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, Keihanna Science City, Kyoto, Japan. E-mail: akisato@ieee.org
† Graduate School of Information Science and Engineering, Tokyo Institute of Technology E-mail: sugi@cs.titech.ac.jp
‡ Graduate School of Information Science and Technology, the University of Tokyo 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 overfittng. 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 (generalized) eigenvalue problem as the original CCA. Preliminary experiments with artificially generated samples and PASCAL VOC data sets demonstrate the effectiveness of the proposed method.**

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

I. Introduction

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 automatic audio tagging [1] and image annotation retrieval[1] [2]. *Canonical correlation analysis* (CCA) [3] is a classical but still powerful method for analyzing multivariate paired samples. CCA finds projection directions $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ so 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, where we often encounter in various real-world applications. Even in such cases, a large number of additional *unpaired* samples (i.e., $\boldsymbol{x}$-only samples and $\boldsymbol{y}$-only samples) are often available. To utilize such additional unpaired samples, several *semi-supervised* [4] extensions of (mainly kernelized) 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. SemiCCA is a generalized (and intuitively comprehensible) variant of semi-supervised CCA with Tikhonov regularization.

[1] In such cases, $\boldsymbol{x}$ corresponds to an audio/image feature, and $\boldsymbol{y}$ corresponds to a feature derived from the associated text information.

II. Canonical Correlation Analysis (CCA)

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

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

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

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

Taking the top $d_z \leq \min(d_x, d_y)$ generalized eigenvectors as row vectors, we obtain d_z-dimensional mappings $(\boldsymbol{W}_x, \boldsymbol{W}_y)$.

III. The proposed method: SemiCCA

A. Concept

When the number of paired samples is small, CCA tends to overfit the given paired samples. Here, let us consider the situation where *unpaired* samples $\boldsymbol{X}^{(U)} = \{\boldsymbol{x}_n\}_{n=N+1}^{N_x}$ and $\boldsymbol{Y}^{(U)} = \{\boldsymbol{y}_n\}_{n=N+1}^{N_y}$ are additionally provided[2], where $\boldsymbol{X}^{(U)}$ and $\boldsymbol{Y}^{(U)}$ might be 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 $(\boldsymbol{X}^{(L)}, \boldsymbol{Y}^{(L)})$ are

[2] In the context of automatic image annotation, $\boldsymbol{X}^{(U)}$ only exists, whereas $\boldsymbol{Y}^{(U)}$ is empty. However, the proposed method SemiCCA can be plausible even in the case of the presence of $\boldsymbol{X}^{(U)}$ and/or $\boldsymbol{Y}^{(U)}$

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.719

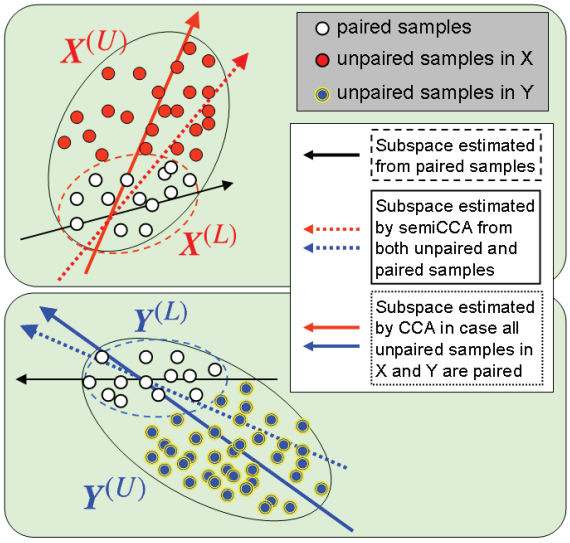

Fig. 1. Effects of unpaired samples in SemiCCA

used, poor projection directions may be obtained by CCA due to overfittng. In contrast, unpaired samples reveal the global structure of whole the samples 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[3]. More specifically, the solution of SemiCCA is given by the leading generalized eigenvectors of the following generalized eigenvalue problem:

$$\boldsymbol{B}\begin{pmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\end{pmatrix} = \lambda \boldsymbol{C}\begin{pmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\end{pmatrix}, \tag{3}$$

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

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

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

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

$\boldsymbol{I}_d$ is the $d \times d$ identity matrix, and β_B and β_C are constants named *trade-off parameters* taking a value in $[0, 1]$. From now on, we focus on using the single shared trade-off parameter β instead of the individual parameters β_B and β_C since the individual parameters makes the parameter choice laborious.

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

The trade-off parameters control 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 $\boldsymbol{X} = (\boldsymbol{X}^{(L)}, \boldsymbol{X}^{(U)})$ and $\boldsymbol{Y} = (\boldsymbol{Y}^{(L)}, \boldsymbol{Y}^{(U)})$ 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.

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 [6], [8]. Also, the proposed method can be easily extended to non-linear or non-vectorial domains by introducing the *kernel trick* [6]. From the formulation of kernelized SemiCCA, we can prove that semi-supervised CCA with Tikhonov regularization presented in [6] is a special case of SemiCCA, where some elements of kernel matrices are forced to be 0. We omit the details for this issue due to the limited space.

IV. Experiment with Artificial Data

We first 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 observations are denoted by X and Y. We drew samples $\{\boldsymbol{z}_i\}_{i=1}^{N_z}$ from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_z})$ of dimension $d_z = 10$ and number of samples $N_z = 10000$, where $\mathcal{N}(\overline{\boldsymbol{x}}, \Sigma_X)$ is a Gaussian probability density function (PDF) with mean $\boldsymbol{x}$ and covariance matrix Σ_X. The means and covariance matrices of the conditional (Gaussian) densities $p(X|Z)$ and $p(Y|Z)$ were determined randomly for each trial. More specifically, we randomly generated each component of transformation matrices $\boldsymbol{T}_x$ and $\boldsymbol{T}_y$ and means $\overline{\boldsymbol{x}}$ and $\overline{\boldsymbol{y}}$ following $\mathcal{N}(0, 1)$. Then complete paired samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_z}$ were created as

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

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 $d_y = 20$.

We removed several samples from $\{\boldsymbol{y}_i\}_{i=1}^{N_z}$ by a simple linear discrimination. As a discriminant function, we used $f(\boldsymbol{y}) = \sum_{d=1}^{d_y} a_d(y_d - \overline{y}_d) - \theta$, where $\boldsymbol{a} = (a_1, \ldots, a_{d_y})^\top$ is a coefficient vector satisfying $\|\boldsymbol{a}\| = 1$, and θ is the *discrimination threshold* such that the larger θ we set, the more samples removed. A sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ was kept paired if $f(\boldsymbol{y}_i) > 0$, and $\boldsymbol{y}_i$ 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:

$$C(\boldsymbol{W}_x, \boldsymbol{W}_x^*, \Lambda^*) = \sum_{i=1}^{r} \lambda_i^* \frac{\boldsymbol{w}_{x,i}^\top \boldsymbol{w}_{x,i}^*}{\|\boldsymbol{w}_{x,i}\| \cdot \|\boldsymbol{w}_{x,i}^*\|},$$

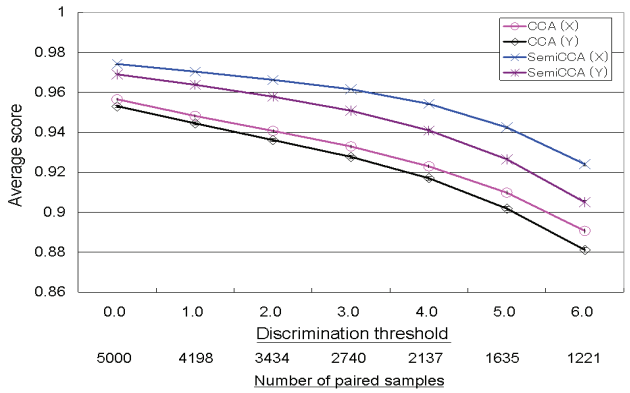

Fig. 2. Average evaluation score for artificial data

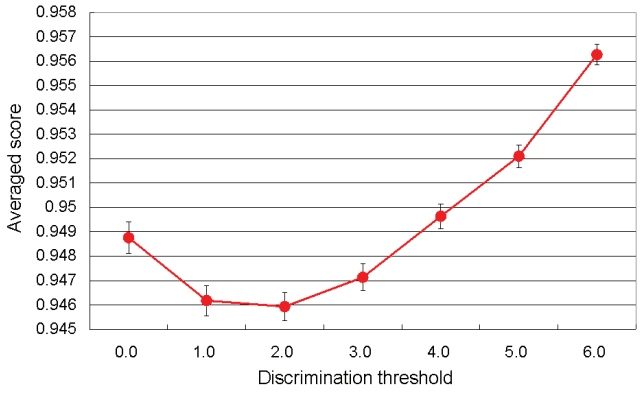

Fig. 3. Average trade-off parameter taking the highest score.

where $\boldsymbol{W}_x^* = (\boldsymbol{w}_{x,1}^*, \boldsymbol{w}_{x,2}^*, \ldots, \boldsymbol{w}_{x,d_z}^*)^\top$ and $\Lambda^* = \mathrm{diag}(\lambda_1^*, \lambda_2^*, \ldots, \lambda_{d_z}^*)$ 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 2 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 3 shows the trade-off parameter taking the highest score averaged over all the trials, and Figure 4 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. 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 2.

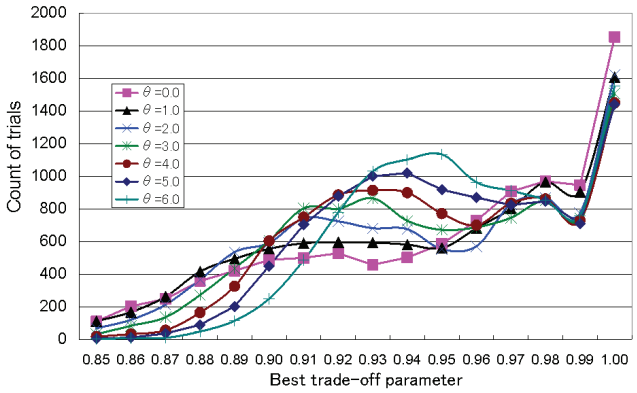

Fig. 4. Histogram of trade-off parameters taking the highest score.

V. Applications to Automatic Image Annotation

A. Method

We describe the method of automatic image annotation based on the one by Nakayama et al. [9], [10] with the help of SemiCCA.

First of all, feature vectors are extracted from images and associated text labels. Each text label is composed of text words selected from a word set given in advance. We utilize Bag of Features (BoF) as image features $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{N_x}$, where SURF [11] is used for key-point detection and descriptor extraction, and binary word vectors $\boldsymbol{Y} = \{\boldsymbol{y}_n\}_{n=1}^{N}$ as text features. Note that $\{\boldsymbol{x}_n\}_{n=1}^{N}$ are labeled images, while $\{\boldsymbol{x}_n\}_{n=N+1}^{N_x}$ are unlabeled.

Next, a topic model is constructed from feature vectors $(\boldsymbol{X}, \boldsymbol{Y})$ with the help of SemiCCA. The first step is to generate latent variables $\boldsymbol{Z} = \{\boldsymbol{z}_n\}_{n=1}^{N}$ with SemiCCA. More specifically, a function $f_x : \mathcal{R}^{d_x} \to \mathcal{R}^{d_z}$ is derived from $\boldsymbol{X}$ as training samples with SemiCCA as $f_x(\boldsymbol{x}) = \boldsymbol{\Lambda}^{1/2}\boldsymbol{W}_x\boldsymbol{x}$, and latent variables $\boldsymbol{Z}$ are generated from $\boldsymbol{X}$ with f_x. The second step is to set up a topic model. The topic model is described by the following equation:

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

where $y_{n,d}$ is the d-th element of $\boldsymbol{y}_n$, N_d is the number of the images containing the d-th word in labeled samples, $\delta(\cdot)$ is Dirac delta, an operator $\propto$ stands for proportion, and γ and μ are constants.

Once the model estimation has been finished, we can execute automatic image annotation 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 an image feature $\boldsymbol{x}^{(g)}$ extracted from a given image, as follows:

$$\widehat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y} \in [0,1]^{d_y}} p(\boldsymbol{y}|\boldsymbol{x}^{(g)}) \tag{4}$$

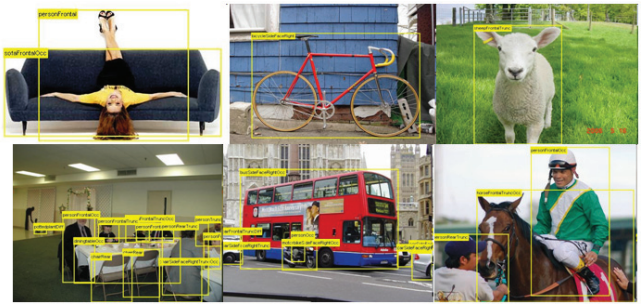

Fig. 5. Example images in PASCAL VOC data set

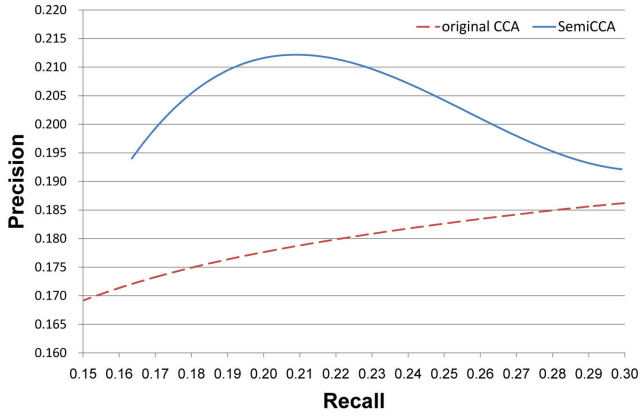

Fig. 6. Experiment results for automatic image annotation

$$= \underset{\boldsymbol{y}\in[0,1]^{d_y}}{\operatorname{argmax}} \frac{\sum_{n=1}^{N} p(\boldsymbol{x}^{(g)}|\boldsymbol{z}_n)p(\boldsymbol{y}|\boldsymbol{z}_n)}{\sum_{n=1}^{N} p(\boldsymbol{x}^{(g)}|\boldsymbol{z}_n)}. \quad (5)$$

B. Experiments

We next evaluated the effectiveness of SemiCCA under the task of automatic image annotation with PASCAL Visual Object Challenge (VOC) 2008/2009 data sets [2]. This data set 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 are put on Fig. 5. As you can see, each image has a bounding box and object class label for each object presented in the image. However, we *removed all the bounding boxes* and only utilized class labels to simulate "weak labeling" settings [12], where images are weakly related to multiple words without region information.

We utilized all of the 5096 images in VOC2008 training data set, and separated them into 500 labeled images for training, 500 unlabeled images for evaluation and the rest (4096 images) as unlabeled images for training. Also, 9647 images in VOC2009 training/test data set were added to unlabeled images for training. In total, 13743 unlabeled images for training were utilized. We adopted the precision rate PR and recall rate RE as the evaluation measures.

Fig. 6 shows the experimental results for the automatic annotation task. 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 the one by the original CCA. Fig. 6 indicates that a topic model build with the help of SemiCCA outperformed that of the original CCA. The total performance seemed to be poor for VOC participants, however, it should be noted again that we *removed all the bounding boxes* and only utilized class labels.

VI. Concluding Remarks

In this paper, we proposed a new semi-supervised variant of CCA that we named *SemiCCA*. Unlike the previous semi-supervised CCA, our formulation is quite simple and also intuitively comprehensive. Namely, SemiCCA smoothly bridges CCA with paired samples and PCA with paired and unpaired samples by a trade-off parameter. We evaluated its performance by using artificially generated samples and PASCAL VOC data set, and revealed the effectiveness of SemiCCA against the original CCA.

Our future work includes some comparison of SemiCCA with other semi-supervised variants of CCA, especially based on Tikhonov regulirization [5], [6], mutually complementary integration of SemiCCA into some methods of semi-supervised learning based on graph Laplacian reguralization, and applications to various challenging real-world problems e.g. automatic music/image/video annotation and retrieval, and multi-model event correlation analysis for audio-video synchronization and audio-visual speech recognition.

References

[1] J. S. 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 et al., "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 et al. *Semi-Supervised Learning*, MIT Press, 2006.

[5] D. R. Hardoon et al. "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[6] M. B. Blaschko et al. "Semi-supervised laplacian regularization of kernel canonical correlation analysis," in Proc. ECML PKDD '08, Berlin, Heidelberg, 2008, pp. 133–145.

[7] M. Sugiyama et al. "Semi-supervised local Fisher discriminant analysis for dimensionality reduction," *Machine Learning*, vol. 78, no. 1–2, pp. 35–61, 2010.

[8] 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*, pp. 253–264, 1993.

[9] H. Nayayama et al. "High-performance image annotation and retrieval for weakly labeled images," in *Proc. PCM*, pp. 601–610. 2008,

[10] T. Harada et al. "Image annotation retrieval based on efficient learning of contextual latent space," in *Proc. ICME*, 2009.

[11] H. Bay et al. "Speeded-up robust features (surf)," *CVIU*, vol. 110, no. 3, pp.346–359, 2008.

[12] G. Carneiro et al. "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans PAMI*, vol. 29, no. 3, pp. 394–410, 2007.

[13] M. Wang et al. "Semi-supervised kernel density estimation for video annotation," *CVIU*, vol. 113, no. 3, pp.384–396, 2009.