

SemiCCA: Efficient semi-supervised learning of canonical correlations

Akisato Kimura*, Hirokazu Kameoka*, Masashi Sugiyama†, Takuho Nakano*,
Eisaku Maeda*, Hitoshi Sakano* and Katsuhiko Ishiguro*

* NTT Communication Science Laboratories, NTT Corporation
Keihanna Science City, Kyoto, Japan. E-mail: akisato@ieee.org

† Graduate School of Information Science and Engineering, Tokyo Institute of Technology
Meguro, Tokyo, Japan. E-mail: sugi@cs.titech.ac.jp

‡ Graduate School of Information Science and Technologies, the University of Tokyo
Bunkyo, Tokyo, Japan. E-mail: t-nakano@hil.t.u-tokyo.ac.jp

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. The proposed method smoothly bridges the 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.*

Index Terms—*Canonical correlation analysis, semi-supervised learning, generalized eigenproblem, automatic image annotation*

I. INTRODUCTION

Analyzing high-dimensional co-occurring data (\mathbf{x}, \mathbf{y}) is an important challenge in machine learning and pattern recognition communities, e.g., in the context of audio [1] and image annotation¹ [2]. *Canonical correlation analysis* (CCA) [3] is a classical but still powerful method for analyzing multivariate paired samples. CCA finds projection directions \mathbf{w}_x and \mathbf{w}_y so that correlation between projected samples $\mathbf{w}_x^\top \mathbf{x}$ and $\mathbf{w}_y^\top \mathbf{y}$ is maximized.

However, the performance of CCA tends to be degraded when the number of paired samples (\mathbf{x}, \mathbf{y}) is limited. On the other hand, a large number of additional *unpaired* samples (i.e., \mathbf{x} -only samples and \mathbf{y} -only samples) are often available in real-world applications. To utilize such additional unpaired samples, several *semi-supervised* [4] extensions of CCA have been proposed, e.g., based on Tikhonov regularization [5] and graph-Laplacian regularization [6].

In this paper, we propose a yet another semi-supervised variant of CCA called *semiCCA*. SemiCCA utilizes additional unpaired samples by smoothly bridging CCA and *principal component analysis* (PCA). More specifically, the 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.

¹In such cases, \mathbf{x} corresponds to an audio/image feature, and \mathbf{y} corresponds to a feature derived from the annotated information.

II. CANONICAL CORRELATION ANALYSIS (CCA)

Consider a set of paired samples of size N , $\mathbf{X}^{(L)} = \{\mathbf{x}_n\}_{n=1}^N$ and $\mathbf{Y}^{(L)} = \{\mathbf{y}_n\}_{n=1}^N$. Without loss of generality, we assume that $\mathbf{X}^{(L)}$ and $\mathbf{Y}^{(L)}$ are both centered, which can always be achieved by subtracting the sample means from each sample. CCA is a method of finding bases \mathbf{w}_x and \mathbf{w}_y for $\mathbf{X}^{(L)}$ and $\mathbf{Y}^{(L)}$ such that their correlation is maximized as

$$\max_{(\mathbf{w}_x, \mathbf{w}_y)} \frac{\mathbf{w}_x^\top \mathbf{S}_{xy}^{(L)} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^\top \mathbf{S}_{xx}^{(L)} \mathbf{w}_x} \sqrt{\mathbf{w}_y^\top \mathbf{S}_{yy}^{(L)} \mathbf{w}_y}}, \quad (1)$$

where $\mathbf{S}_{xx}^{(L)} = 1/N \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top$, and $\mathbf{S}_{yy}^{(L)}$ and $\mathbf{S}_{xy}^{(L)}$ are defined similarly. The solution $(\mathbf{w}_x, \mathbf{w}_y)$ is given as the solution of the following generalized eigenvalue problem:

$$\begin{pmatrix} \mathbf{0} & \mathbf{S}_{xy}^{(L)} \\ \mathbf{S}_{yx}^{(L)} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{S}_{xx}^{(L)} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{yy}^{(L)} \end{pmatrix} \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix}. \quad (2)$$

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

III. SEMICCA

A. Semi-supervised Setup

When the number of paired samples is small, CCA tends to overfit the given samples. Here, let us consider the situation where *unpaired* samples $\mathbf{X}^{(U)} = \{\mathbf{x}_n\}_{n=N+1}^{N_x}$ and $\mathbf{Y}^{(U)} = \{\mathbf{y}_n\}_{n=N+1}^{N_y}$ are additionally provided², where $\mathbf{X}^{(U)}$ and $\mathbf{Y}^{(U)}$ are independently generated. Since the original CCA cannot directly incorporate such unpaired samples, we propose a novel method named *SemiCCA* that can avoid overfitting by utilizing the additional unpaired samples.

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 $(\mathbf{X}^{(L)}, \mathbf{Y}^{(L)})$ are

²In the context of image annotation retrieval, unpaired samples $\mathbf{X}^{(U)}$ only exist, whereas $\mathbf{Y}^{(U)}$ is empty.

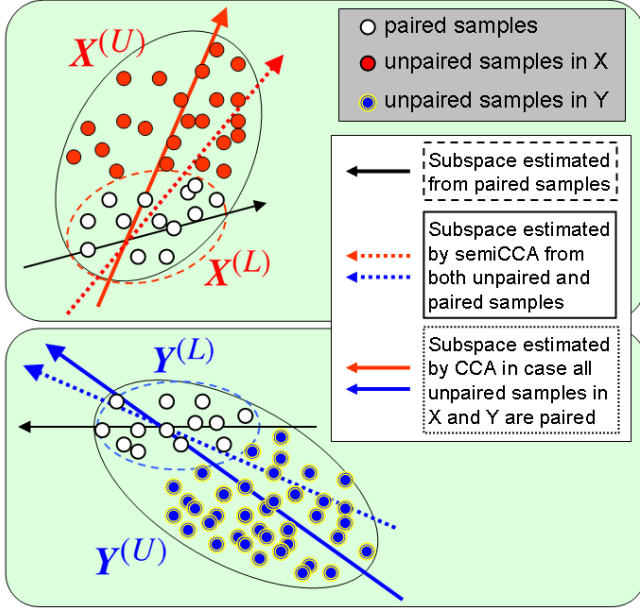


Fig. 1. Effects of unpaired samples in SemiCCA

used, poor projection directions may be obtained by CCA due to overfitting. In contrast, unpaired samples may be used for revealing the global structure in each domain. Note once a basis in one sample space is rectified, the corresponding bases in the other sample space is also rectified so that correlations between two bases are maximized.

B. Definition

Motivated by the above illustration, we propose to combine CCA with principal component analysis (PCA) for utilizing unpaired samples. There are various possibilities to combine CCA and PCA. Here we combine the eigenvalue problems of CCA and PCA since this allows us to compute the combined solution efficiently. More specifically, the solution of SemiCCA is given by the leading generalized eigenvectors of the following generalized eigenvalue problem:

$$\mathbf{B} \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix} = \lambda \mathbf{C} \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix}, \quad (3)$$

where

$$\mathbf{B} = \beta \begin{pmatrix} \mathbf{0} & \mathbf{S}_{xy}^{(L)} \\ \mathbf{S}_{yx}^{(L)} & \mathbf{0} \end{pmatrix} + (1 - \beta) \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{yy} \end{pmatrix},$$

$$\mathbf{C} = \beta \begin{pmatrix} \mathbf{S}_{xx}^{(L)} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{yy}^{(L)} \end{pmatrix} + (1 - \beta) \begin{pmatrix} \mathbf{I}_{D_x} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{D_y} \end{pmatrix},$$

$$\mathbf{S}_{xx} = 1/N_x \sum_{n=1}^{N_x} \mathbf{x}_n \mathbf{x}_n^\top,$$

$$\mathbf{S}_{yy} = 1/N_y \sum_{n=1}^{N_y} \mathbf{y}_n \mathbf{y}_n^\top,$$

and β is a constant named a *trade-off parameter* taking a value in $[0, 1]$. This parameter controls the trade-off between CCA and PCA. Namely, when $\beta = 1$, Eq. (3) is reduced to the CCA eigenvalue problem Eq. (2) while when $\beta = 0$ Eq. (3)

is reduced to the PCA eigenvalue problem, under the assumption that $\mathbf{X} = (\mathbf{X}^{(L)}, \mathbf{X}^{(U)})$ and $\mathbf{Y} = (\mathbf{Y}^{(L)}, \mathbf{Y}^{(U)})$ 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.

Motivated by the above illustration, we propose to combine CCA with PCA for utilizing unpaired samples. There are various possibilities to combine CCA and PCA. Here we combine the eigenvalue problems of CCA and PCA since this allows us to compute the combined solution efficiently³. More specifically, the solution of semiCCA is given by the leading generalized eigenvectors of the following generalized eigenvalue problem:

$$\mathbf{B} \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix} = \lambda \mathbf{C} \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix}, \quad (4)$$

where

$$\mathbf{B} = \beta \begin{pmatrix} \mathbf{0} & \hat{\mathbf{S}}_{xy} \\ \hat{\mathbf{S}}_{xy}^\top & \mathbf{0} \end{pmatrix} + (1 - \beta) \begin{pmatrix} \hat{\mathbf{S}}'_{xx} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{S}}'_{yy} \end{pmatrix},$$

$$\mathbf{C} = \beta \begin{pmatrix} \hat{\mathbf{S}}_{xx} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{S}}_{yy} \end{pmatrix} + (1 - \beta) \begin{pmatrix} \mathbf{I}_{d_x} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d_y} \end{pmatrix},$$

$$\hat{\mathbf{S}}'_{xx} = \frac{1}{N + N_x} \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top + \sum_{i=1}^{N_x} \mathbf{x}'_i \mathbf{x}'_i{}^\top \right),$$

$$\hat{\mathbf{S}}'_{yy} = \frac{1}{N + N_y} \left(\sum_{i=1}^N \mathbf{y}_i \mathbf{y}_i^\top + \sum_{i=1}^{N_y} \mathbf{y}'_i \mathbf{y}'_i{}^\top \right),$$

\mathbf{I}_d is the $d \times d$ identity matrix, and β is a constant named a *trade-off parameter* taking a value in $[0, 1]$.

Here, we show some operational interpretations of the constant β , which is the only free parameter of semiCCA. Roughly speaking, it controls the trade-off between CCA and PCA. Namely, when $\beta = 1$, Eq.(4) is reduced to the CCA eigenvalue problem Eq.(2). On the other hand, when $\beta = 0$, Eq.(4) is reduced to the PCA eigenvalue problem, under the assumption that $\mathbf{X} = (\mathbf{X}^{(L)}, \mathbf{X}^{(U)})$ and $\mathbf{Y} = (\mathbf{Y}^{(L)}, \mathbf{Y}^{(U)})$ 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. This intuition can be supported by some relationships between the operation of semiCCA and a generative model that combines a probabilistic latent semantic analysis (pLSA) model [8] (i.e., a generative model that the original CCA assumes [9]) and two latent models (i.e., a generative models the PCA assumes [10]) We omit the details for this issue due to the limited space.

One may use different trade-off parameters in \mathbf{B} and \mathbf{C} to increase the flexibility. However, this in turn makes the trade-off parameter choice laborious. For this reason, we focus on

³This idea is motivated by [7], which combines a variant of Fisher discriminant analysis with PCA by blending the eigenvalue problems.

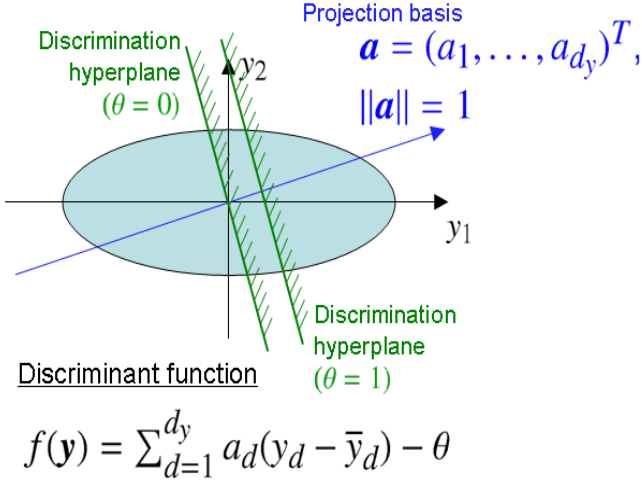


Fig. 2. Artificial data.

using the single shared trade-off parameter β for both \mathbf{B} and \mathbf{C} .

We focused on the case where two sets of samples are given, but semiCCA can be easily extended to multiple data sets by considering correlations over all pairs of samples [11]. It is also easy to show that semiCCA can be *kernelized* by applying the *kernel trick* (see e.g. [6]).

IV. EXPERIMENTS

In this section, we report the results of numerical experiments.

We evaluated the performance of the proposed method using the artificial data set created as follows: We considered a Gaussian pLSA 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 denoted by X and Y . We drew samples $\{z_i\}_{i=1}^{N_z}$ from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{d_z})$, where $d_z = 10$ is the dimension of the random variable Z . The number N_z of samples was set to $N_z = 10000$. The means and covariance matrices of the conditional (Gaussian) densities $p(X|Z)$ and $p(Y|Z)$ were determined randomly. More specifically, we randomly generated each component of transformation matrices \mathbf{T}_x and \mathbf{T}_y and means $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ following $\mathcal{N}(0, 1)$. Then complete paired samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_z}$ were created as

$$\begin{aligned} \mathbf{x}_i &= \mathbf{T}_x \mathbf{z}_i + \bar{\mathbf{x}} + \boldsymbol{\delta}_{x,i}, & \boldsymbol{\delta}_{x,i} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{X|Z}), \\ \mathbf{y}_i &= \mathbf{T}_y \mathbf{z}_i + \bar{\mathbf{y}} + \boldsymbol{\delta}_{y,i}, & \boldsymbol{\delta}_{y,i} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{Y|Z}), \end{aligned}$$

where each component of $\boldsymbol{\Sigma}_{X|Z}$ and $\boldsymbol{\Sigma}_{Y|Z}$ was generated from the folded standard normal distribution. The dimensions of the samples are set to $d_x = 15$ and $d_y = 20$.

Then, we removed several samples from $\{\mathbf{y}_i\}_{i=1}^{N_z}$ as depicted in Figure 2. Here, we used a linear discriminant function $f(\cdot)$

$$f(\mathbf{y}) = \sum_{d=1}^{d_y} a_d(y_d - \bar{y}_d) - \theta, \quad (5)$$

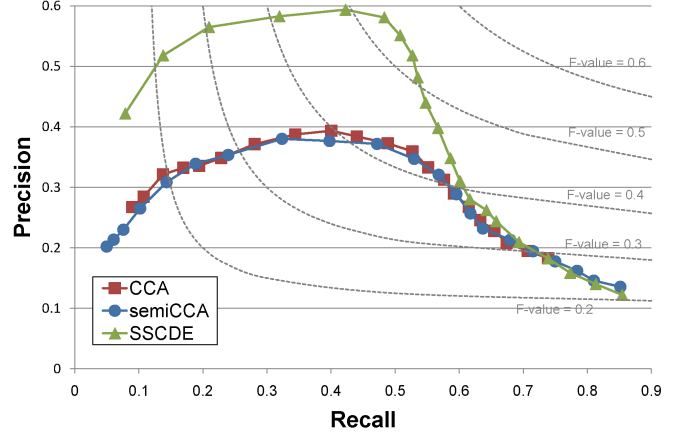


Fig. 3. Average evaluation score for automatic image annotations.

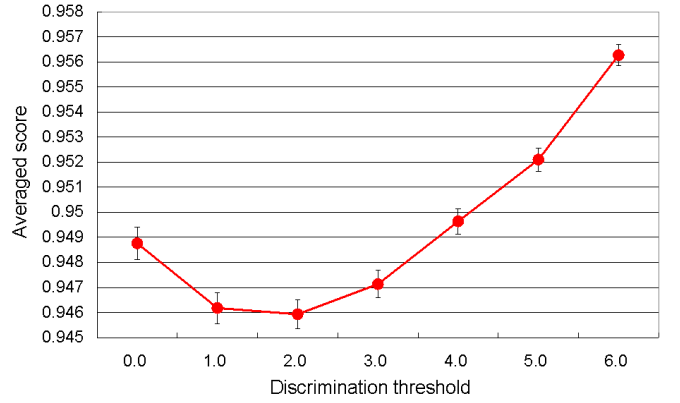


Fig. 4. Average trade-off parameter taking the highest score.

where $\mathbf{a} = (a_1, \dots, a_{d_y})^\top$ is a coefficient vector satisfying $\|\mathbf{a}\| = 1$, and θ is a threshold value such that the larger θ is, the more samples are removed. A sample $(\mathbf{x}_i, \mathbf{y}_i)$ was kept paired if $f(\mathbf{y}_i) > 0$, and \mathbf{y}_i was removed otherwise.

We compare the proposed semiCCA with the original CCA. We evaluated the performance of (semi)CCA by the weighted sum of cosine distances defined as follows:

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

where $\mathbf{w}_{x,i}^*$ and λ_i^* are the “true” eigenvectors and eigenvalues. 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.

Figure 3 shows the evaluation scores averaged over 10000 independent trials for several discrimination thresholds θ and also shows the average number of paired samples for each discrimination threshold. The results indicate that semiCCA tends to outperform the ordinary CCA; it is note worthy that even when the number of unpaired samples is not so large, semiCCA performs better than the original CCA.

Figure 4 shows the trade-off parameter taking the highest score averaged over all the trials, and Figure 5 depicts the histogram of the best trade-off parameters. The results imply

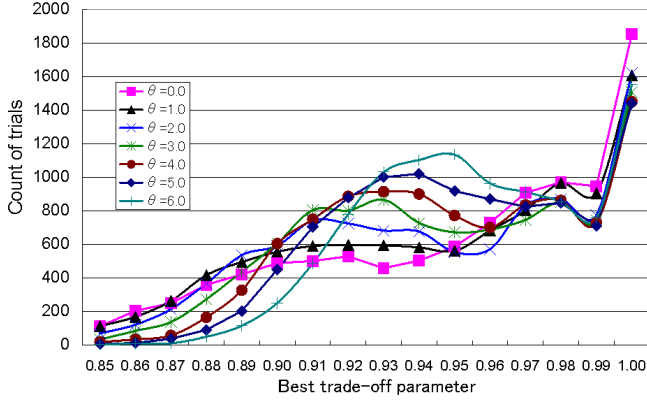


Fig. 5. Histogram of trade-off parameters taking the highest score.

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. The results also indicate that the best trade-off parameters were usually close to 1, i.e., the effect of PCA is only mildly incorporated. Nevertheless, the performance is much improved, as shown in Figure 3.

V. APPLICATIONS TO AUTOMATIC IMAGE ANNOTATION

A. Framework

Fig. V-A shows the framework of image annotation retrieval with semiCCA.

First of all, feature vectors are extracted from images $\mathbf{G} = \{\mathbf{g}_n\}_{n=1}^{N_x}$ and associated text labels $\mathbf{W} = \{\mathbf{w}_n\}_{n=1}^N$, where N is the number of labeled images and N_x is the total number of images including labeled and unlabeled images (should be $N \leq N_x$ and in most cases $N \ll N_x$). Each text label \mathbf{w}_n is composed of text words selected from a word set given in advance. We utilize Bag of Features (BoF) as image features $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N_x}$, where SURF [12] is used for key-point detection and descriptor extraction, and word existence vectors $\mathbf{Y} = \{\mathbf{y}_n\}_{n=1}^N$ as text features, where each dimension represents an existence or absence of a specific word.

Next, a topic model is estimated from feature vectors (\mathbf{X}, \mathbf{Y}) by the proposed method called SSCDE, which consists of two steps: latent variable generation with our new method named *SemiCCA* [13], and model density estimation with multi-class SSKDE [14].

The first step is to generate latent variables $\mathbf{Z} = \{\mathbf{z}_n\}_{n=1}^{N_x}$ with SemiCCA. More specifically, a pair (f_x, f_{xy}) of functions $f_x: \mathcal{R}^{D_x} \rightarrow \mathcal{R}^{D_z}$ and $f_{xy}: \mathcal{R}^{D_x} \times \mathcal{R}^{D_y} \rightarrow \mathcal{R}^{D_z}$ is derived from (\mathbf{X}, \mathbf{Y}) as training samples with SemiCCA, and latent variables \mathbf{Z} are generated from (\mathbf{X}, \mathbf{Y}) with (f_x, f_{xy}) , where D_x, D_y and D_z is the dimension of vectors $\mathbf{x}_n, \mathbf{y}_n$ and \mathbf{z}_n , respectively. We will describe the detail of the first step in Section ??.

The second step is to set up a topic model by utilizing a multi-class SSKDE. The topic model is described by the

following equation:

$$p(\mathbf{x}, \mathbf{y}) = 1/N_x \sum_{n=1}^{N_x} p(\mathbf{x}|\mathbf{z}_n)p(\mathbf{y}|\mathbf{z}_n). \quad (6)$$

The main procedure in this step is to construct the conditional densities $p(\mathbf{x}|\mathbf{z}_n)$ and $p(\mathbf{y}|\mathbf{z}_n)$ of features (\mathbf{x}, \mathbf{y}) for every latent variable \mathbf{z}_n . Note that in Fig. V-A a conditional density $p(\mathbf{y}|\mathbf{z})$ can be derived even though the corresponding text label does not exist. We will show the detailed procedure in Section ??.

Once the model estimation has been finished, we can execute image annotation and retrieval within the same framework through maximum a posteriori (MAP) estimation.

[Annotation]

By using an image feature $\mathbf{x}^{(g)}$ extracted from a given image $\mathbf{g}^{(g)}$, the text feature $\hat{\mathbf{y}}$ of the most probable text label $\hat{\mathbf{w}}$ can be derived as

$$\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y} \in [0,1]^{D_y}} p(\mathbf{y}|\mathbf{x}^{(g)}) \quad (7)$$

$$= \operatorname{argmax}_{\mathbf{y} \in [0,1]^{D_y}} \frac{\sum_{n=1}^{N_x} p(\mathbf{x}^{(g)}|\mathbf{z}_n)p(\mathbf{y}|\mathbf{z}_n)}{\sum_{n=1}^{N_x} p(\mathbf{x}^{(g)}|\mathbf{z}_n)}. \quad (8)$$

As shown later in Section ??, a conditional density $p(\mathbf{y}|\mathbf{z}_n)$ for a text feature $\mathbf{y} = (y_1, \dots, y_{D_y})^\top$ ($y_d \in \{0,1\}$, $^\top$ denotes the transpose of a vector of a matrix) is modeled as $p(\mathbf{y}|\mathbf{z}_n) = \prod_{d=1}^{D_y} p(y_d|\mathbf{z}_n)$. Therefore, the annotation problem can be rewritten to the following simple form:

$$\hat{y}_d = \frac{\sum_{n=1}^{N_x} p(\mathbf{x}^{(g)}|\mathbf{z}_n)p(y_d = 1|\mathbf{z}_n)}{\sum_{n=1}^{N_x} p(\mathbf{x}^{(g)}|\mathbf{z}_n)}. \quad (9)$$

When \hat{y}_d exceeds a pre-defined threshold θ_d , the text word of index d is provided to the given image $\mathbf{g}^{(g)}$.

[Retrieval]

By using a text feature $\mathbf{y}^{(g)}$ extracted from a given text label $\mathbf{w}^{(g)}$, the image feature $\hat{\mathbf{x}}$ of the most probable image $\hat{\mathbf{g}}$ in the database can be derived as

$$\hat{\mathbf{x}} = \operatorname{argmax}_{\mathbf{x} \in DB} p(\mathbf{x}|\mathbf{y}^{(g)}) \quad (10)$$

$$= \operatorname{argmax}_{\mathbf{x} \in DB} \sum_{n=1}^{N_x} p(\mathbf{y}^{(g)}|\mathbf{z}_n)p(\mathbf{x}|\mathbf{z}_n) \quad (11)$$

where DB is a set of registered images in a given database.

B. Experiments

This section describes the results for the automatic image annotation task with the dataset used in PASCAL Visual Object Challenge (VOC) 2008 [15] and 2009 [2]. The VOC dataset is composed of images including objects from 20 visual object classes related to people, animals, vehicles and furniture. Multiple objects from multiple classes may be present in a single image. Example images included in VOC2008 training dataset is shown in Fig. 7. As shown in this figure, each training image has a bounding box and object class label for each object presented in the image, we *removed all the bounding boxes* and only utilized class labels associated with bounding boxes to simulate “weak labeling” settings [16],

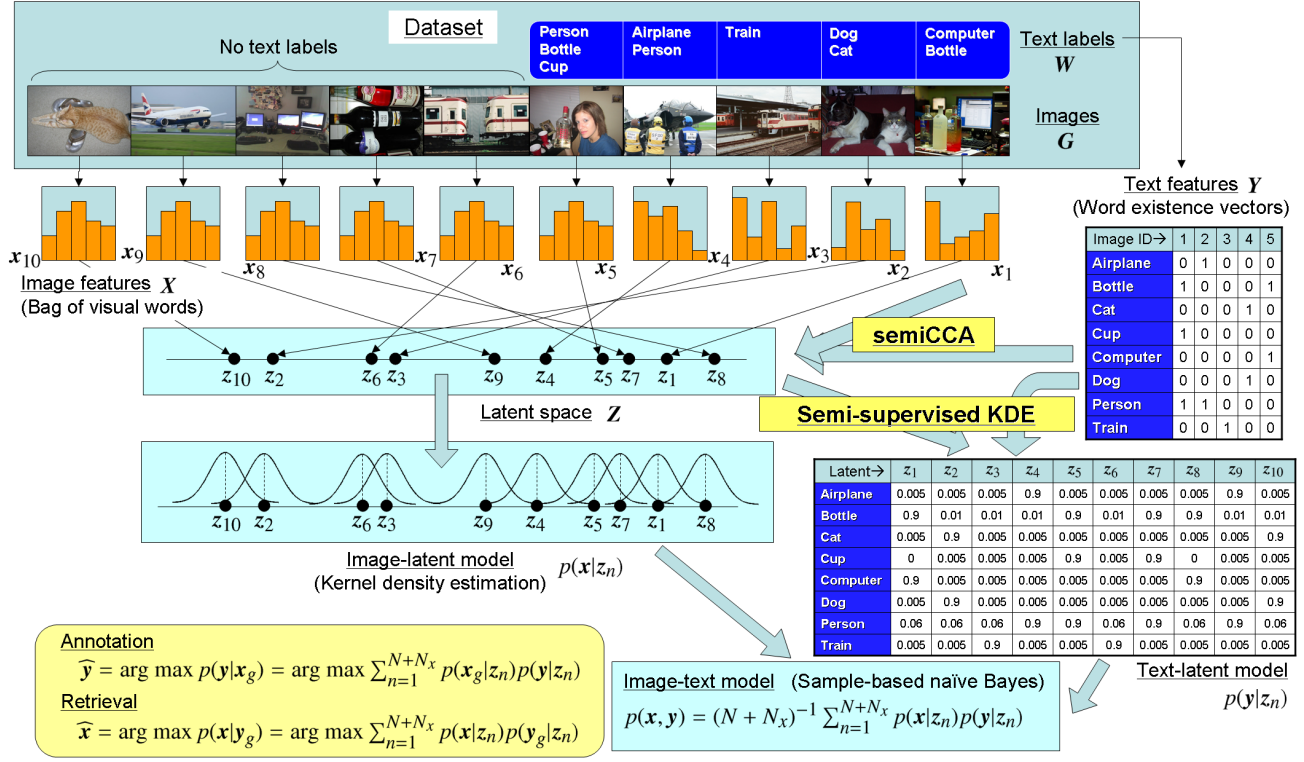


Fig. 6. Framework of the proposed method for image annotation retrieval

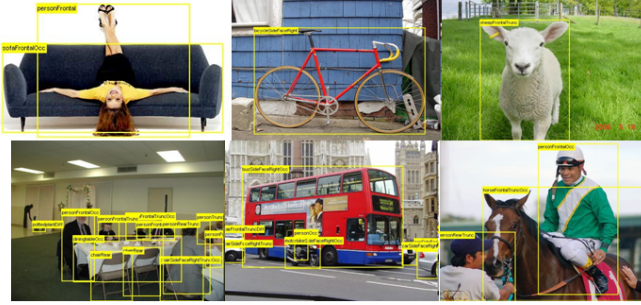


Fig. 7. Example images in VOC2008 dataset

where images are weakly related to multiple words without region information.

We utilized all of the 5096 images in VOC2008 training dataset, and separated them into 1000 labeled images for training ($G^{(L)}$), 500 unlabeled images for evaluation ($G^{(E)}$) and the rest (3596 images) as unlabeled images for training and also images for determining hyper parameters D_z , β , t and μ with 7-fold cross validation. Also, 2722 images in VOC2009 training dataset⁴ were added to unlabeled images for training. In total, 6318 unlabeled images for training ($G^{(U)}$) were utilized. We adopted the precision rate PR , recall rate RE

and F-value F as the evaluation measures, defined as

$$PR = \frac{\sum_{n=1}^{N_e} TP_n}{\sum_{n=1}^{N_e} (TP_n + FP_n)}, \quad (12)$$

$$RE = \frac{\sum_{n=1}^{N_e} TP_n}{\sum_{n=1}^{N_e} (TP_n + FN_n)}, \quad (13)$$

$$F = 2(1/PR + 1/RE)^{-1}, \quad (14)$$

where N_e ($= 500$) is the number of images in $G^{(E)}$, TP_n , FP_n and FN_n is respectively true positives, false positives and false negatives for the n -th image in $G^{(E)}$. The hyper parameter γ was determined by an adaptive variant of SSKDE called SSKADE [14].

Fig. V-B shows the experimental results for the automatic annotation task, where the threshold $\theta = (\theta_1, \dots, \theta_{D_y})^T$ varies from 0.5θ to 2.0θ . The horizontal axis stands for the recall rate, the vertical axis represents the precision rate, and each gray dash line is a precision-recall curve with the constant F-value ($F = 0.2$ to 0.6 from the bottom left to the top right). We compared the proposed method (green line, triangular markers) with the method [17] using only labeled images $G^{(L)}$ and CCA-based model learning (red line, square markers), and the method using labeled $G^{(L)}$ and unlabeled images $G^{(U)}$ and SemiCCA-based model learning, namely multi-class SSKDE was removed from the proposed method (blue line, circle markers). Since SemiCCA includes CCA as a special case of $\beta = 1.0$, annotation with topic models learned by SemiCCA would achieve at least as the same accuracy as

⁴VOC2009 training dataset contains all the images in VOC2008 training dataset, however, those duplicated images were removed in advance.

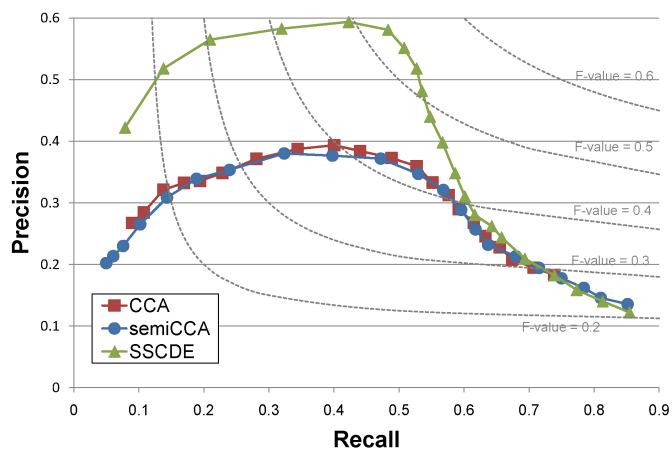


Fig. 8. Experiment results for automatic image annotation

the one by CCA. From this viewpoint, Fig. V-B implies that latent space extraction based on SemiCCA was not effective so much against the dataset used in the experiment. However, the proposed method greatly improved the annotation accuracy by additionally introducing multi-class SSKDE.

VI. 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 compare *semiCCA* with other semi-supervised variants of CCA such as [5], [6], and apply *semiCCA* to challenging real-world problems such as multi-label classification for automatic image/music annotation and retrieval, and multi-model event correlation analysis for audio-video synchronization and audio-visual speech recognition.

ACKNOWLEDGMENTS

The authors thank Dr. Mark Everingham of University of Leeds for providing all the test images of PASCAL VOC2009 datasets. The fourth author contributed to this work during the internship in NTT Communication Science Laboratories. The authors also thank Dr. Yoshinobu Tonomura (currently Ryukoku University), Dr. Naonori Ueda, Dr. Futoshi Naya, Dr. Kenji Nakazawa (currently with NTT Advance Technologies Inc.), Dr. Junji Yamato and Dr. Kunio Kashino of NTT Communication Science Laboratories, Prof. Nobutaka Ono and Prof. Shigeki Sagayama of the University of Tokyo for their help to the internship.

REFERENCES

[1] J. Stephen Downie, "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.

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

[3] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psych.*, vol. 24, 1933.

[4] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*, MIT Press, Cambridge, 2006.

[5] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[6] M. B. Blaschko, C. H. Lampert, and A. Gretton, "Semi-supervised laplacian regularization of kernel canonical correlation analysis," in *ECML PKDD '08: Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I*, Berlin, Heidelberg, 2008, pp. 133–145, Springer-Verlag.

[7] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese, "Semi-supervised local Fisher discriminant analysis for dimensionality reduction," *Machine Learning*, vol. 78, no. 1–2, pp. 35–61, 2010.

[8] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 41, no. 2, pp. 177–196, 2001.

[9] Francis R. Bach and Michael I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Tech. Rep. 688, Department of Statistics, University of California, Berkeley, 2005.

[10] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysers," *Journal of the Royal Statistical Society B*, vol. 61, no. 3, pp. 611–622, 1999.

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

[12] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[13] Akisato Kimura, Hirokazu Kameoka, Masashi Sugiyama, Eisaku Maeda, Hitoshi Sakano, and Katsuhiko Ishiguro, "Semicca: Efficient semi-supervised learning of canonical correlations," in *Proc. International Conference on Pattern Recognition (ICPR)*, 2010, submitted.

[14] Meng Wang, Xian-Sheng Hua, Tao Mei, Richang Hong, Guojun Qi, Yan Song, and Li-Rong Dai, "Semi-supervised kernel density estimation for video annotation," *Computer Vision and Image Understanding*, vol. 113, no. 3, pp. 384–396, 2009.

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

[16] Gustavo Carneiro, Antoni B. Chan, Pedro J. Moreno, and Nuno Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.

[17] Tatsuya Harada, Hideki Nakayama, and Yasuo Kuniyoshi, "Image annotation retrieval based on efficient learning of contextual latent space," in *Proc. International Conference on Multimedia and Expo (ICME)*, 2009.