Regular Paper

SemiCCA: Efficient Semi-supervised Learning of Canonical Correlations*

Akisato Kimura[1,a)] Masashi Sugiyama[2] Takuho Nakano[1,3]
Hirokazu Kameoka[1,3] Hitoshi Sakano[1] Eisaku Maeda[1]
Katsuhiko Ishiguro[1]

Received: August 22, 2012, Revised: October 12, 2012,
Accepted: November 26, 2012

Abstract: Canonical correlation analysis (CCA) is a powerful tool for analyzing multi-dimensional paired data. However, CCA tends to perform poorly when the number of paired samples is limited, which is often the case in practice. To cope with this problem, we propose a semi-supervised variant of CCA named *SemiCCA* that allows us to incorporate additional unpaired samples for mitigating overfitting. Advantages of the proposed method over previously proposed methods are its computational efficiency and intuitive operationality: it smoothly bridges the generalized eigenvalue problems of CCA and principal component analysis (PCA), and thus its solution can be computed efficiently just by solving a single eigenvalue problem as the original CCA.

Keywords: Canonical correlation analysis, semi-supervised learning, generalized eigenproblem, principal component analysis, multi-label prediction

1. Introduction

The goal of dimensionality reduction is to obtain a low-dimensional representation of high-dimensional data samples, while preserving most of the intrinsic information contained in the original data. If dimensionality reduction is carried out appropriately, the compact representation of the data can be used for various tasks, such as visualization, noise reduction and classification.

Analyzing high-dimensional co-occurring data $(\boldsymbol{x}, \boldsymbol{y})$ is an important challenge in machine learning and pattern recognition communities, e.g., in the context of multi-view learning [1], automatic annotation of music, image and video [2], [3], [4], and sensor data mining [5], [6], [7], [8]. Canonical correlation analysis (CCA) [9] is a classical but still powerful tool for analyzing multivariate paired samples. CCA finds projection bases $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ such that correlation between projected samples $\boldsymbol{w}_x^\top \boldsymbol{x}$ and $\boldsymbol{w}_y^\top \boldsymbol{y}$ is maximized. However, the performance of CCA tends to be degraded when the number of paired samples $(\boldsymbol{x}, \boldsymbol{y})$ is limited, while a large number of additional *unpaired* samples (i.e., $\boldsymbol{x}$-only samples and $\boldsymbol{y}$-only samples) are often a lot in real-world applications. For example, in the case of automatic image annotation, collecting many labeled images (= paired samples $(\boldsymbol{x}, \boldsymbol{y})$) is often hard, while unlabeled images (= unpaired samples $\boldsymbol{x}$) can be easily obtained abundantly. In the case of sensor data mining, data tends to be lost due to faulty devices and unstable transmissions, which produces a lot of unpaired samples.

To utilize such additional unpaired samples, Blaschko et al. [10] proposed a semi-supervised extension of kernelized CCA [11], [12] by the use of Laplacian regularization. This method enables us to find highly correlated directions that are also located on high variance directions along the data manifold. However, it is specialized to kernelized CCA, and deriving semi-supervised variants of the standard (linear) CCA is not necessarily straightforward.

This paper proposes quite a simple method to extend linear CCA to semi-supervised one, that we call *SemiCCA*. The proposed method SemiCCA utilizes additional unpaired samples by smoothly bridging CCA and principal component analysis (PCA). More specifically, the generalized eigenvalue problems of CCA and PCA are combined using a trade-off parameter. Thus the solution of SemiCCA can still be obtained just by solving the combined eigenvalue problem, which is the same computational complexity as the original CCA. We note that several publications discuss other types of semi-supervised variants of canonical correlation analysis (e.g., Refs. [13], [14]) that require fully paired samples $(\boldsymbol{x}, \boldsymbol{y})$ and additional side information associated with samples. Thus, their settings are different from the current paper.

The rest of this paper is organized as follows: Section 2 reviews the standard CCA briefly as an introduction of the proposed

[1] NTT Communication Science Laboratories, NTT Corporation, Soraku, Kyoto 619–0237, Japan
[2] Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Meguro, Tokyo 152–8552, Japan
[3] Graduate School of Information Science and Technologies, the University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[a)] akisato@ieee.org

* A prerilimary version of this work have been presented in Conference on International Association for Pattern Recognition (ICPR2010).

method SemiCCA. Section 3 describes the proposed method SemiCCA in detail. This section also introduces several extensions of SemiCCA to more than two samples sets and nonlinear analysis with a kernel trick. Section 4 reports several experimental results with randomly generated data, and examines several fundamental properties and the effectiveness of the proposed method. Section 5 considers some applications of the proposed method to automatic image/audio annotation, and reports the quantitative evaluations. Section 6 concludes this paper and discusses promising future work.

2. Reviewing Canonical Correlation Analysis (CCA)

Consider a set of paired samples of size N,

$$\boldsymbol{X}_P = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N),$$
$$\boldsymbol{Y}_P = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N),$$

where each sample is a real-valued vector with dimension D_x and D_y, and a pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ of samples with the same suffix is co-occurring. Without loss of generality, we assume that $\boldsymbol{X}_P$ and $\boldsymbol{Y}_P$ are both centered, that can always be achieved by subtracting the sample means from each sample. CCA is a method of finding a pair $(\boldsymbol{w}_x, \boldsymbol{w}_y) \in \mathcal{R}^{D_x} \times \mathcal{R}^{D_y}$ of basis vectors for a given set $(\boldsymbol{X}_P, \boldsymbol{Y}_P)$ of paired samples such that their normalized correlation is maximized as follows:

$$\begin{aligned}\rho(\boldsymbol{X}_P, \boldsymbol{Y}_P) &= \max_{(\boldsymbol{w}_x,\boldsymbol{w}_y)\in\mathcal{R}^{D_x}\times\mathcal{R}^{D_y}} \frac{\left\langle \boldsymbol{X}_P^\top \boldsymbol{w}_x, \boldsymbol{Y}_P^\top \boldsymbol{w}_y \right\rangle}{\left\|\boldsymbol{X}_P^\top \boldsymbol{w}_x\right\|_F \cdot \left\|\boldsymbol{Y}_P^\top \boldsymbol{w}_y\right\|_F} \\ &= \max_{(\boldsymbol{w}_x,\boldsymbol{w}_y)\in\in\mathcal{R}^{D_x}\times\mathcal{R}^{D_y}} \frac{\boldsymbol{w}_x^\top \boldsymbol{S}_{Pxy} \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^\top \boldsymbol{S}_{Pxx} \boldsymbol{w}_x}\sqrt{\boldsymbol{w}_y^\top \boldsymbol{S}_{Pyy} \boldsymbol{w}_y}},\end{aligned}$$

where $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is the inner product of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, $\|X\|_F$ is the Frobenius norm of a matrix $\boldsymbol{X}$, $\boldsymbol{X}^\top$ is a transpose of a matrix $\boldsymbol{X}$, $\boldsymbol{S}_{Pxx}$, $\boldsymbol{S}_{Pyy}$ and $\boldsymbol{S}_{Pxy}$ are sample covariance matrices of paired samples

$$\boldsymbol{S}_{Pxx} = \boldsymbol{X}_P \boldsymbol{X}_P^\top / N, \qquad \boldsymbol{S}_{Pyy} = \boldsymbol{Y}_P \boldsymbol{Y}_P^\top / N,$$
$$\boldsymbol{S}_{Pxy} = \boldsymbol{X}_P \boldsymbol{Y}_P^\top / N.$$

The maximizers of the function $\rho(\boldsymbol{X}_P, \boldsymbol{Y}_P)$ with respect to $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ are not affected by re-scaling $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$. Therefore, the maximization of $\rho(\boldsymbol{X}_P, \boldsymbol{Y}_P)$ is equivalent to maximizing the numerator $\boldsymbol{w}_x^\top \boldsymbol{S}_{Pxy} \boldsymbol{w}_y$ of $\rho(\boldsymbol{X}_P, \boldsymbol{Y}_P)$ subject to

$$\boldsymbol{w}_x^\top \boldsymbol{S}_{Pxx} \boldsymbol{w}_x = \boldsymbol{w}_y^\top \boldsymbol{S}_{Pyy} \boldsymbol{w}_y = 1.$$

Taking derivatives of the corresponding Lagrangian with respect to $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$, we obtain

$$\boldsymbol{S}_{Pxy} \boldsymbol{w}_y - \lambda \boldsymbol{S}_{Pxx} \boldsymbol{w}_x = \boldsymbol{0},$$
$$\boldsymbol{S}_{Pxy}^\top \boldsymbol{w}_x - \lambda \boldsymbol{S}_{Pyy} \boldsymbol{w}_y = \boldsymbol{0}.$$

Therefore, the solution $(\boldsymbol{w}_x, \boldsymbol{w}_y)$ is given as the solution of the following generalized eigenvalue problem:

$$\begin{pmatrix} \boldsymbol{0} & \boldsymbol{S}_{Pxy} \\ \boldsymbol{S}_{Pxy}^\top & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix} = \lambda \begin{pmatrix} \boldsymbol{S}_{Pxx} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{Pyy} \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix}. \qquad (1)$$

Picking up the top $D_z(\leq \min(D_x, D_y))$ generalized eigenvectors as row vectors, we can obtain D_z-dimensional mappings $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$.

3. Proposed Method: SemiCCA

3.1 Semi-supervised Setup

As described in the previous section, CCA can handle paired samples, any types of co-occurring real-valued sample pairs, where several additional unpaired samples are available.

Let us explain the idea of SemiCCA using an illustrative two-dimensional data set depicted in **Fig. 1**, where paired (resp. unpaired) samples are plotted with white (resp. red and blue). When only the paired samples $(\boldsymbol{X}_P, \boldsymbol{Y}_P)$ are used, poor projection bases may be obtained by CCA due to overfitting, as shown by the black arrows in Fig. 1. In contrast, unpaired samples

$$\begin{aligned}\boldsymbol{X}_U &= (\boldsymbol{x}_{N+1}, \boldsymbol{x}_{N+2}, \ldots, \boldsymbol{x}_{N+N_x}) \\ &= (\boldsymbol{x}_{U,1}, \boldsymbol{x}_{U,2}, \ldots, \boldsymbol{x}_{U,N_x}), \\ \boldsymbol{Y}_U &= (\boldsymbol{y}_{N+N_x+1}, \boldsymbol{y}_{N+N_x+2}, \ldots, \boldsymbol{y}_{N+N_x+N_y}) \\ &= (\boldsymbol{y}_{U,1}, \boldsymbol{y}_{U,2}, \ldots, \boldsymbol{y}_{U,N_y})\end{aligned}$$

can be used for revealing the global structure in each domain, as shown by the colored arrows in Fig. 1. Note once rectification of a basis in one sample space affects its counterpart in the other sample space because of the correlation maximizing nature of CCA (cf. the dotted arrows in Fig. 1).

3.2 Algorithm

Motivated by the above illustration, we develop a novel method for effectively incorporating unpaired samples into the original CCA. The proposed method, SemiCCA, combines CCA for only the paired samples and PCA, one of the major tools to capture the global structure of samples in an unsupervised manner. More specifically, we integrate the eigenvalue problems of CCA and PCA since this allows us to compute the combined solution efficiently. The solution of SemiCCA is given by the leading generalized eigenvectors of the following generalized eigenvalue problem:

$$\overline{\boldsymbol{C}} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix} = \lambda \underline{\boldsymbol{C}} \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{pmatrix}, \qquad (2)$$

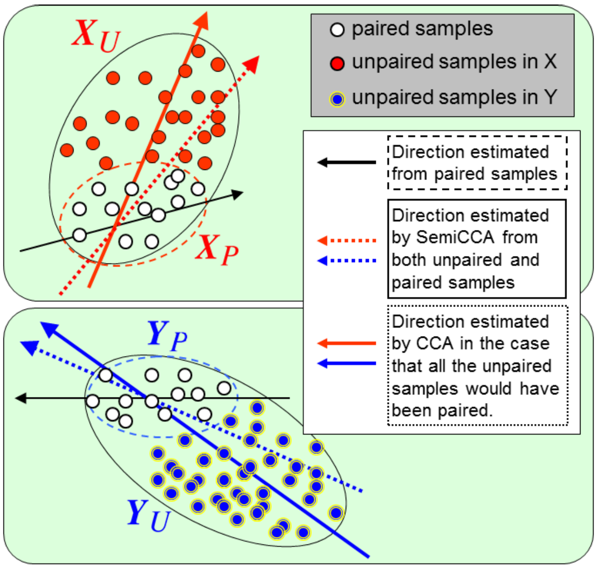
X_U
X_P
Y_P
Y_U
○ paired samples
● unpaired samples in X
● unpaired samples in Y
Direction estimated from paired samples
Direction estimated by SemiCCA from both unpaired and paired samples
Direction estimated by CCA in the case that all the unpaired samples would have been paired.

Fig. 1 Effects of unpaired samples in SemiCCA.

where

$$\overline{\boldsymbol{C}} = \beta\begin{pmatrix}\boldsymbol{0} & \boldsymbol{S}_{Pxy}\\ \boldsymbol{S}_{Pxy}^{\top} & \boldsymbol{0}\end{pmatrix} + (1-\beta)\begin{pmatrix}\boldsymbol{S}_{xx} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}_{yy}\end{pmatrix},$$

$$\underline{\boldsymbol{C}} = \beta\begin{pmatrix}\boldsymbol{S}_{Pxx} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}_{Pyy}\end{pmatrix} + (1-\beta)\begin{pmatrix}\boldsymbol{I}_{D_x} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_{D_y}\end{pmatrix},$$

S_{xx} and S_{yy} are sample covariance matrices of all the pairs

$$\boldsymbol{S}_{xx} = \left(\boldsymbol{X}_P\boldsymbol{X}_P^{\top} + \boldsymbol{X}_U\boldsymbol{X}_U^{\top}\right)\Big/ N_x,$$

$$\boldsymbol{S}_{yy} = \left(\boldsymbol{Y}_P\boldsymbol{Y}_P^{\top} + \boldsymbol{Y}_U\boldsymbol{Y}_U^{\top}\right)\Big/ N_y,$$

and β is a constant named *a trade-off parameter* taking a value in $[0, 1]$. The parameter β controls the trade-off between CCA and PCA. Namely, when $\beta = 1$, Eq. (2) is reduced to the CCA eigenvalue problem Eq. (1), while when $\beta = 0$ Eq. (2) is reduced to the PCA eigenvalue problem, under the assumption that $\boldsymbol{X} = (\boldsymbol{X}_P, \boldsymbol{X}_U)$ and $\boldsymbol{Y} = (\boldsymbol{Y}_P, \boldsymbol{Y}_U)$ are uncorrelated. In general, SemiCCA with a trade-off parameter $0 < \beta < 1$ inherits the properties of both CCA and PCA so that the global structure in each domain and the co-occurrence information of paired samples are smoothly controlled.

One may use different trade-off parameters for $\overline{\boldsymbol{C}}$ and $\underline{\boldsymbol{C}}$ to increase the flexibility. However, this makes the trade-off parameter choice laborious. For this reason, we focus on using the single shared trade-off parameter β for both $\overline{\boldsymbol{C}}$ and $\underline{\boldsymbol{C}}$, as the first step.

3.3 Some Extensions

We have focused on the case where two sets of samples are given so far. However, the proposed method SemiCCA can be easily extended to multiple data sets by considering correlations over all pairs of samples [15]. For example, we can formulate SemiCCA for a triad $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ of sample sets, as follows:

$$\overline{\boldsymbol{C}}\begin{pmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\\ \boldsymbol{w}_z\end{pmatrix} = \lambda\underline{\boldsymbol{C}}\begin{pmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\\ \boldsymbol{w}_z\end{pmatrix},$$

where

$$\overline{\boldsymbol{C}} = \beta\begin{pmatrix}\boldsymbol{0} & \boldsymbol{S}_{(P)xy} & \boldsymbol{S}_{(P)xz}\\ \boldsymbol{S}_{(P)xy}^{\top} & \boldsymbol{0} & \boldsymbol{S}_{(P)yz}\\ \boldsymbol{S}_{(P)xz}^{\top} & \boldsymbol{S}_{(P)yz}^{\top} & \boldsymbol{0}\end{pmatrix} + (1-\beta)\begin{pmatrix}\boldsymbol{S}_{xx} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}_{yy} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{S}_{zz}\end{pmatrix},$$

$$\underline{\boldsymbol{C}} = \beta\begin{pmatrix}\boldsymbol{S}_{(P)xx} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}_{(P)yy} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{S}_{(P)zz}\end{pmatrix} + (1-\beta)\begin{pmatrix}\boldsymbol{I}_{D_x} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_{D_y} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{D_z}\end{pmatrix}.$$

Of course, the above discussion can be applied to more than 3 sample set in the same way.

We can also obtain a kernelized variant of SemiCCA by using the standard kernel trick and the technique of pairwise expression [16]. A covariance matrix $\boldsymbol{S}_{xy}$ can be converted to the following pairwise expression (see Ref. [16] for details):

$$\boldsymbol{S}_{xy} = \boldsymbol{X}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{X}^{\top} = \boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\top},$$

where $\boldsymbol{W}$ is a matrix so that all the elements are 1, $\boldsymbol{D}$ is a diagonal matrix so that the n-th diagonal element is $D_{n,n} = \sum_{m=1}^{N} W_{n,m}$, and $\boldsymbol{L}$ is called a graph Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$. In the same way, the identity matrix can be expressed with the following pairwise form:

$$\boldsymbol{I}_{D_x} = \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{\dagger}\boldsymbol{X}^{\top},$$

where $\boldsymbol{X}^{\dagger}$ denotes the Moore-Penrose generalized inverse of a matrix $\boldsymbol{X}$. Therefore, we can express the eigenvalue problem (Eq. (2)) solved in SemiCCA as

$$\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{pmatrix}\overline{\boldsymbol{L}}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{pmatrix}^{\top}\begin{pmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\end{pmatrix} = \lambda\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{pmatrix}\underline{\boldsymbol{L}}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{pmatrix}^{\top}\begin{pmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\end{pmatrix}, \qquad (3)$$

where

$$\overline{\boldsymbol{L}} = \beta\begin{pmatrix}\boldsymbol{0} & \boldsymbol{L}_{(P)xy}\\ \boldsymbol{L}_{(P)xy}^{\top} & \boldsymbol{0}\end{pmatrix} + (1-\beta)\begin{pmatrix}\boldsymbol{L}_{xx} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{L}_{yy}\end{pmatrix},$$

$$\underline{\boldsymbol{L}} = \begin{pmatrix}\boldsymbol{L}_{(P)xx} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{L}_{(P)yy}\end{pmatrix} + (1-\beta)\begin{pmatrix}(\boldsymbol{X}^{\top}\boldsymbol{X})^{\dagger} & \boldsymbol{0}\\ \boldsymbol{0} & (\boldsymbol{Y}^{\top}\boldsymbol{Y})^{\dagger}\end{pmatrix}.$$

Here, we introduce the following expressions with appropriate vectors $\boldsymbol{\alpha}_X, \boldsymbol{\alpha}_Y \in \mathcal{R}^N$ as follows:

$$\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{pmatrix}^{\top}\begin{pmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{pmatrix}^{\top}\begin{pmatrix}\boldsymbol{X}\boldsymbol{\alpha}_x\\ \boldsymbol{Y}\boldsymbol{\alpha}_y\end{pmatrix} = \begin{pmatrix}\boldsymbol{K}_x\boldsymbol{\alpha}_x\\ \boldsymbol{K}_y\boldsymbol{\alpha}_y\end{pmatrix}$$

where $\boldsymbol{K}_x = \{K_{x(i,j)}\}_{i,j=1}^{N}$ and $\boldsymbol{K}_y = \{K_{y(i,j)}\}_{i,j=1}^{N}$ are $N \times N$ matrices with

$$K_{x(i,j)} = \boldsymbol{x}_i^{\top}\boldsymbol{x}_j, \qquad K_{y(i,j)} = \boldsymbol{y}_i^{\top}\boldsymbol{y}_j.$$

Then, multiplying Eq. (3) by $(\boldsymbol{X}^{\top}, \boldsymbol{Y}^{\top})$ from the left-hand side yields

$$\begin{pmatrix}\boldsymbol{K}_x\\ \boldsymbol{K}_y\end{pmatrix}\overline{\boldsymbol{L}}\begin{pmatrix}\boldsymbol{K}_x\\ \boldsymbol{K}_y\end{pmatrix}^{\top}\begin{pmatrix}\boldsymbol{\alpha}_x\\ \boldsymbol{\alpha}_y\end{pmatrix} = \lambda\begin{pmatrix}\boldsymbol{K}_x\\ \boldsymbol{K}_y\end{pmatrix}\underline{\boldsymbol{L}}\begin{pmatrix}\boldsymbol{K}_x\\ \boldsymbol{K}_y\end{pmatrix}^{\top}\begin{pmatrix}\boldsymbol{\alpha}_x\\ \boldsymbol{\alpha}_y\end{pmatrix}.$$

The above equation implies that the samples appear only via their inner products, which means that $\boldsymbol{K}_x = \{K_{x(i,j)}\}_{i,j=1}^{N}$ and $\boldsymbol{K}_y = \{K_{y(i,j)}\}_{i,j=1}^{N}$ can be replaced by Gram matrices, each of whose component can be decomposed with a pair (ϕ_x, ϕ_y) of functions as

$$K_{x(i,j)} = K_x(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi_x(\boldsymbol{x}_i)^{\top}\phi_x(\boldsymbol{x}_j),$$

$$K_{y(i,j)} = K_y(\boldsymbol{y}_i, \boldsymbol{y}_j) = \phi_y(\boldsymbol{y}_i)^{\top}\phi_y(\boldsymbol{y}_j).$$

The kernelized version of SemiCCA can be integrated into the work by Blaschko et al. [10] with the introduction of Laplacian regularization to inhibit overfitting.

4. Preliminary Investigation

We first investigated the fundamental characteristics of the proposed method exhaustively using artificial data sets created as follows: Consider a simple Gaussian latent model, where the latent random variable (corresponding to a canonical variable in the framework of CCA) is denoted by Z and the observable random variables are X and Y. We drew samples $\{z_i\}_{i=1}^{N_z}$ from a standard normal distribution independently, $z_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{D_z})$, where $D_z = 10$ is the dimension of the latent random variable Z. The number N_z of samples was set to $N_z = 10{,}000$. Then complete paired samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_z}$ were created as

$$\boldsymbol{x}_i = \boldsymbol{T}_x z_i + \overline{\boldsymbol{x}} + \boldsymbol{\delta}_{x,i}, \qquad \boldsymbol{\delta}_{x,i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{X|Z}),$$

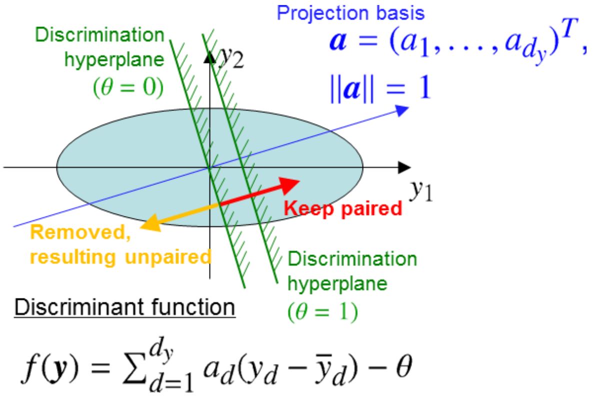
Projection basis
$a = (a_1, \ldots, a_{d_y})^T$,
$\|a\| = 1$
Discrimination hyperplane
($\theta = 0$)
y_2
y_1
Keep paired
Removed, resulting unpaired
Discrimination hyperplane
($\theta = 1$)
Discriminant function
$f(y) = \sum_{d=1}^{d_y} a_d(y_d - \bar{y}_d) - \theta$

Fig. 2 How to generate artificial data.

$$y_i = T_y z_i + \bar{y} + \delta_{y,i}, \qquad \delta_{y,i} \sim \mathcal{N}(\mathbf{0}, \Sigma_{Y|Z}),$$

where each component of transformation matrices (T_x and T_y), means ($\bar{x}$ and $\bar{y}$), and covariance matrices ($\Sigma_{X|Z}$ and $\Sigma_{Y|Z}$) was generated from the folded standard normal distribution. The dimensions of the samples were set to $D_x = 15$ and $D_y = 20$.

Then, we removed several samples from $\{y_i\}_{i=1}^{N_z}$ to artificially generate unpaired samples, as depicted in **Fig. 2**. Here, we used the following linear hyperplane $f(\cdot)$ to remove samples:

$$f(y) = \sum_{d=1}^{D_y} a_d(y_d - \bar{y}_d) - \eta, \tag{4}$$

where $a = (a_1, \ldots, a_{D_y})^\top$ is a coefficient vector satisfying $\|a\| = 1$, and η is a threshold value such that the larger η is, the more samples are removed. A sample (x_i, y_i) was kept paired if $f(y_i) > 0$, and y_i was removed otherwise. The above scheme can simulate the situation where the distributions of paired and unpaired samples are quite different with each other. Note that when their distributions are almost the same, CCA with only the paired samples asymptotically yields the same output as that with complete paired samples (before removing several samples).

We measured the performance of (Semi)CCA by the weighted sum of cosine distances defined as follows:

$$\sum_{i=1}^{r} \lambda_i^* \frac{w_{x,i}^\top w_{x,i}^*}{\|w_{x,i}\| \cdot \|w_{x,i}^*\|},$$

where $w_{x,i}$ $(i = 1, 2, \ldots, r)$ are the eigenvectors derived by (Semi)CCA from the paired and unpaired samples, and $w_{x,i}^*$ and λ_i^* are the eigenvectors and eigenvalues derived by the standard CCA from the complete paired samples. We took an oracle setting for selecting the trade-off parameter β. Namely, we adopted the trade-off parameter β marking the highest score for each trial. Note that our proposed method SemiCCA is directly formulated as a generalized eigenproblem, which imples that it does not have any explicit objective functions. Due to the lack of explicit objective functions, we cannot compare the performance of SemiCCA for different trade-off paramters β. This is the main reason why we took the oracle setting for this preliminary investigation.

Figure 3 shows the evaluation scores averaged over 10,000 independent trials for several discrimination thresholds θ, each of which corresponds to the average number of paired samples. The results indicate the potential of the proposed method SemiCCA: if we can appropriately select the trade-off parameters β, SemiCCA can outperform the standard CCA. It is noteworthy that even when the number of unpaired samples is not so large, SemiCCA has a potential to perform better than the original CCA.

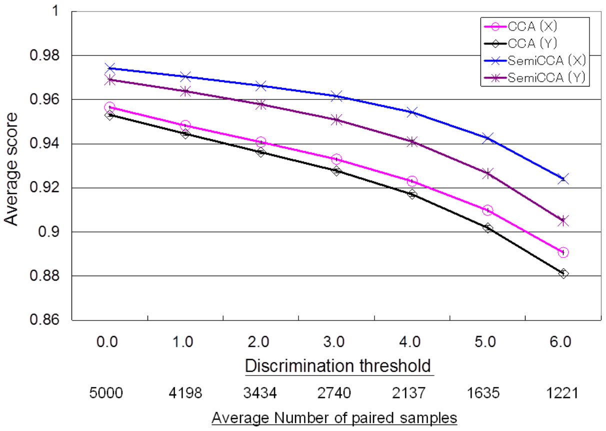

Fig. 3 Average evaluation score for artificial data.

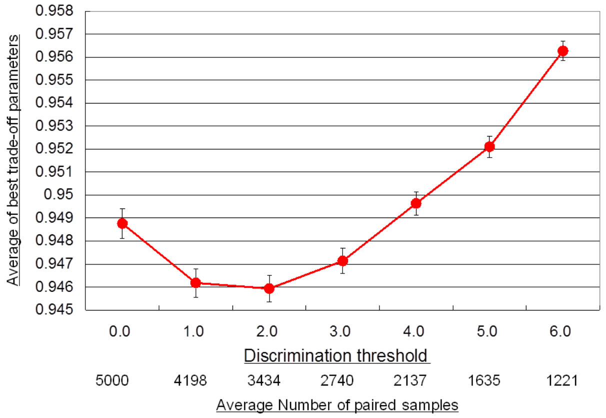

Fig. 4 Average trade-off parameter taking the highest score.

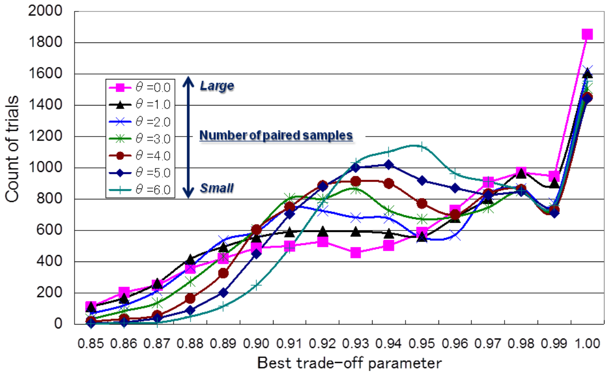

Fig. 5 Histogram of trade-off parameters taking the highest score.

Figure 4 shows the trade-off parameter taking the highest score averaged over all the trials, and **Fig. 5** depicts the histogram of the best trade-off parameters. The results imply that the best trade-off parameters have a concave profile with respect to the number of paired samples. Since standard errors of the best trade-off parameters were relatively small, we expect to obtain similar results not only for oracle settings but also for cross validation scenarios. Namely, this result provides us a generic guideline to estimate promising trade-off parameters from the ratio of paired and unpaired samples. The guideline can be useful for several applications, as shown in the next section. The experimental results also indicate that the average of the best trade-off parameters were usually close to 1. Namely, the PCA term scaled by $(1 - \beta)$ took a role of regularizing CCA with only paired samples.

5. Applications to Multi-label Prediction

5.1 Method

We applied the proposed method SemiCCA to multi-label prediction, and evaluated its performance with automatic annotation of images and audios. The baseline was proposed by Nakayama et al. [17] and Harada et al. [18], which is based on a simple latent model with the same structure as *probabilistic Latent Semantic Analysis* (pLSA) [19], [20].

Feature vectors were extracted from images/audios $\boldsymbol{G} = \{\boldsymbol{g}_n\}_{n=1}^{N_x}$ and associated text labels $\boldsymbol{W} = \{\boldsymbol{w}_n\}_{n=1}^{N}$, where N is the number of labeled samples and N_x is the total number of samples including labeled and unlabeled samples (should be $N \leq N_x$ and in most cases $N \ll N_x$). Each text label $\boldsymbol{w}_n$ was composed of text words selected from a word set given in advance. We utilized *Bag of Features* (BoF) with dimension $D_x = 1{,}024$ as image/audio features $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{N_x}$. As a fundamental feature comprising a BoF, the *Speed Up Robust Features* (SURF) algorithm [21] was used for image local descriptors, and *Mel-frequency Cepstral Coefficients* (MFCCs), the first and second instantaneous derivatives (Δ- and ΔΔ-MFCC) were used for audio frame features. We adopted word existence vectors $\boldsymbol{Y} = \{\boldsymbol{y}_n\}_{n=1}^{N}$ as text features, where each element represents an existence or absence of a specific word and thus the dimension D_y of text features was equal to the number of classes (= 20).

Next, a latent model was estimated from feature vectors $(\boldsymbol{X}, \boldsymbol{Y})$ with the help of (Semi)CCA. The first step was to generate latent variables $\boldsymbol{Z} = \{z_n\}_{n=1}^{N_x}$ with (Semi)CCA. More specifically, a function $f_x : \mathcal{R}^{D_x} \rightarrow \mathcal{R}^{D_z}$ was derived from $(\boldsymbol{X}, \boldsymbol{Y})$ as training samples with SemiCCA, and latent variables $\boldsymbol{Z}$ are generated from $(\boldsymbol{X}, \boldsymbol{Y})$ with f_x. The dimension D_z of latent variables was experimentally determined as $D_z = 20$. Here, we set the function f_x as

$$f_x(\boldsymbol{x}) = \boldsymbol{\Lambda}^{1/2} \boldsymbol{W}_x \boldsymbol{x},$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix that contains eigenvalues in diagonal components. The second step was to set up a latent model. The latent model was described by the following equations:

$$\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{y}) &= \frac{1}{N_x} \sum_{n=1}^{N_x} p(\boldsymbol{x}|z_n) p(\boldsymbol{y}|z_n), \\
p(\boldsymbol{x}|z_n) &\propto \exp\left(-\frac{\|f_x(\boldsymbol{x}) - z_n\|}{2\gamma^2}\right), \\
p(\boldsymbol{y}|z_n) &= \prod_{d=1}^{d_y} p(y_d|z_n), \\
p(y_d = 1|z_n) &= \mu\delta(1 - y_{n,d}) + (1 - \mu)N_d/N, \\
p(y_d = 0|z_n) &= 1 - p(y_d = 1|z_n),
\end{aligned}$$

where $y_{n,d}$ is the d-th element of $\boldsymbol{y}_n$, N_d is the number of images containing the d-th word in labeled samples, μ is a parameter representing how reliable a given label is, δ is the Kronecker delta, an operator $\propto$ stands for proportion, and γ is a positive constant. According to the preceding study [17], we set $\mu = 0.99$ and $\gamma = 1.0$.

Once the model estimation has been finished, we can execute image annotation within the same framework through *maximum a posteriori* (MAP) estimation. More specifically, the text feature $\widehat{\boldsymbol{y}}$ of the most probable text label $\widehat{\boldsymbol{w}}$ can be derived by using a feature $\boldsymbol{x}^{(g)}$ extracted from a given image or audio, as follows:

$$\begin{aligned}
\widehat{\boldsymbol{y}} &= \operatorname*{argmax}_{\boldsymbol{y} \in [0,1]^{D_y}} p(\boldsymbol{y}|\boldsymbol{x}^{(g)}) \\
&= \operatorname*{argmax}_{\boldsymbol{y} \in [0,1]^{D_y}} \frac{\sum_{n=1}^{N_x} p(\boldsymbol{x}^{(g)}|z_n) p(\boldsymbol{y}|z_n)}{\sum_{n=1}^{N_x} p(\boldsymbol{x}^{(g)}|z_n)}.
\end{aligned}$$

Since a conditional density $p(\boldsymbol{y}|z_n)$ for a text feature $\boldsymbol{y}$ is modeled as an element-wise independent distribution

$$p(\boldsymbol{y}|z_n) = \prod_{d=1}^{D_y} p(y_d|z_n),$$

the annotation problem can be rewritten to the following simple form:

$$\widehat{y}_d = \frac{\sum_{n=1}^{N_x} p(\boldsymbol{x}^{(g)}|z_n) p(y_d = 1|z_n)}{\sum_{n=1}^{N_x} p(\boldsymbol{x}^{(g)}|z_n)} \quad (d = 1, 2, \ldots, D_y)$$

When $\widehat{y}_d$ $(d = 1, 2, \ldots, D_y)$ exceeds a pre-defined threshold θ_d, the text word of index d is provided to the given image or audio.

5.2 Automatic Image Annotation

We use the dataset used in PASCAL Visual Object Challenge (VOC) 2008 [22] and 2009 [23] for the experiments of automatic image annotation, which consists of 20 binary classification tasks of identifying the existence of a specific object in each image. Image examples included in the VOC2008 training dataset is shown in **Fig. 6**. We utilized all of the 5,096 images in the VOC2008 training dataset, and separated them into 1,000 labeled images for training, 500 unlabeled images for evaluation and the rest as unlabeled images for training. Also, 9,647 images in the VOC2009 training/test dataset [*1] were added to unlabeled images for training. In total, 13,743 unlabeled images for training were utilized. We removed all the bounding boxes and only utilized class labels associated with bounding boxes to simulate weak labeling settings [24], where images are weakly related to multiple words without region information. We adopted the precision rate *PR* and recall rate *RE* as the evaluation measures, defined as

Fig. 6 Example images in VOC2008 dataset.

[*1] The VOC2009 training dataset contains all the images in VOC2008 training dataset, however, those duplicated images were removed in advance.

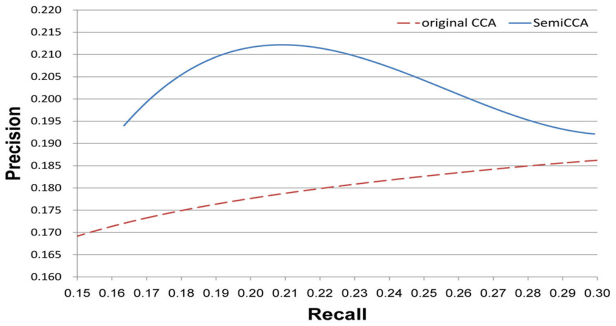

Fig. 7 Precision-recall curve for automatic image annotation with PASCAL VOC 2008/2009 dataset.

$$PR = \frac{\sum_{n=1}^{N_e} TP_n}{\sum_{n=1}^{N_e}(TP_n + FP_n)}, \tag{5}$$

$$RE = \frac{\sum_{n=1}^{N_e} TP_n}{\sum_{n=1}^{N_e}(TP_n + FN_n)}, \tag{6}$$

where N_e (= 500) is the number of images for evaluation, TP_n, FP_n and FN_n is respectively true positives, false positives and false negatives for the n-th image for evaluation. We measured the precision and recall rate for various threshold vectors $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{D_y})^\top$ whose range was from $0.5\hat{\boldsymbol{\theta}}$ to $2.0\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is the threshold vector that achieved the best balance of the precision and recall rates for the training dataset. In general the large threshold would achieve high precision rate but low recall rate.

Figure 7 shows the experimental results, where the horizontal axis stands for the recall rate, and the vertical axis represents the precision rate. We compared the proposed method SemiCCA utilizing both labeled and unlabeled samples with the standard CCA utilizing only labeled samples. Figure 7 indicates that latent space extraction based on SemiCCA with labeled and unlabeled images was effective against the standard CCA with only labeled images.

5.3 Automatic Audio Annotation

For experiments of automatic audio annotation, we use the data collected from a audio sharing service called Freesound [*2], which consists of various audio files annotated with word tags such as "people", "noisy", and "restaurant." The goal is to predict the existence of each tag for a new audio clip. We downloaded 2012 audio clips from among all files containing any of pre-defined 230 text labels, 3–60 seconds in length and with a sampling rate 44.1 kHz. We then randomly selected 1,000 clips as labeled training samples, 912 clips as unlabeled training samples and the rest (= 100 samples) as samples for evaluation. In the same way as image annotation, we adopted the precision rate and recall rate as the evaluation measures.

Figure 8 shows the experimental results, where the threshold vector $\boldsymbol{\theta}$ varied from $0\hat{\boldsymbol{\theta}}$ to $5.0\hat{\boldsymbol{\theta}}$. The horizontal axis stands for the recall rate, and the vertical axis represents the precision rate. We compared the proposed method SemiCCA utilizing both labeled and unlabeled samples (red line) with (1) the standard CCA utilizing only labeled samples (blue line) and (2) the standard CCA in the case that all the unlabeled samples would have been labeled (purple line). We note that the second opponent, the standard CCA in the case that all the unlabeled samples would have been labeled, would be the essential upper bound of the performance of SemiCCA. All the audio clips in the Freesound dataset used in this experiment have text labels, and a part of them were forced to be considered as unlabeled to build a semi-supervised setup. Thus, the experiments in this section can acquire the upper bound of the semi-supervised performance. This is not the same case as the PASCAL VOC dataset shown in the previous section, whose test samples did not have any ground-truth text labels. Figure 8 indicates that latent space extraction based on SemiCCA was effective also for automatic audio annotation.

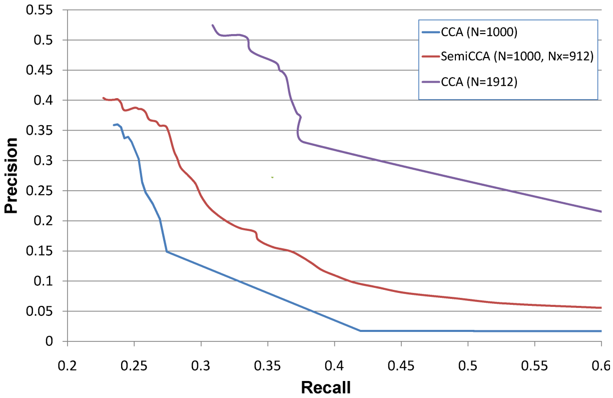

Fig. 8 Precision-recall curve for automatic audio annotation with Freesound dataset.

6. Concluding Remarks

In this paper, we proposed a semi-supervised extension of CCA that we call *SemiCCA*. Our formulation is quite simple and also intuitively understandable. Namely, SemiCCA smoothly bridges CCA with paired samples and PCA with paired and unpaired samples by a trade-off parameter. We evaluated its experimental performance, and revealed the effectiveness of SemiCCA against the original CCA.

In our future work, we will clarify some relationships between the proposed method SemiCCA and Bayesian modeling [25], [26], [27], and apply SemiCCA to other challenging real-world problems such as multi-modal event correlation analysis for audio-video synchronization, audio-visual speech recognition and sensor data mining.

References

[1] Hardoon, D.R., Szedmak, S. and Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods, *Neural Computation*, Vol.16, No.12, pp.2639–2664 (2004).

[2] Downie, J.S.: The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research, *Acoustical Science and Technology*, Vol.29, No.4, pp.247–255 (2008).

[3] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results (2012), available from ⟨http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html⟩.

[4] Smeaton, A.F., Over, P. and Kraaij, W.: Evaluation campaigns and trecvid, *Proc. ACM International Workshop on Multimedia Information Retrieval* (*MIR*), pp.321–330 (2006).

[5] Correa, N., Adali, T., Li, Y.O. and Calhoun, V.: Canonical correlation analysis for data fusion and group inferences, *IEEE Signal Processing Magazine*, Vol.27, No.4, pp.39–50 (2010).

[6] Pezeshki, A., Azimi-Sadjadi, M.R. and Scharf, L.L.: Undersea target classification using canonical correlation analysis, *IEEE Journal*

[*2] http://www.freesound.org

of Oceanic Engineering, Vol.32, No.4, pp.948–955 (2007).

[7] Nielsen, A.A.: Multiset canonical correlations analysis and multispectral, truly multi-temporal remote sensing data, *IEEE Trans. Image Processing*, Vol.11, No.3, pp.293–305 (2002).

[8] Schizas, I., Giannakis, G. and Luo, Z.Q.: Distributed estimation using reduced-dimensionality sensor observations, *IEEE Trans. Signal Processing*, Vol.55, No.8, pp.4284 –4299 (2007).

[9] Hotelling, H.: Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, Vol.24 (1933).

[10] Blaschko, M.B., Lampert, C.H. and Gretton, A.: Semi-supervised Laplacian regularization of kernel canonical correlation analysis, *Proc. European Conference on Machine Learning and Knowledge Discovery in Databases* (*ECML-PKDD*), pp.133–145 (2008).

[11] Lai, P.L. and Fyfe, C.: Kernel and nonlinear canonical correlation analysis, *Proc. International Joint Conference on Neural Networks* (*IJCNN*), p.4614 (2000).

[12] Akaho, S.: A kernel method for canonical correlation analysis, *Proc. International Meeting of the Psychometric Society* (*IMPS*), pp.1–5 (2001).

[13] Kursun, O. and Alpaydin, E.: Canonical correlation analysis for multiview semisupervised feature extraction, *Proc. International Conference on Artificial Intelligence and Soft Computing* (*ICAISC*), pp.430–436 (2010).

[14] Peng, Y. and Zhang, D.Q.: Semi-supervised canonical correlation analysis algorithm (in Chinese), *Journal of Software*, Vol.19, No.11, pp.2822–2832 (2008).

[15] Yanai, H. and Puntanen, S.: Partial canonical correlation associated with the inverse and some generalized inverse of a partitioned dispersion matrix, *Proc. Pacific Area Statistical Conference on Statistical Sciences and Data Analysis*, pp.253–264 (1993).

[16] Sugiyama, M., Idé, T., Nakajima, S. and Sese, J.: Semi-supervised local Fisher discriminant analysis for dimensionality reduction, *Machine Learning*, Vol.78, No.1, pp.35–61 (2010).

[17] Nayayama, H., Harada, T., Kuniyoshi, Y. and Otsu, N.: High-performance image annotation and retrieval for weakly labeled images, *Proc. Pacific-Rim Conference on Multimedia* (*PCM*), pp.601–610 (2008).

[18] Harada, T., Nakayama, H. and Kuniyoshi, Y.: Image annotation retrieval based on efficient learning of contextual latent space, *Proc. IEEE International Conference on Multimedia and Expo* (*ICME*), pp.858–861 (2009).

[19] Hofmann, T.: Probabilistic latent semantic indexing, *Proc. International ACM SIGIR Conference on Research and development in Information Retrieval* (*SIGIR*), pp.50–57 (1999).

[20] Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis, *Machine Learning*, Vol.42, No.1, pp.177–196 (2001).

[21] Bay, H., Ess, A., Tuytelaars, T. and Van, Gool, L.: Speeded-up robust features (SURF), *Computer Vision and Image Understanding* (*CVIU*), Vol.110, No.3, pp.346–359 (2008).

[22] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A.: The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results (2008), available from ⟨http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html⟩.

[23] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A.: The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results (2009), available from ⟨http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html⟩.

[24] Carneiro, G., Chan, A.B., Moreno, P.J. and Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval, *IEEE Trans. Pattern Analysis and Machine Intelligence* (*PAMI*), Vol.29, No.3, pp.394–410 (2007).

[25] Bach, F.R. and Jordan, M.I.: A probabilistic interpretation of canonical correlation analysis, Tech. Rep. 688, Department of Statistics, University of California, Berkeley (2005).

[26] Wang, C.: Variational bayesian approach to canonical correlation analysis, *IEEE Trans. Neural Networks*, Vol.18, No.3, pp.905–910 (2007).

[27] Virtanen, S., Klami, A. and Kaski, S.: Bayesian CCA via group sparsity, *Proc. IEEE International Conference on Machine Learning* (*ICML*), pp.457–464 (2011).

Akisato Kimura received his B.E., M.E. and D.E. degrees in communications and integrated systems from Tokyo Institute of Technology, Japan in 1998, 2000 and 2007, respectively. Since 2000, he has been with NTT Communication Science Laboratories, NTT Corporation, where he is currently a senior research scientist in Innovative Communication Laboratory. He has been engaged in work on multimedia content identification, automatic multimedia annotation, human visual attention modeling and social media mining. He is a member of IEICE, IEEE and ACM SIGMM/SIGKDD.

Masashi Sugiyama received his B.E., M.E., and Ph.D. degrees from Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan, in 1997, 1999, and 2001, respectively. In 2001, he was appointed as a Research Associate in the same institute, and from 2003, he is an Associate Professor. His research interests include theory and application of machine learning.

Takuho Nakano received his B.E. and M.E. degrees from the University of Tokyo, Japan, in 2009 and 2011, respectively. His research interests include statistical signal processing, music processing, and machine learning. He is a member of IEICE and IPSJ.

Hirokazu Kameoka received his B.E., M.E. and Ph.D. degrees all from the University of Tokyo, Japan, in 2002, 2004 and 2007, respectively. He is currently a research scientist at NTT Communication Science Laboratories and an Adjunct Associate Professor at the University of Tokyo. His research interests include computational auditory scene analysis, statistical signal processing, speech and music processing, and machine learning. He is a member of IEEE, IEICE, IPSJ and ASJ. He received 13 awards over the past 9 years, including the IEEE Signal Processing Society 2008 SPS Young Author Best Paper Award.

Hitoshi Sakano received his B.S. degree in physics from Chuo University Tokyo and M.S. degree in physics from Saitama University, and Ph.D. degree in applied physics from Waseda University, Tokyo in 1988, 1990 and 2008, respectively. He joined NTT Communication Science Laboratories in 2008 and studied pattern recognition technology. He is a member of IEEE, IEICE and Physical Society of Japan (PSJ).

Eisaku Maeda received his B.S. and M.S. degrees in biological science and Ph.D. degree in mathematical engineering, from the University of Tokyo, Tokyo, Japan, in 1984, 1986 and 1993, respectively. He joined NTT Corporation in 1986, was a visiting scholar of Physiological Laboratory, the University of Cambridge, UK from 1995 to 1996, and is currently an Executive Research Scientist in Research Planning Section, Communication Science Laboratories, NTT Corporation. His research interests are in pattern recognition, statistical learning and bioinformatics. He is a member of IEEE, JSBI and IPSJ.

Katsuhiko Ishiguro has been a researcher at NTT Communication Science Laboratories, NTT Corporation, Kyoto since 2006. He received his B.E. and M.Info. degrees from the University of Tokyo, Japan, in 2000 and 2004, respectively, and Ph.D. degree from University of Tsukuba, Ibaraki, Japan in 2010. His research interests include multimedia data modeling with Bayesian approaches, probabilistic models for data mining, and time series analysis. He is a member of IEEE, IEICE and IPSJ.