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

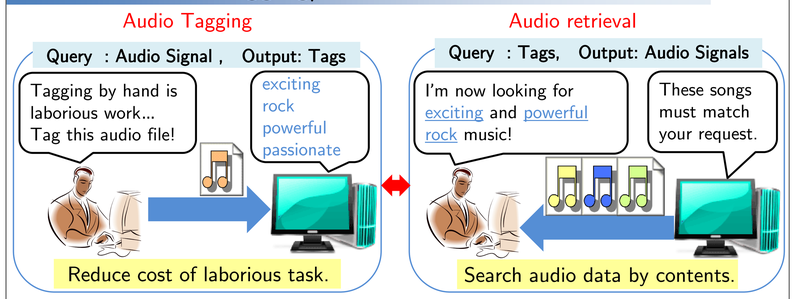
Contact information
E-mail:
takagi@sg.cs.titech.ac.jp

Overview

Apply semi-supervised method to audio tagging/retrieval.

Utilize untagged samples and reduce the needed number of expensive tagged samples.

What is Audio Tagging/Retrieval?



SSCDE can cope with both tasks in the same framework!

Technical Challenges : Use Inexpensive Untagged Audios

Audio samples having high-quality tags are very expensive!

Semi-supervised method utilizes inexpensive untagged audio samples!

Easy to collect at low cost!

Technical Challenges : Use Tag Co-occurrence Information

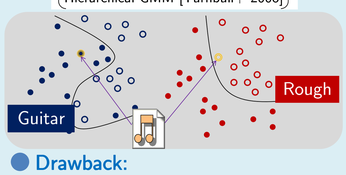
Tag co-occurrence information seems to be useful for tagging/retrieval task.

But almost all of existing method cannot utilize this information.

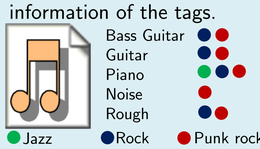
Typical approach

- Train a dedicated classifier for each tag separately.

Linear regression [Whiteman+ 2004]
Hierarchical GMM [Turnbull+ 2008]



- Drawback: Difficult to use co-occurrence information of the tags.



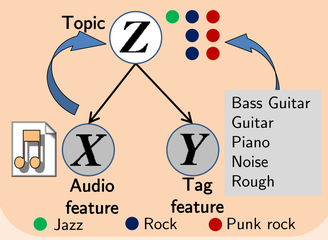
→ How to incorporate the co-occurrence information into the classifier?

Topic model based approach

- Design probabilistic generative model for audio signals and tags.

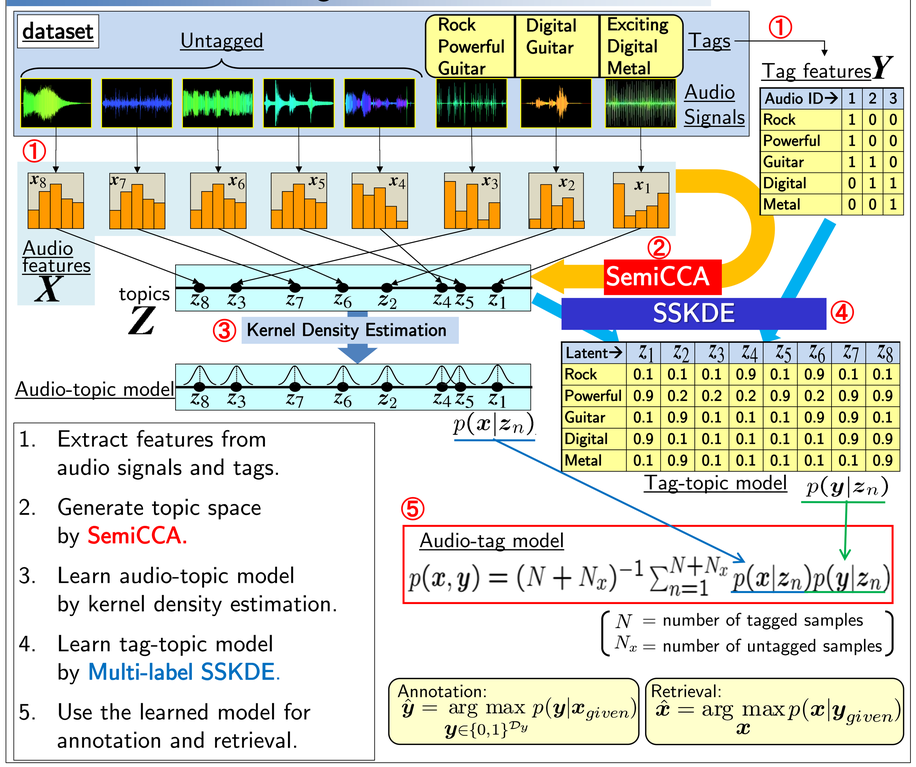
pLSA: NLP [Hofmann 1999] Image [Barnard+ 2001]
LDA : NLP[Blei+ 2003], Image [Li+ 2005]

$$p(x, y) = \int_z p(z)p(x|z)p(y|z)dz$$



- Merits:
  1. Tag co-occurrence is considered in the model.
  2. Scalable to many features.

SSCDE : Model Learning Framework



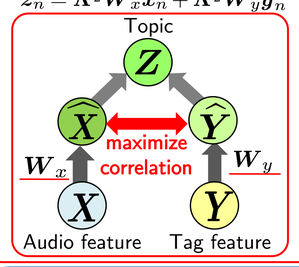
1. Extract features from audio signals and tags.
2. Generate topic space by SemiCCA.
3. Learn audio-topic model by kernel density estimation.
4. Learn tag-topic model by Multi-label SSKDE.
5. Use the learned model for annotation and retrieval.

Technical Points of SSCDE

1. Learn topic space with tagged and untagged samples : SemiCCA [Kimura+ ICPR2010]

Learn the map :

$$z_n = \Lambda^{\frac{1}{2}} W_x x_n + \Lambda^{\frac{1}{2}} W_y y_n$$



SemiCCA can utilize untagged samples differently from standard CCA.

Solution

generalized eigenproblem.

$$B \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda C \begin{bmatrix} w_x \\ w_y \end{bmatrix}$$

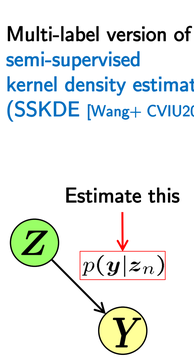
CCA (tagged samples) PCA (tagged and untagged samples)

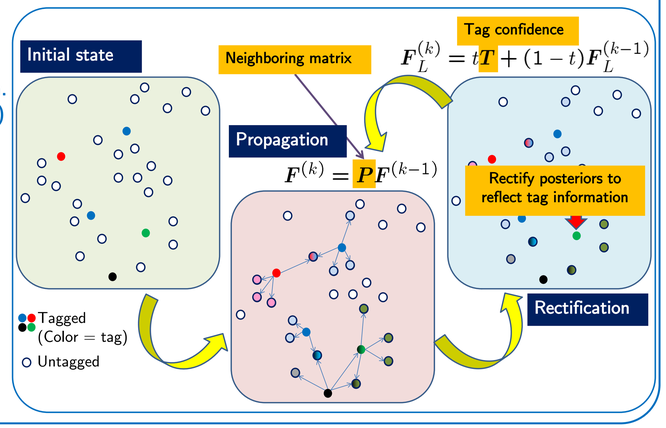
$$B = \beta \begin{bmatrix} 0 & S_{xy}^{(T)} \\ S_{yx}^{(T)} & 0 \end{bmatrix} + (1-\beta) \begin{bmatrix} S_{xx} & 0 \\ 0 & S_{yy} \end{bmatrix}$$

$$C = \beta \begin{bmatrix} S_{xx}^{(T)} & 0 \\ 0 & S_{yy}^{(T)} \end{bmatrix} + (1-\beta) \begin{bmatrix} I_{D_x} & 0 \\ 0 & I_{D_y} \end{bmatrix}$$

2. Propagate tag information : Multi-label SSKDE

Multi-label version of semi-supervised kernel density estimation. (SSKDE [Wang+ CVIU2009])





Experiment

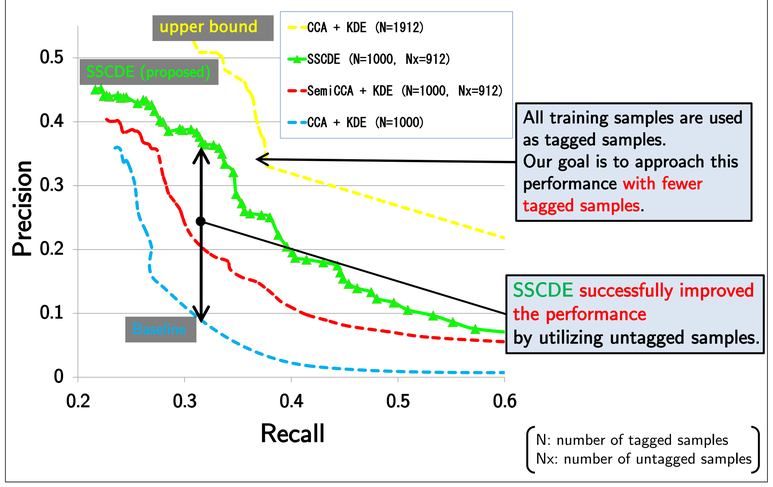
Annotation performance of SSCDE is evaluated under following condition.

- **Dataset**: 2012 audio files taken from "Freesound" (http://www.freesound.org/).
  - Database of Creative Commons licensed sounds.
  - Annotated with vocabulary.
- **Evaluation condition:**
  - 2012 audio clips with WAV format.
  - Each clip has multiple tags selected from 230 words.

Tagged audio signal	Untagged audio signal	Evaluation sample
1000	912	100

- **Audio feature**: bag-of-feature-vectors extracted by the following process.
  1. Audio signals are splitted into half-overlapping 23ms windows, and 39-dimensional vector (including first 13 MFCC, MFCC-Δ, MFCC-Δ-Δ) is extracted from each window.
  2. 500 vectors are sampled from each audio signal (about 1,000,000 vectors in total). LBG algorithm (Linde-Buzo-Gray algorithm, algorithm for vector quantization : VQ) is applied to them, and VQ codebook (size :1024) is obtained.
  3. 1024-dimensional normalized vector representing each audio signal is created by VQ.
- **Tag feature**: 230-dimensional binary vector. (Each element of the vector corresponds to specific tag.)

Result



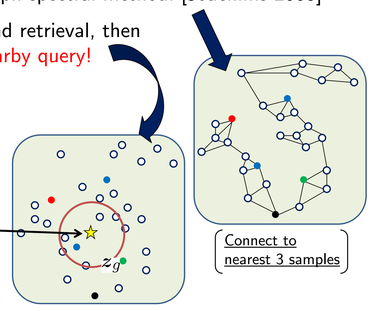
Use Sparse Matrix as Neighboring Matrix

- Fix the number of non-zero elements in each row, then required memory size to hold the neighboring matrix decreases. $(\mathcal{O}(N^2) \longrightarrow \mathcal{O}(N))$
- In this case, SSKDE is equivalent to Graph spectral method. [Joachims 2003]
- Apply the same idea to audio tagging and retrieval, then calculation needs only a few samples nearby query!

$$\hat{y} = \sum_{n=1}^{N} \kappa(z_g, z_n)\ p(y = all\ 1|z_n)$$

Gaussian kernel — Posterior probability of y given Zn

$$\hat{y} = \sum_{z_n \in \mathcal{N}(z_g)} \kappa(z_g, z_n)\ p(y = all\ 1|z_n)$$



Annotation/Retrieval task becomes equivalent to neighboring search problem! (Computational complexity is $\mathcal{O}(\log N)$!)