AUTOMATIC AUDIO TAG CLASSIFICATION VIA SEMI-SUPERVISED CANONICAL DENSITY ESTIMATION

Jun Takagi†, *Yasunori Ohishi*‡, *Akisato Kimura*‡, *Masashi Sugiyama*†, *Makoto Yamada*†, *Hirokazu Kameoka*‡

†Graduate School of Information Science and Engineering, Tokyo Institute of Technology
‡NTT Communication Science Laboratories, NTT Corporation

ABSTRACT

We propose a novel semi-supervised method for building a statistical model that represents the relationship between sounds and text labels ("tags"). The proposed method, named *semi-supervised canonical density estimation*, makes use of unlabeled sound data in two ways: 1) a low-dimensional latent space representing topics of sounds is extracted by a semi-supervised variant of canonical correlation analysis, and 2) topic models are learned by multi-class extension of *semi-supervised kernel density estimation* in the topic space. Real-world audio tagging experiments indicate that our proposed method improves the accuracy even when only a small number of labeled sounds are available.

Index Terms— Audio tag classification, semi-supervised learning, topic model, canonical correlation analysis, kernel density estimation

1. INTRODUCTION

A central goal of music information retrieval is to develop a system that can efficiently store and retrieve sounds from a large database of musical contents. The most common way is to use *metadata* such as the name of composers and artists, the title of songs, and the release date of albums; furthermore, one may also use various additional information such as genres, instruments, song lyrics, and reviews. These metadata can be used as input to query-by-keyword systems and collaborative recommender systems [1, 2, 3]. However, all these systems have a common drawback called the *new item problem*—when a new song is added to the database, it needs to be annotated manually. This requires a large amount of human labor, and thus keeping the systems updated is highly costly.

To cope with this problem, most previous works have tried to automatically associate sounds with words for query-by-text retrieval or music annotation [4, 5, 6, 7, 8, 9, 10, 11]. Recently, inference techniques based on *topic models*, such as *probabilistic latent semantic analysis* (pLSA) and *latent Dirichlet allocation* (LDA), have been exploited for automatic image annotation and retrieval [12, 13]. Since *canonical correlation analysis* (CCA) [14] can be interpreted as an approximation to Gaussian pLSA and also be regarded as an extension of *Fisher linear discriminant analysis* (FDA) to multi-label classification [15], learning topic models through CCA is not only computationally efficient, but also promising for multi-label audio annotation and retrieval.

All of the methods explained above considered a supervised learning setup, where annotated sounds are used as training data. Thus, in order to further improve the annotation and retrieval accuracy, high-quality semantic information about sounds would be necessary. However, gathering such high-quality semantic information is expensive, and thus a big raise in the number of annotated sounds cannot be expected in reality. On the other hand, a large number of unannotated sounds which have not yet been registered to the system can be readily collected. Thus, a semi-supervised learning approach which utilizes a small number of annotated sounds and a large number of unannotated sounds would be promising.

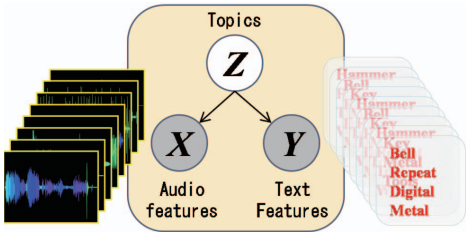

Fig. 1. Topic model for audio tag classification

In this paper, we propose a semi-supervised learning method for generic topic models named *semi-supervised canonical density estimation* (SSCDE). SSCDE fully makes use of unannotated sounds for both feature extraction and model estimation. More specifically, we estimate a low-dimensional latent space representing topics of music by applying a semi-supervised variant of CCA called *SemiCCA* [16], which extends the ordinary CCA to be able to utilize both paired and unpaired samples. Then, in the estimated latent space, topic models are learned by multi-class extension of semi-supervised non-parametric density estimation called *semi-supervised kernel density estimation* (SSKDE) [17]. Through real-world audio tagging experiments, we demonstrate the effectiveness of the proposed approach.

2. FRAMEWORK

Figs. 2 and 3 show the framework of the proposed method briefly. Let $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{N+N_x}$ and $\boldsymbol{Y} = \{\boldsymbol{y}_n\}_{n=1}^{N}$ be a set of audio and semantic features with D_x and D_y dimensions, where N and N_x are the number of labeled and unlabeled sounds, respectively. A topic model is estimated from feature vectors $(\boldsymbol{X}, \boldsymbol{Y})$, which consists of two steps.

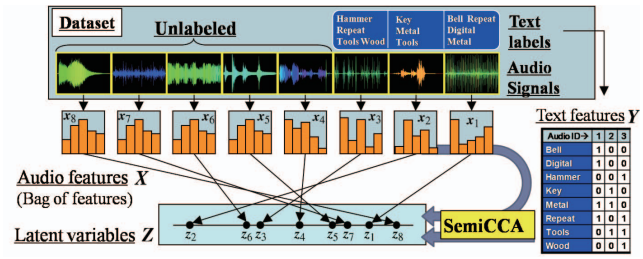

Fig. 2. Framework of the proposed method (1st part)

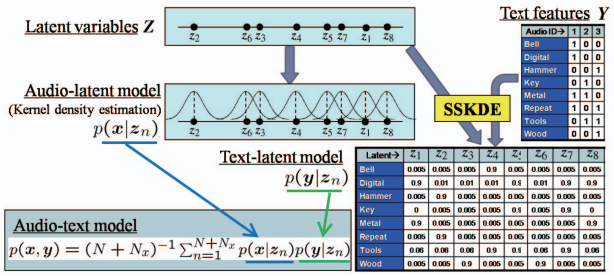

Fig. 3. Framework of the proposed method (2nd part)

The first step is to generate a latent variable $\boldsymbol{Z} = \{\boldsymbol{z}_n\}_{n=1}^{N+N_x}$ of D_z dimensions with SemiCCA. More specifically, a pair (f_x, f_{xy}) of functions $f_x(\boldsymbol{x})$ and $f_{xy}(\boldsymbol{x},\boldsymbol{y})$ is derived from $(\boldsymbol{X},\boldsymbol{Y})$ as training samples via SemiCCA, and latent variables $\boldsymbol{Z}$ are generated from $(\boldsymbol{X},\boldsymbol{Y})$ with (f_x, f_{xy}). We describe the detail of the first step in Section 4.1.

The second step is to set up a topic model with the help of kernel density estimation (KDE) on the latent space:

$$p(\boldsymbol{x},\boldsymbol{y}) = (N+N_x)^{-1}\sum_{n=1}^{N+N_x} p(\boldsymbol{x}|\boldsymbol{z}_n)p(\boldsymbol{y}|\boldsymbol{z}_n). \quad (1)$$

The main procedure in this step is to estimate the conditional densities $p(\boldsymbol{x}|\boldsymbol{z}_n)$ and $p(\boldsymbol{y}|\boldsymbol{z}_n)$ of features $(\boldsymbol{x},\boldsymbol{y})$ for every latent variable $\boldsymbol{z}_n$. Note that a conditional density $p(\boldsymbol{y}|\boldsymbol{z}_n)$ $(n = N+1, N+2, \ldots, N+N_x)$ can be derived even though the corresponding text label $\boldsymbol{y}_n$ does not exist. We show the detailed procedure in Section 4.2.

Once the model estimation has been finished, we are ready to annotate an unseen sound s through maximum a posteriori (MAP) estimation. The most probable semantic feature $\widehat{\boldsymbol{y}}$ can be derived by using an audio feature $\boldsymbol{x}^{(s)}$ extracted from a given sound s as

$$\widehat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y}\in[0,1]^{D_y}} p(\boldsymbol{y}|\boldsymbol{x}^{(s)}) \quad (2)$$

$$= \operatorname*{argmax}_{\boldsymbol{y}\in[0,1]^{D_y}} \sum_{n=1}^{N+N_x} p(\boldsymbol{x}^{(s)}|\boldsymbol{z}_n)p(\boldsymbol{y}|\boldsymbol{z}_n). \quad (3)$$

When the d-th element of $\widehat{\boldsymbol{y}}$ exceeds a pre-defined threshold θ, the text word of index d is provided to the given sound s.

3. FEATURE REPRESENTATIONS

We describe an annotated sound corpus "Freesound" used in our experiments and introduce audio and semantic feature representations.

The Freesound Project is a collaborative database of Creative Commons licensed sounds [18]. Each sound sample has been annotated using a vocabulary (e.g., genre, instrumentation, emotion, style, rhythm). We extract 2,012 WAV files with a 44.1 kHz sampling rate, a 16 bit depth, and mono or stereo. The stereo audio files are converted to the mono ones by taking the average of the left and right channels. Then, we obtain a vocabulary of 230 words used for annotating these sounds.

Each sound is represented as a *bag-of-feature-vectors* calculated by analyzing a short-time segment of the audio signal. Specifically, we represent the audio with a time series of *Mel-frequency cepstral coefficients* (MFCCs) feature vectors which are popular features for speech recognition and music classification [9]. A time series of MFCC vectors is extracted by sliding a half-overlapping, short-time window (23 ms) over the audio file. The dynamic features (MFCC-Delta) are derived by calculating the first and second instantaneous derivatives of each element over consecutive vectors and are appended them to the vector of MFCCs. We use the first 13 MFCCs resulting in about 5,200 39-dimensional feature vectors per minute of audio contents. We create a vector quantization (VQ) codebook of size D_x using about 1,000,000 feature vectors which are sampled randomly so that each sound is represented by 500 feature vectors. Then, normalized code histograms of VQ results are used as D_x-dimensional audio feature vectors $\boldsymbol{X} = \{\{\boldsymbol{x}_n\}_{n=1}^{N}, \{\boldsymbol{x}_n\}_{n=N+1}^{N+N_x}\}$ (D_x=1,024), representing the acoustic characteristics of sounds, where $\{\boldsymbol{x}_n\}_{n=1}^{N}$ are labeled sounds, while $\{\boldsymbol{x}_n\}_{n=N+1}^{N+N_x}$ are unlabeled sounds.

On the other hand, we also extract D_y-dimensional binary annotation vectors $\boldsymbol{Y} = \{\boldsymbol{y}_n\}_{n=1}^{N}$ ($D_y = 230$) from the 2,012 sounds. Each element of y is set to 1 if the corresponding word is annotated and 0 otherwise.

4. ASSOCIATING SOUNDS WITH WORDS

We describe a semi-supervised learning method for generic topic models called SSCDE to associate souonds with words.

4.1. Semi-supervised CCA

We have proposed SemiCCA [16] that combines CCA with principal component analysis (PCA) for utilizing unlabeled samples. Let us explain the idea of SemiCCA using an illustrative two-dimensional data set depicted in Fig. 4, where labeled (resp. unlabeled) samples are plotted with white (resp. red and blue). When only the labeled samples are used, poor projection directions may be obtained by CCA due to overfittng. In contrast, unlabeled samples may be used for revealing the global structure in each domain. Note that once a basis in one sample space is rectified, the corresponding basis in the other sample space is also rectified so that correlations between two bases are maximized.

Motivated by the above illustration, we smoothly combine the eigenvalue problems of CCA and PCA. More specif-

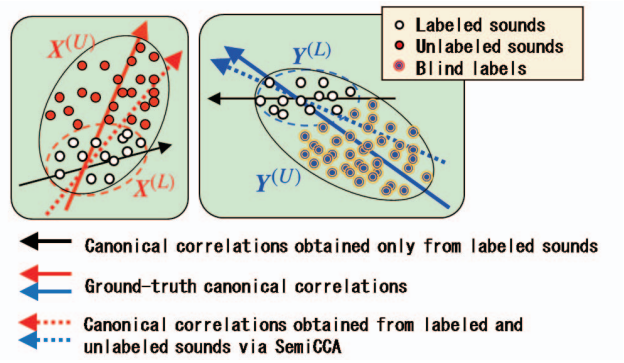

Fig. 4. Effects of unlabeled samples in SemiCCA

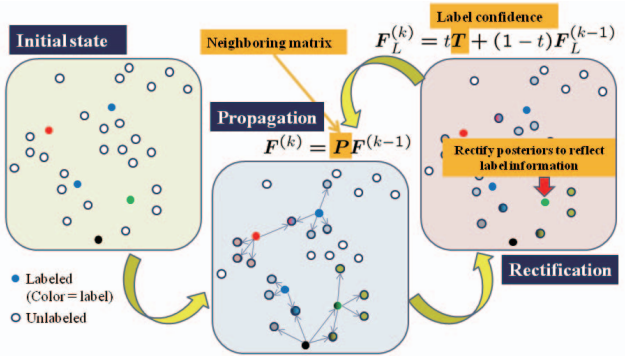

Fig. 5. SSKDE procedure

ically, the solution of SemiCCA is given by the leading generalized eigenvectors of the following generalized eigenvalue problem:

$$\boldsymbol{B}\boldsymbol{w} = \lambda \boldsymbol{C}\boldsymbol{w}, \quad \boldsymbol{w} = (\boldsymbol{w}_x, \boldsymbol{w}_y)^\top \tag{4}$$

$$\boldsymbol{B} = \beta \begin{pmatrix} \boldsymbol{0} & \boldsymbol{S}_{xy}^{(L)} \\ \boldsymbol{S}_{yx}^{(L)} & \boldsymbol{0} \end{pmatrix} + (1-\beta) \begin{pmatrix} \boldsymbol{S}_{xx} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{yy} \end{pmatrix}, \tag{5}$$

$$\boldsymbol{C} = \beta \begin{pmatrix} \boldsymbol{S}_{xx}^{(L)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{yy}^{(L)} \end{pmatrix} + (1-\beta) \begin{pmatrix} \boldsymbol{I}_{D_x} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{D_y} \end{pmatrix}, \tag{6}$$

where $\boldsymbol{S}_{xx}$ is a scatter matrix of $\boldsymbol{X}$, $\boldsymbol{S}_{xy}^{(L)}$ is a scatter matrix obtained from a pair of labeled samples in $(\boldsymbol{X}, \boldsymbol{Y})$, (all the other scatter matrices can be defined similarly), and β $(0 \leq \beta \leq 1)$ is a constant named *a trade-off parameter*. The trade-off parameter controls the trade-off between CCA and PCA. Namely, when $\beta = 1$, the problem is reduced to the CCA eigenvalue problem, while when $\beta = 0$ the problem is reduced to the PCA eigenvalue problem, under the assumption that $\boldsymbol{X}$ and $\boldsymbol{Y}$ 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. Picking up the top D_z generalized eigenvectors as row vectors, we can obtain D_z-dimensional mappings $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$.

By introducing a stochastic interpretation of CCA [15], we can derive two types of functions f_{xy} and f_x to obtain a latent variable $\boldsymbol{z}$ as follows:

$$f_{xy}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Lambda}^{1/2}(\boldsymbol{I}_{D_z} + \boldsymbol{\Lambda})^{-1}(\boldsymbol{W}_x \boldsymbol{x} + \boldsymbol{W}_y \boldsymbol{y}), \tag{7}$$

$$f_x(\boldsymbol{x}) = \boldsymbol{\Lambda}^{1/2}\boldsymbol{W}_x \boldsymbol{x}, \tag{8}$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix with the d-th diagonal component being the d-th largest singular value λ_d $(d = 1, 2, \ldots, D_z)$.

4.2. Semi-supervised KDE

Based on the idea of KDE, we introduce the following topic model $p(\boldsymbol{x}, \boldsymbol{y})$ [19] describing the relationship between an audio feature $\boldsymbol{x}$ and a semantic feature $\boldsymbol{y}$ as

$$p(\boldsymbol{x}, \boldsymbol{y}) = (N + N_x)^{-1} \textstyle\sum_{n=1}^{N+N_x} p(\boldsymbol{x}|\boldsymbol{z}_n) p(\boldsymbol{y}|\boldsymbol{z}_n), \tag{9}$$

$$p(\boldsymbol{x}|\boldsymbol{z}_n) = \kappa(f_x(\boldsymbol{x}) - \boldsymbol{z}_n), \tag{10}$$

$$p(\boldsymbol{y}|\boldsymbol{z}_n) = \textstyle\prod_{d=1}^{D_y} p(y_d|\boldsymbol{z}_n), \tag{11}$$

$$p(y_d|\boldsymbol{z}_n) = \mu\delta(y_d - y_{n,d}) + (1-\mu)N_d/N, \tag{12}$$

where each latent variable $\boldsymbol{z}_n$ is calculated as

$$\boldsymbol{z}_n = \begin{cases} f_{xy}(\boldsymbol{x}_n, \boldsymbol{y}_n), & n = 1, \ldots, N \\ f_x(\boldsymbol{x}_n), & n = N+1, \ldots, N_x \end{cases} \tag{13}$$

$\kappa(\cdot)$ is a Gaussian kernel with a paramter γ, $\delta(\cdot)$ is the Dirac delta, $y_{n,d}$ is the d-th element of $\boldsymbol{y}_n$, N_d is the number of the audio signals containing the d-th word in labeled sounds, and μ $(0 < \mu < 1)$ is a parameter representing how reliable a given label is.

Note that this strategy with the traditional KDE can utilize only labeled samples to estimate a topic model $p(\boldsymbol{x}, \boldsymbol{y})$ due to lack of labels (cf. Eq. (12)). This indicates that the accuracy of density estimation heavily relies on the number of labeled samples. Thus, we introduce an idea of semi-supervised KDE [17] used for discrimination tasks, and extend it to multi-label classification.

From Eq. (11), each conditional density $p(y_d|\boldsymbol{z}_n)$ can be viewed as a posterior of the d-th "class" given a "feature" $\boldsymbol{z}_n$. Then, we can apply the idea of SSKDE to the estimation of this conditional density. For simplicity, we introduce the following matrix forms

$$\boldsymbol{P} = \{P_{n,m}\}_{n,m=1}^{N+N_x}, \quad \boldsymbol{F} = \{F_{n,d}\}_{n=1}^{N+N_x \; D_y}_{d=1}, \tag{14}$$

$$P_{n,m} = \frac{\kappa(\boldsymbol{z}_n - \boldsymbol{z}_m)}{\sum_{m'=1}^{N+N_x} \kappa(\boldsymbol{z}_n - \boldsymbol{z}'_m)}, \tag{15}$$

$$F_{n,d} = p(y_d|\boldsymbol{z}_n). \tag{16}$$

What we want to derive is the matrix $\boldsymbol{F}$ of conditional densities, which can be obtained through an iterative procedure (see [17] for detail). Fig. 5 briefly depicts the procedure of SSKDE.

5. EXPERIMENTS

This section describes the results for the automatic audio annotation task. We used 2,012 sounds in the Freesound dataset,

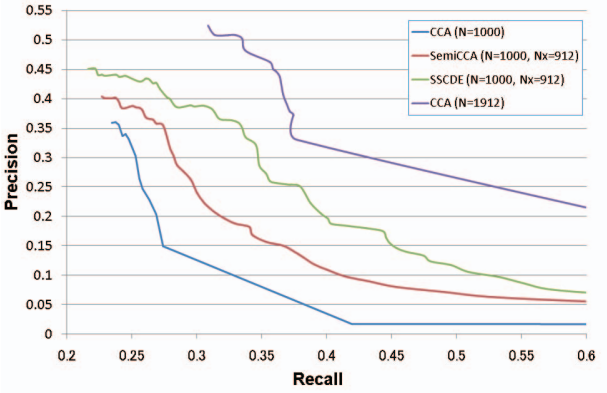

Fig. 6. Results for the automatic annotation task

and separated them into 1,912 sounds ($N + N_x = 1,912$) for training and 100 sounds for evaluation. Parameters D_z, β, γ, μ were set to 100, 0.99, 0.8, 0.99, respectively. We determined these parameters experimentally. As the evaluation measures, we calculated the precision, recall and F-value based on evaluation dataset.

Fig. 6 shows the experimental results for the automatic annotation task, where the threshold θ (see the last sentence in Section 2) is varied from 0 to 5.0. In Tab. 1, we show recall and precision when the F-value reach a maximum value. We separated 1912 sounds for training into 1000 labeled sounds ($N = 1,000$) and 912 unlabeled sounds ($N_x = 912$), and learned the proposed model using these sounds. We compared the proposed model with the CCA-based model using all of training data as labeled sounds ($N = 1,912$), the CCA-based model using only $\boldsymbol{X}^{(L)}$ ($N = 1,000$), and the SemiCCA-based model using $\boldsymbol{X}^{(L)}$ and $\boldsymbol{X}^{(U)}$, namely multi-class SSKDE was removed from the proposed method. As the table shows, the topic model built with the help of SSCDE outperformed that of SemiCCA. This is because the proposed method incorporate semi-supervised latent space estimation with SemiCCA and semi-supervised nonparametric model estimation with SSKDE into the existing learning method of topic models.

6. CONCLUDING REMARKS

We have developed a new and efficient method to learn topic models in a semi-supervised manner, named semi-supervised canonical density estimation (SSCDE), and presented a way to integrate it to audio annotation/retrieval. The proposed method contained two novel contributions: (a) semi-supervised latent space estimation with SemiCCA and (b) semi-supervised non-parametric model estimation with SSKDE. Experiments with thousands of audio signals have demonstrated that the proposed method is promising.

In the future, we plan to evaluate the proposed method in detail on a larger database. We also want to compare the performance of the proposed method with that of the conventional methods [9, 11] and apply to various challenging real-world problems e.g., multi-modal event correlation analysis for audio-video synchronization and audio-visual speech recognition.

Table 1. Precision, Recall, and F-value

	Precision	Recall	F-value
CCA (N=1,000)	0.355	0.240	0.287
SemiCCA (N=1,000, N_x=912)	0.355	0.274	0.310
SSCDE (N=1,000, N_x=912)	0.359	0.333	0.345
CCA (N=1,912)	0.462	0.356	0.402

7. ACKNOWLEDGMENT

The authors are grateful to Dr. Gordon Wichern of Arizona State University for assistance with experiments. The authors would like to thank Mr. Takuho Nakano of the University of Tokyo, Dr. Hitoshi Sakano, Dr. Eisaku Maeda, Dr. Katsuhiko Ishiguro, Dr. Kunio Kashino and Dr. Hidehisa Nagano of NTT Communication Science Laboratories for their valuable discussions.

8. REFERENCES

[1] W. W. Cohen and W. Fan, "Web-collaborative filtering: Recommending music by crawling the Web," in *Proc. WWW9*, 2000.

[2] A. Uitdenbogerd and R. van Schyndel, "A review of factors affecting music recommender success," in *Proc. ISMIR 2002*.

[3] P. Knees et al., "Artist classification with web-based data," in *Proc. ISMIR 2004*.

[4] M. Slaney, "Semantic-audio retrieval," in *Proc. ICASSP 2002*.

[5] B. Whitman et al., "Automatic record reviews," in *Proc. ISMIR 2004*.

[6] P. Knees et al., "A music search engine built upon audio-based and web-based similarity measures," in *Proc. SIGIR 2007*.

[7] D. Torres et al., "Identifying words that are musically meaningful," in *Proc. ISMIR 2007*.

[8] L. Barrington et al., "Combining feature kernels for semantic music retrieval," in *Proc. ISMIR 2008*.

[9] D. Turnbull et al., "Semantic annotation and retrieval of music and sound effects," *IEEE TASLP*, vol. 16, no. 2, pp. 467–476, 2008.

[10] R. Takahashi et al., "Building and combining document and music spaces for music Query-By-Webpage system," in *Proc. INTERSPEECH 2008*.

[11] D. Turnbull et al., "Combining audio content and social context for semantic music discovery," in *Proc. SIGIR 2009*.

[12] K. Barnard et al., "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

[13] Li Fei-Fei and P. Pietro, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. CVPR 2005*.

[14] 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.

[15] F. R. Bach and M. Jordan, "A probabilistic interpretation of canonical correlation analysis," Tech. Rep. 688, Department of Statistics, University of California, Berkeley, 2005.

[16] A. Kimura et al., "SemiCCA: Efficient semi-supervised learning of canonical correlations," in *Proc. ICPR 2010*.

[17] M. Wang et al., "Semi-supervised kernel density estimation for video annotation," *Computer Vision and Image Understanding*, vol. 113, no. 3, pp. 384–396, 2009.

[18] The Freesound Project, "http://www.freesound.org/," .

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