporo, Japan, in 2015, where he is currently working toward the M.S. degree with the Graduate School of Information Science and Technology. His research interests include image retrieval and clustering.



Takahiro Ogawa received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Japan, in 2003, 2005, and 2007, respectively. He joined the Graduate School of Information Science and Technology, Hokkaido University, in 2008, where he is currently an Associate Professor. His research interests are multimedia signal processing and its applications. He has been an Associate Editor of the ITE Transactions on Media Technology and Applications. He is a member of the ACM, the EURASIP, the IEICE, and the ITE.

Miki Haseyama received the B.S., M.S., and Ph.D. degrees in electronics from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University, as an Associate Professor in 1994. She was a Visiting Associate Professor with Washington University, USA, from 1995 to 1996. She is currently a Professor with the Graduate School of Information Science and Technology, Hokkaido University. Her current research interests include image and video processing and its development into semantic analysis. She has been the Vice President of the Institute of Image Information and Television Engineers (ITE), Japan, an Editor-in-Chief of the ITE Transactions on Media Technology and Applications, and the Director of the International Coordination and Publicity, Institute of Electronics, Information, and Communication Engineers (IEICE). She is a member of the IEICE, the ITE, and the ASJ.