COLLECTIVE ANNOTATION OF MUSIC FROM MULTIPLE SEMANTIC CATEGORIES

Zhiyao Duan[1,2] **Lie Lu**[1] **Changshui Zhang**[2]
[1]Microsoft Research Asia (MSRA), Sigma Center, Haidian District, Beijing 100080, China.
[2]State Key Laboratory of Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Automation, Tsinghua University, Beijing 100084, China.
duanzhiyao00@mails.tsinghua.edu.cn, llu@microsoft.com, zcs@tsinghua.edu.cn

ABSTRACT

Music semantic annotation aims to automatically annotate a music signal with a set of semantic labels (words or tags). Existing methods on music semantic annotation usually take it as a multi-label binary classification problem, and model each semantic label individually while ignoring their relationships. However, there are usually strong correlations between some labels. Intuitively, investigating this correlation can be helpful to improve the overall annotation performance. In this paper, we report our attempts to collective music semantic annotation, which not only builds a model for each semantic label, but also builds models for the pairs of labels that have significant correlations. Two methods are exploited in this paper, one based on a generative model (Gaussian Mixture Model), and another based on a discriminative model (Conditional Random Field). Experiments show slight but consistent improvement in terms of precision and recall, compared with the individual-label modeling methods.

1 INTRODUCTION

Semantic annotation of music signals have become an important direction in music information retrieval. With music annotation approaches, a music signal is associated with a set of semantic labels (text, words), which is a more compact and efficient representation than the raw audio or low level features. It can also potentially facilitate a number of music applications, such as music retrieval and recommendation, since it is more natural for a user to describe a song by semantic words, and it is more flexible to measure music similarities with vectors of semantic labels.

Several methods have been proposed for automatic music semantic annotation, which basically can be classified into two categories: *non-parametric* and *parametric*. Non-parametric methods model the text-audio relations implicitly. For example, Slaney [10] created separate hierarchical models in the acoustic and text spaces and then linked the two spaces for annotation and retrieval. Cano and Koppenberger [2] proposed an approach to predict the semantic words based on nearest neighbor classification. On the other hand, parametric methods explicitly model the text-audio relations. For instance, Whitman *et al.* [14, 13] trained a one-versus-all discriminative model (a regularized least-square classifier or a support vector machine) for each word, based on which the audio frames were classified. Turnbull *et al.* [11] built a generative model for each semantic word, and calculated a multinomial distribution over the word vocabulary for each song. Eck *et al.* [3] used AdaBoost to predict the strength of the occurrence of each social tag (a word) on a large audio data set.

The methods abovementioned achieve good results by modeling the text-audio relations, however, they share two drawbacks. First, there lacks of a taxonomy to organize the semantic labels. Music has a number of important aspects affecting music perception and music similarity, such as genre, instrumentation, emotion, and tempo, etc. The classification or detection of these aspects, were also investigated in many previous methods [12, 4, 8, 9]. Semantic labels used for music annotation can be also naturally divided into groups corresponding to these aspects. However, although most of the annotation methods use a rather large semantic vocabulary that covers almost all the aspects, the words are not structurally organized. One consequence is that they cannot make sure that a song is annotated from all the aspects. For example, Turnbull *et al.* [11] calculated the posterior probability for each word in the vocabulary given a song, and selected the A (a constant) words with the largest posterior probability to annotate the song. Thus, suppose for some songs, the posterior probabilities of some words describing genre are larger than those of all the words describing instrumentation, this may cause the words describing instrumentation being absent in the annotation.

Second, in the previous methods, semantic labels are modeled individually, that is, the methods only build text-audio

This wok was performed when the first author was a visiting student in Microsoft Research Asia.

relations, but ignore the text-text relations between two labels. However, some labels do have strong correlations, and this information can be investigated to improve annotation schemes. For example, “hard rock” and “electronic guitar” tend to co-occur in the annotations of a song, while “happy” and “minor key”, “fast tempo” and “gloomy” tend rarely to co-occur. Using the text-text relation information, strong evidence for the occurrence of “hard rock” may help to predict the presence of “electronic guitar”. On the other hand, conflicts in the annotated labels such as the co-occurrence of “fast tempo” and “gloomy”, which may happen using the individually annotation methods, could be mostly avoided by employing the text-text correlation.

To address the two issues above, this paper divides the semantic vocabulary into a number of categories, and proposes two collective annotation methods by exploiting the correlations within label pairs. Specifically, 50 web-parsed semantic labels are used to form the vocabulary, and are divided into 10 categories, each of which describes an aspect of music attributes, including genre, instrumentation, texture, vocal, arousal, affectivity, tempo, rhythm, tonality and production process. We also pose the restriction that the obtained annotation should contain labels from all the categories. In order to estimate the text-text relations, the normalized mutual information (NormMI) between the labels in the vocabulary is calculated. The label pairs whose NormMI values are larger than a threshold are selected to be modeled. Two methods are then exploited to integrate correlation modeling: one is a generative method, in which each selected label pair is modeled by a Gaussian Mixture Model (GMM); the other is a discriminative method, which is based on Conditional Random Field (CRF).

The rest of the paper is organized as follows: Section 2 describes the semantic vocabulary and the selection process of important word pairs. Section 3 describes audio feature extraction. The two proposed annotation methods are presented in Section 4. Section 5 presents the experimental results, and Section 6 concludes this paper.

2 SEMANTIC VOCABULARY

A vocabulary lists all the labels that can be used for semantic annotation. Currently there is not a standard vocabulary, and most researchers build their own vocabularies. Cano and Koppenberger [2] used the taxonomy provided by WordNet [1]. Whitman and Rifkin [14] extracted about 700 words from web documents associated with artists. Turnbull *et al.* [11] extracted 135 musically relevant words spanning six semantic categories from song reviews. These vocabularies are usually large enough to cover all the aspects of music. However, due to the large vocabulary, it is usually hard to avoid preference over words when acquiring the ground truth annotations.

Category	Words	Num
Genre	Blues, Country, Electronica, Folk, Funk, Gospel, HardRock, Jazz, Pop, Punk, Rap, R&b, Rock-roll, SoftRock	1-2
Instrument	Acoustic guitar, Acoustic piano, Bass, Drum, Electric guitar, Electric piano, Harmonica, Horn, Organ, Percussion, Sax, String	1-5
Texture	Acoustic, Electric, Synthetic	1-2
Vocal	Group, Male, Female, No	1-2
Affective	Positive, Neutral, Negative	1
Arousal	Strong, Middle, Weak	1
Rhythm	Strong, Middle, Weak	1
Tempo	Fast, Moderato, Slow	1
Tonality	Major, Mixed, Minor	1
Production	Studio, Live	1

Table 1. The vocabulary contains 50 quantized labels spanning 10 semantic categories. Each song is annotated using labels from all the categories with a number limitation.

In this paper, we build a simplified (but still general) vocabulary from a list of web-parsed musically relevant words. 50 commonly used labels are manually selected and quantized, covering 10 semantic categories (aspects) to describe characteristics of music signals. Table 1 lists the vocabulary. Using this vocabulary, each song will be annotated by labels from each category with label number limitations. This solves the problem that the annotations of a music signal are missing in some music aspects when the vocabulary is not organized as categories. It is noticed that multiple labels can be selected from the categories of genre, instrument, texture and vocal, while the labels within the other categories are exclusive with each other.

The same as existing methods, each label can be viewed as a binary variable in modeling and annotation. As mentioned previously, some labels have strong relations: *positive* or *negative* correlations. For example, the labels within some categories (e.g. Rhythm and Tempo) have negative correlations and are exclusive with each other. Moreover, some labels from different categories may have positive or negative correlations. For example, “Genre.HardRock” and “Arousal.Strong” tend to co-occur, while “Tonality.Major” and “Affective.Negative” tend rarely to co-occur.

The normalized mutual information (NormMI) is used to measure the correlations of each label pair (X, Y) as

$$\text{NormMI}(X, Y) = \frac{I(X, Y)}{\min\{H(X), H(Y)\}} \qquad (1)$$

[1] http://wordnet.princeton.edu/

Word pair	NormMI
(Production.Live, Production.Studio)	1.00
(Vocal.Female, Vocal.Male)	0.79
(Tonality.Major, Tonality.Minor)	0.69
(Tempo.Fast, Tempo.Moderato)	0.62
(Rhythm.Middle, Rhythm.Strong)	0.56
(Genre.Electronica, Texture.Synthetic)	0.25
(Arousal.Weak, Rhythm.Weak)	0.24
(Instrument.AcousticGuitar, Texture.Acoustic)	0.23
(Instrument.Drum, Rhythm.Weak)	0.23
(Genre.HardRock, Instrument.ElectricGuitar)	0.19

Table 2. Selected word pairs and their normalized mutual information. The label pairs in the first five rows are from the same semantic category, and those in the last five rows are from different categories.

where $I(X, Y)$ is the mutual information between X and Y

$$I(X,Y) = \sum_{x\in\{+1,-1\}} \sum_{y\in\{+1,-1\}} P(x,y) \log \frac{P(x,y)}{P_X(x)P_Y(y)} \tag{2}$$

and $H(x)$ is the entropy of label X defined by

$$H(X) = -\sum_{x\in\{+1,-1\}} P_X(x) \log P_X(x) \tag{3}$$

Here $+1$ and -1 represents the presence and absence of a label, respectively. The probability $P_X(x)$ and $P_Y(y)$, and $P(X, Y)$ can be estimated from a training set.

NormMI(X, Y) has the following properties:

1. $0 \leq$ NormMI$(X, Y) \leq 1$;
2. NormMI$(X, Y) = 0$ when X and Y is statistically independent;
3. NormMI$(X, X) = 1$.

NormMI considers label correlations only, and is irrelevant to the distributions of individual labels. The larger NormMI is, the stronger the correlation is. In our approach, only the label pairs whose NormMI values are larger than a threshold are selected to be modeled (see in Section 4). Table 2 lists some of the selected pairs and their NormMI values.

3 AUDIO FEATURE EXTRACTION

Each song is divided into frames with 20ms length and 10ms overlap. Tempo and beats are detected, then the song is divided into beat segments. Each segment contains a number of successive frames. A bag of beat-level feature vectors are used to represent a song. Each vector contains two sets of features: timbre features and rhythm features. Beat-level timbre features are the mean and standard deviation of the timbre features extracted in each frame. Rhythm features are extracted from the beat segments.

The reason to use beat-level features is that they are much more compact than the frame-level features to represent a song, and hence result in much lower computational complexity. Besides, the beat-level features cover a long period information, and may represent some high-level music characteristics so that it may be helpful for the annotation task.

3.1 Timbre Features

For each audio frame, three classes of spectral features are calculated. They are 8-order Mel-frequency cepstral coefficients (MFCCs), spectral shape features and spectral contrast features. The spectral shape features, including brightness, bandwidth, rolloff, and spectral flux, are commonly used in genre classification [12]. The spectral contrast features which was originally proposed in [6], are designed to be a complement of MFCCs on the sub-band information, and are shown successful in mood classification [8]. These three classes of features constitute a 47-dimensional vector. Finally, the mean and standard deviation of the frame-level timbre features in each beat segment compose the beat-level timbre feature vector, which is a 94-dimensional vector.

3.2 Rhythm Features

Rhythm is an important aspect of music. In our approach, a 20-second window (with current beat-segment in the middle) is used for rhythm feature extraction. Following Lu *et al.* [8], eight rhythm features are extracted, including average tempo, average onset frequency, rhythm regularity, rhythm contrast, rhythm strength, average drum frequency, drum amplitude and drum confidence.

In the end, a 102-dimensional (timbre plus rhythm) beat-level feature vector for each beat segment is extracted. Then the vectors are normalized to zero mean and unit variance along each dimension. Principle Component Analysis (PCA) is further employed to reduce the dimensionality of the feature vectors to 65, reserving 95% energy.

4 SEMANTIC ANNOTATION

Given a vocabulary $\mathcal{V}$ consisting of $|\mathcal{V}|$ labels (or words) $w_i \in \mathcal{V}$, and a song s represented by a bag of T real-valued feature vectors $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$, the goal of semantic annotation is to find a set $\mathcal{W} = \{w_1, \cdots, w_A\}$ of A words describing the song. It is convenient to represent the set $\mathcal{W}$ as an annotation vector $\mathbf{y} = (y_1, ..., y_{|\mathcal{V}|})$. Here y_i is a binary variable, valued 1 or -1 to represent "presence" or "absence" of label w_i. Therefore, a data set $\mathcal{D}$ is a collection of song-annotation pairs $\mathcal{D} = \{(\mathcal{X}_1, \mathbf{y}_1), \cdots, (\mathcal{X}_{|\mathcal{D}|}, \mathbf{y}_{|\mathcal{D}|})\}$.

In general, this annotation problem can be addressed by Maximum A Posterior (MAP), that is, to choose an annotation vector with maximum posterior: $\hat{\mathbf{y}} = \arg\max P(\mathbf{y}|\mathcal{X})$ [11]. In existing methods, the labels are treated independent, so that the posterior probability of the annotation vector can be decomposed into the multiplication of the posterior probability of each label as

$$P(\mathbf{y}|\mathcal{X}) = \prod_{i}^{|\mathcal{V}|} P(y_i|\mathcal{X}) \propto \prod_{i}^{|\mathcal{V}|} p(\mathcal{X}|y_i)P(y_i) \quad (4)$$

If further assume that feature vectors $\mathbf{x}_1, \cdots, \mathbf{x}_T$ in the bag $\mathcal{X}$ are independent, then

$$p(\mathcal{X}|y_i) = \prod_{t}^{T} p(\mathbf{x}_t|y_i) \quad (5)$$

where T is the number of feature vectors in the bag $\mathcal{X}$. The likelihood $p(\mathbf{x}_t|y_i)$ can be estimated using a parametric model such as a GMM from the training data. The prior probability $P(y_i)$ can also be estimated or as usual set to a uniform distribution.

However, as mentioned above, the labels are not independent, and we need to consider their correlations. In the following subsections, two approaches are exploited for correlation modeling: a GMM-based method, and a Conditional Random Field (CRF)-based method.

4.1 The GMM-based method

When the labels are not independent, the joint posterior probability $P(\mathbf{y}|\mathcal{X})$ cannot be decomposed into single label posteriors. Instead, we approximate it using the multiplication of single label posteriors and label-pair posteriors.

$$\begin{aligned} P(\mathbf{y}|\mathcal{X}) &\sim \prod_{i}^{|\mathcal{V}|} P(y_i|\mathcal{X}) \left(\prod_{j}^{|\mathcal{E}|} P(y_{e_j^1}, y_{e_j^2}|\mathcal{X}) \right)^{\alpha} && (6) \\ &\propto \prod_{i}^{|\mathcal{V}|} p(\mathcal{X}|y_i)P(y_i) \\ &\quad \left(\prod_{j}^{|\mathcal{E}|} p(\mathcal{X}|y_{e_j^1}, y_{e_j^2}) P(y_{e_j^1}, y_{e_j^2}) \right)^{\alpha} && (7) \end{aligned}$$

where $\mathcal{E}$ is the set of the selected label pairs that have large normalized mutual information; e_j^1 and e_j^2 are the two labels in the pair; α is a the trade off between label posteriors and label pair posteriors. In our experiments it is set to 1 typically.

The likelihood $p(\mathcal{X}|y_i)$ and $p(\mathcal{X}|y_{e_j^1}, y_{e_j^2})$ can be computed based on Eq.(5), assuming the feature vectors within a bag are independent. The likelihood of each feature vector is estimated using a GMM model, where 8 kernels are arbitrarily selected in this paper.

Although the right part of Eq.(6) is not the exact decomposition of $P(\mathbf{y}|\mathcal{X})$, it represents the intuitive idea that the annotation should not only maximize the posterior probabilities of individual words, but should also consider the posteriors of the correlated label pairs.

4.2 The CRF-based method

Conditional Random Field (CRF) was firstly proposed by Lafferty *et al.* [7] to segment and label sequence data such as natural language. It is an undirected graphical model, where the nodes represent the label variables and the edges represent the relations between labels. Compared with Hidden Markov Model (HMM), one advantage of the chain CRF model is that it relaxes the strong independence assumptions. In fact, a general CRF can naturally model arbitrary dependencies between features and labels. Further, Ghamrawi and McCallum [5] proposed two multi-label classification models based on CRF, which directly parameterized correlations among labels and features.

Generally, given a sample $\mathbf{x}$ and its output label vector $\mathbf{y}$, using the multi-label classification CRF models, the posterior probability $p(\mathbf{y}|\mathbf{x})$ can be written as

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_k \lambda_k f_k(\mathbf{x}, \mathbf{y}) + \sum_l \mu_l g_l(\mathbf{x}, \mathbf{y}) \right) \quad (8)$$

where $Z(\mathbf{x})$ is the normalizing factor. $f_k(\mathbf{x}, \mathbf{y})$ and $g_l(\mathbf{x}, \mathbf{y})$, are two predefined real-valued functions, corresponding to a node and an edge respectively. They are usually referred to as *features* of the CRF. In principle, any real-valued function of sample $\mathbf{x}$ and label $\mathbf{y}$ can be treated as a feature. For example, it can be the frequency of a phrase in a text document, or one dimension of the beat-level feature in a music signal. λ_k and μ_l are the parameters to be estimated to maximize Eq. (8) using training data, where k and l enumerate the following indexes of features,

$$k \in < r_i, y_j >: 1 \leq i \leq |R|, 1 \leq j \leq |\mathcal{Y}| \quad (9)$$

$$l \in < r_i, y_j, y_{j'} >: 1 \leq i \leq |R|, 1 \leq j, j' \leq |\mathcal{Y}| \quad (10)$$

where R is a set of music characteristics (we do not call them features in order to avoid confusion with CRF features like f_k and g_l above), and $r_i \in R$; $|\mathcal{Y}|$ is the length of the label vector, and y_j is a label variable. It can be seen that each feature f_k corresponds to a pair consisting of a label and a characteristic, and each feature g_l corresponds to a triplet consisting of a label pair and a characteristic.

Note that k in Eq.(9) enumerates all the nodes, and l in Eq.(10) enumerates all the edges. Therefore, Eq. (8) corresponds to a full connected graph, where $\sum_k \lambda_k f_k(\mathbf{x}, \mathbf{y})$ represents the overall potential of nodes, and $\sum_l \mu_l g_l(\mathbf{x}, \mathbf{y})$ represents the overall potential of edges. However, in practice, not all the label pairs have close relations, and we only

%	Individual GMM	Collective GMM	Individual CRF	Collective CRF
Overall	60.7 / **61.0** / 60.8	61.2 / **61.0** / 61.1	68.0 / 60.5 / 64.0	**68.4 / 61.0 / 64.5**
Genre	43.4 / **50.9** / 46.9	44.5 / 50.2 / 47.2	54.2 / 40.9 / 46.6	**55.4** / 41.8 / **47.7**
Instrument	54.3 / 53.5 / 53.9	54.9 / **53.8** / 54.3	72.8 / 48.4 / 58.1	**72.9** / 48.6 / **58.3**
Texture	73.2 / **71.8** / 72.5	74.0 / 71.7 / 72.8	75.1 / 71.0 / 73.0	**75.2** / 71.0 / **73.1**
Vocal	76.5 / 71.3 / 73.8	76.7 / 71.2 / 73.8	**80.6** / 84.3 / 82.4	**80.6 / 85.4 / 82.9**
Affective	46.8	**47.5**	42.1	43.1
Arousal	56.5	56.6	57.0	**58.5**
Rhythm	63.0	62.3	64.0	**64.5**
Tempo	59.2	59.4	**63.3**	63.0
Tonality	56.9	57.4	**60.4**	60.1
Production	93.0	93.0	94.5	**94.6**

Table 3. Average per-category performance in the whole vocabulary and each semantic category. For each item, the three numbers are arranged in the format "Precision / Recall / F-measure". Note that for the lower 6 semantic categories, precision, recall and F-measure are the same, and hence written as one number.

consider those with strong correlations. In this case, the edges of the graph are sparse. Suppose the number of edges is E, and all the label variables have C possible values (in our case, C is 2, representing the presence and absence of each label.), then the number of parameters to be estimated in total is $|R||Y|C + |R|EC^2$.

In Eq.(10), the potential of edges are feature-dependent. It can also be degenerated to feature-independent as

$$l \in < y_j, y_{j'} >: 1 \leq j, j' \leq |\mathcal{Y}| \qquad (11)$$

In this case, the number of parameters to be estimated in total is $|R||Y|C + EC^2$.

In order to reduce the computational complexity, we adopt the degenerated CRF model, where the edges are feature-independent, as in Eq.(11). Besides, each song is treated as a sample, and a 115-dimensional song-level feature vector is calculated and set as the characteristic set R. It consists of two parts: a 65-dimensional vector, which is the mean of the beat-level features, and a 50-dimensional vector, with each dimension representing the likelihood of the song given an semantic label in vocabulary, calculated using Eq.(5). The 50-dimensional vector can also be seen as a song representation in an anchor space [1], where each anchor represents one of the 50 semantic labels in the vocabulary.

5 EXPERIMENTS

In this section, the two proposed collective annotation methods are evaluated and compared with two individual annotation methods: the individual GMM-based method with Eq.(4), and the individual CRF-based method which does not consider the edges (or label pairs) in the graph. The experimental data set consists of 4,951 Western popular songs, each of which was manually annotated with semantic labels from the vocabulary in Table 1, as described in Section 2. 25% of the songs are randomly selected as the training set and the left as the test set. For the collective methods, 49 label pairs in total, whose NormMI values in the training set are larger than 0.1, are selected to be modeled.

The annotation performance is measured from two aspects: the *category* aspect and the *song* aspect. From the category aspect, the annotation performances in different categories are measured. From the song aspect, the average annotation performance of each song is evaluated. For both aspects, the average *precision*, *recall* and *F-measure* are used as evaluation metrics, where F-measure is defined as

$$\text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{(\text{precision} + \text{recall})} \qquad (12)$$

Table 3 lists the average per-category performances of the four methods. It can be seen that the CRF-based methods outperform the GMM-based methods generally, which accords with previous experiences that discriminative methods generally outperform generative methods in classification problems. Besides, the collective annotation methods slightly but consistently improve the performance, compared with their individual counterpart, both for GMM-based methods and CRF-based methods. This indicates that the label pair modeling helps annotatoin in some cases.

Table 4 presents the performances of song annotations, comparing with four methods. It can be seen that, while the recalls are similar for all the methods, the precision is improved significantly from the generative models to discriminative models. Besides, the collective methods slightly outperform their individual counterparts, which are consistent with the observations made above.

However, the performance improvements from individual modeling to collective modeling is not so much. Although the level of improvement accords with the experiments in [5], we still need to further discover reasons and

%	Precision	Recall	F-measure
Individual GMM	60.9	61.4	60.9
Collective GMM	61.4	61.4	61.2
Individual CRF	68.1	60.9	64.0
Collective CRF	**68.5**	**61.3**	**64.4**

Table 4. Performance of song annotation, comparing with four methods.

exploit solutions. One possible reason may be that, in individual modeling methods, although each label is modeled individually, the labels which are "correlated" share many songs in their training set (since each song has multiple labels). This makes the trained models of "correlated" labels are also "correlated", or in other words, the correlation is implicitly modeled.

6 CONCLUSION

In this paper, we presents our attempts to collective annotation of music signals, which not only models individual semantic labels , but also their correlations. In our approach, 50 musically relevant labels are manually selected for music annotation, covering 10 semantic aspects of music perception. Then, normalized mutual information is employed to measure the correlation between two semantic labels, and those label pairs with strong correlation are selected and modeled in two methods, one based on GMM, and the other based on CRF. Experimental results show slight but consistent improvement compared with individual label modeling methods.

There is still considerable room to improve the proposed approach. First, we need further exploit better methods to model label correlation in order to get higher performance improvement. Second, we need also exploit better features. In current approach, only a song-level feature vector is used for CRF-based methods. How to choose effective song-level features or to adapt the bag of features to CRF-based methods is still a challenging task. Finally, we will also try to apply the obtained annotations in various applications, such as music similarity measure, music search or music recommendation. We are also like to check the impact of annotation accuracy in these applications.

7 REFERENCES

[1] Berenzweig, A., Ellis, D.P.W. and Lawrence, S. "Anchor space for classification and similarity measurement of music," in *Proc. IEEE International Conference on Multimedian and Expo (ICME)*, 2003, pp. I-29–32.

[2] Cano, P. and Koppenberger, M. "Automatic sound annotation," in *Proc. IEEE Workshop Mach. Learn. Signal Process.*, 2004, pp. 391-400.

[3] Eck, D., Lamere, P., Bertin-Mahieux, T. and Green, S., "Automatic generation of social tags for music recommendation, " in *Proc. Neural Informaiton Processing Systems (NIPS)*, 2007.

[4] Essid, S., Richard, G. and David, B. "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 14, no. 1, 2006.

[5] Ghamrawi, N. and McCallum, A. "Collective multilabel classification," in *Proc. the 14th ACM International Conference on Information and Knowledge Management (CIKM)*, 2005, pp. 195-200.

[6] Jiang, D.N., Lu, L., Zhang, H.J., Tao, J.H. and Cai, L.H., "Music type classification by spectral contrast features," in *Proc. IEEE International Conference on Multimedian and Expo (ICME)*, vol. 1, 2002, pp. 113-116.

[7] Lafferty, J., McCallum, A. and Pereira, F. "Conditional random fields: Probabilistic models for segmenting and labeling seqeunce data," in *Proc. the Eighteenth International Conference on Machine Learning (ICML)*, 2001, pp. 282-289.

[8] Lu, L., Liu, D. and Zhang, H.J. "Automatic mood detection and tracking of music audio signals", *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 14, no. 1, pp. 5-18, 2006.

[9] Peeters, G. "Time variable tempo detection and beat marking," in *Proc. International Computer Music Conference (ICMC)*, 2005.

[10] Slaney, M. "Mixtures of probability experts for audio retrieval and indexing," in *Proc. IEEE International Conference on Multimedian and Expo (ICME)*, 2002, pp. 345-348.

[11] Turnbull, D., Barrington, L., Torres, D. and Lanckriet, G. "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 16, no. 2, pp. 467-476, 2008.

[12] Tzanetakis, G. and Cook, P. "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Process.*, vol. 10, no. 5, pp. 293-302, 2002.

[13] Whitman, B. and Ellis, D.P.W. "Automatic record reviews," in *Proc. ISMIR*, 2004, pp. II-1002 - II-1009.

[14] Whitman, B. and Rifkin, R. "Musical query-by-description as a multiclass learning problem," in *IEEE Workshop Multimedia Signal Process.*, 2002, pp. 153-156.