SPARSE MULTI-LABEL LINEAR EMBEDDING NONNEGATIVE TENSOR FACTORIZATION FOR AUTOMATIC MUSIC TAGGING

Yannis Panagakis[*], *Constantine Kotropoulos*[*], *Gonzalo R. Arce*[†]

[*] Department of Informatics
Aristotle University of Thessaloniki
Box 451, Thessaloniki 54124, GREECE
email: {panagakis,costas}@aiia.csd.auth.gr

[†] Department of Electrical & Computer Engineering
University of Delaware
Newark, DE 19716-3130, U.S.A.
email: arce@ece.udel.edu

ABSTRACT

In this paper, a robust framework for automatic music tagging is proposed. First, each music recording is represented by its auditory temporal modulations. Then, a multilinear subspace learning algorithm based on sparse label coding is proposed to effectively harness the multi-label information for dimensionality reduction. The proposed algorithm is referred to as *Sparse Multi-label Linear Embedding Nonnegative Tensor Factorization*. Finally, a recently proposed sparse representation-based method for multi-label data is employed to propagate the multiple labels of the training auditory temporal modulations to annotate the auditory temporal modulations extracted from a test music recording with the sparse ℓ_1 reconstruction coefficients. The proposed framework outperforms both humans and state-of-the-art computer audition systems in the music tagging task, when applied to the CAL500 dataset.

1. INTRODUCTION

The emergence of Web 2.0 and the success of music-oriented social network websites, such as *last.fm*, has revealed the concept of music tagging. Tags are text-based labels that encode semantic information related to music (i.e., instrumentation, genres, emotions, etc.) resulting into a non-acoustic representation of music, which can be used as input to collaborative filtering systems assisting users to search for music content. However, a drawback of these systems is that a newly added music recording must be tagged manually first, before it can be retrieved [18, 19], which is a time consuming, expensive process. Therefore, an interesting problem in Music Information Retrieval (MIR) community is how to automate the process of tagging music recordings when they become available. This problem is referred to as *automatic music tagging* or *automatic multi-label music annotation*.

MIR has mainly focused on content-based classification of music by genre [11, 12, 13], and emotion [14]. Such classification systems effectively annotate music with class labels, such as "rock", "happy", etc by assuming a predefined taxonomy and explicit labeling of a music recording into mutually exclusive classes. However, this assumption is unrealistic and results into a number of problems since music perception is inherently subjective [19]. These problems can be overcome by the less restrictive approach of annotating the audio content by more than one labels, which reflect more aspects of music. However, has been made little work on multi-label automatic music annotation compared to that on the multi-label automatic image annotation (refer to [2, 20] and the references therein). Automatic music tagging algorithms can be roughly classified into three categories: 1) classification-based methods, 2) probabilistic modeling-based methods, and 3) web game related methods. The classification-based methods treat audio tag prediction as a set of binary classification problems where standard classifiers such as the Support Vector Machines [17] or AdaBoost [1] can be applied. The probabilistic modeling-based methods [19, 5] attempt to infer the correlations or joint probabilities between the tags and the low-level acoustic features extracted from audio. Web game related methods try to solve the music tagging problem via games [7].

In this paper, the problem of automatic music tagging is addressed as multi-label multi-class classification problem by employing a novel multilinear subspace learning algorithm and sparse representations. Motivated by the robustness of the auditory representations in the music genre classification [11, 12, 13], each audio recording is represented in terms of its slow temporal modulations by a two-dimensional (2D) auditory representation as in [13]. Consequently, an ensemble of audio recordings is represented by a third-order tensor. The auditory temporal modulations do not explicitly utilize the label set (i.e., the tags) of music recordings. Due to the well-known semantic gap, it is unclear how the semantic similarity between the label sets associated to two music recordings can drive the efficient feature extraction. Based on the automatic multi-label image annotation framework proposed in [20], the semantic similarities between two music recordings with overlapped labels are measured in a sparse representation-based way rather than in one-to-one way as in [17, 1]. To this end, a novel multilinear subspace learning algorithm is developed to efficiently harness the multi-label information for feature extraction. In particular, the proposed method incorporates the Multi-label Linear Embedding (MLE) [20] into the Nonnegative Tensor Factorization (NTF) [11]. It is referred to as *Sparse Multi-label Linear Embedding Nonnegative Tensor Factorization* (SMLENTF). The SMLENTF is adopted in order to reduce the dimensionality of the space, where the high-order data (i.e. auditory temporal modulations representations) lie, by mapping the high-order data onto a lower-dimensional semantic space dominated by the label information. Features extracted by the SMLENTF form an overcomplete dictionary for the semantic space of music. If sufficient training music recordings are available, it is possible to express any test representation of auditory temporal modulations as a compact linear combination of the dictionary atoms, which are semantically close. This representation is designed to be sparse, because it involves only a small fraction of the dictionary atoms and can

 ISSN 2076-1465

be computed efficiently via ℓ_1 optimization. Finally, tags are propagated from the training atoms to a test music recording with the sparse ℓ_1 representation coefficients.

The performance of the proposed automatic music tagging framework is assessed by conducting experiments on the CAL500 dataset [18, 19]. For comparison purposes, the MLE [20] is also tested in this task. The reported experimental results indicate the superiority of the proposed SMLENTF over the MLE, the human performance, as well as that of state-of-the-art computer audition systems in music tagging, on the same dataset.

The paper is organized as follows. In Section 2, basic multilinear algebra concepts and notations are defined. In Section 3, the bio-inspired auditory representation based on a computational auditory model is briefly described. SMLENTF is introduced in Section 4. The sparse representations based multi-label annotation framework is detailed in Section 5. Experimental results are demonstrated in Section 6 and conclusions are drawn in Section 7.

2. NOTATION AND MULTILINEAR ALGEBRA BASICS

Tensors are considered as the multidimensional equivalent of matrices (i.e., second-order tensors) and vectors (i.e., first-order tensors) [6]. Throughout the paper, tensors are denoted by boldface Euler script calligraphic letters (e.g. $\mathcal{X}$, $\mathcal{A}$), matrices are denoted by uppercase boldface letters (e.g. $\mathbf{U}$), vectors are denoted by lowercase boldface letters (e.g. $\mathbf{u}$), and scalars are denoted by lowercase letters (e.g. u). The ith row of $\mathbf{U}$ is denoted as $\mathbf{u}_{i:}$ while its jth column is denoted as $\mathbf{u}_{:j}$.

Let $\mathbb{Z}$ and $\mathbb{R}$ denote the set of integer and real numbers, respectively. A high-order real valued tensor $\mathcal{X}$ of order N is defined over the tensor space $\mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, where $I_n \in \mathbb{Z}$ and $n = 1, 2, \ldots, N$. Each element of $\mathcal{X}$ is addressed by N indices, i.e., $x_{i_1 i_2 \ldots i_N}$. Mode-n unfolding of tensor $\mathcal{X}$ yields the matrix $\mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times (I_1 \ \ldots I_{n-1} I_{n+1} \ldots I_N)}$. In the following, the operations on tensors are expressed in matricized form [6].

An N-order tensor $\mathcal{X}$ has rank-1, when it is decomposed as the outer product of N vectors $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(N)}$, i.e. $\mathcal{X} = \mathbf{u}^{(1)} \circ \mathbf{u}^{(2)} \circ \ldots \circ \mathbf{u}^{(N)}$. That is, each element of the tensor is the product of the corresponding vector elements, $x_{i_1 i_2 \ldots i_N} = u^{(1)}_{i_1} u^{(2)}_{i_2} \ldots u^{(N)}_{i_N}$ for $i_n = 1, 2, \ldots, I_n$. The rank of an arbitrary N-order tensor $\mathcal{X}$ is the minimal number of rank-1 tensors that yield $\mathcal{X}$ when linearly combined. Next, several products between matrices will be used, such as the Khatri-Rao product denoted by $\odot$, and the Hadamard product denoted by $*$, whose definitions can be found in [6] for example.

3. AUDITORY TEMPORAL MODULATIONS REPRESENTATION

A key step for representing music signals in a psychophysiologically consistent manner is to resort on how audio is encoded in the human *primary auditory cortex*. The primary auditory cortex is the first stage of the central auditory system, where higher level mental processes take place, such as perception and cognition [10]. To this end the *representation of auditory temporal modulations* for audio signals is employed [13]. The auditory representation is a joint acoustic and modulation frequency representation [15], that discards much of the spectro-temporal details and focuses on the underlying slow temporal modulations of the music signal. Such a representation has been proven very robust in representing music signals for music genre classification [12, 13].

The 2D representation of auditory temporal modulations can be obtained by modeling the path of auditory processing as detailed in [13]. The computational model of human auditory system consists of two basic processing stages. The first stage models the early auditory system, which converts the acoustic signal into an auditory representation, the so-called *auditory spectrogram*, i.e. a time-frequency distribution along a tonotopic (logarithmic frequency) axis. At the second stage, the temporal modulation content of the auditory spectrogram is estimated by wavelets applied to each row of the auditory spectrogram. Psychophysiological evidence justifies the choice of discrete rate $r \in \{2, 4, 8, 16, 32, 64, 128, 256\}$ (Hz) to represent the temporal modulation content of sound. The cochlear model employed in the first stage, has 96 filters covering 4 octaves along the tonotopic axis (i.e. 24 filters per octave). Accordingly, the auditory temporal modulation of a music recording is represented by a real-valued nonnegative second-order tensor (i.e. a matrix) $\mathbf{X} \in \mathbb{R}_+^{I_1 \times I_2}$, where $I_1 = I_f = 96$ and $I_2 = I_r = 8$. Hereafter, let $\mathbf{x} = \mathrm{vec}(\mathbf{X}) \in \mathbb{R}_+^{I_1 \cdot I_2} = \mathbb{R}_+^{768}$ denote the lexicographically ordered vectorial representation of the auditory temporal modulations.

4. SPARSE MULTI-LABEL LINEAR EMBEDDING NONNEGATIVE TENSOR FACTORIZATION

In order to transform the high-dimensional original tensor space into a lower-dimensional semantic space defined by label information, multilinear subspace learning algorithms are required. In conventional multilinear subspace learning algorithms, such as the General Tensor Discrimininant Analysis [16], the assumption made is that data points annotated by the same label should be close to each other where data bearing different labels should to be far away in the feature space. However, this assumption is not valid in a multi-label task as discussed in [20] and such subspace learning algorithms will fail to produce a lower-dimensional semantic space based on multiple labels.

Let $\{\mathcal{X}_q|_{q=1}^{Q}\}$ be a set of Q nonnegative tensors $\mathcal{X}_q \in \mathbb{R}_+^{I_1 \times I_2 \times \ldots \times I_N}$ of order N. We can represent such a set by a $(N+1)$-order tensor $\mathcal{A} \in \mathbb{R}_+^{I_1 \times I_2 \times \ldots \times I_N \times I_{N+1}}$ with $I_{N+1} = Q$. Furthermore, let us assume that the multi-labels of the training tensor $\mathcal{A}$ are represented by the matrix $\mathbf{C} \in \mathbb{R}_+^{V \times Q}$, where V indicates the cardinality of the tag vocabulary. Accordingly, $c_{ji} = 1$ if the ith tensor is labeled with the jth tag in the vocabulary and 0 otherwise. Since, every tensor object (music recording in this paper) can be labeled by multiple labels, there may exist more than one non-zero elements in a label vector (i.e. $\mathbf{c}_{:i}$).

To overcome the limitation of conventional multilinear subspace learning algorithms, the MLE [20] is incorporated into the NTF. To this end two methods for using multi-label information in order to drive semantically oriented feature extraction from tensor objects are adopted. First, the tensor objects with the same label set, that is $\mathbf{c}_{:i} = \mathbf{c}_{:j}$, are considered to be fully semantically related and thus the similarity graph $\mathbf{W}^1$ has elements $w^1_{ij} = w^1_{ji} = 1$ and 0 otherwise. However, in real-world datasets, data samples with exactly the same label set are rare, especially in music en-

sembles. In such a case, the semantic relationship between data samples can be inferred via the ℓ_1 semantic graph as proposed in [20]. Let us denote by $\mathbf{W}^2$ the ℓ_1 semantic graph. $\mathbf{W}^2$ contains the coefficients that represent each label vector $\mathbf{c}_{:i}$ as a compact linear combination of the remaining semantically related label vectors. Formally, let us define $\hat{\mathbf{C}}_i = [\mathbf{c}_{:1}|\mathbf{c}_{:2}|\ldots|\mathbf{c}_{:i-1}|\mathbf{c}_{:i+1}|\ldots|\mathbf{c}_{:Q}]$. If $V \ll Q$ the linear combination coefficients $\mathbf{a}$ can be obtained by seeking the sparsest solution to the undetermined system of equations $\mathbf{c}_{:i} = \hat{\mathbf{C}}_i\mathbf{a}$. That is, by solving the following optimization problem:

$$\arg\min_{\mathbf{a}} \|\mathbf{a}\|_0 \quad \text{subject to } \hat{\mathbf{C}}_i\mathbf{a} = \mathbf{c}_{:i}, \tag{1}$$

where $\|.\|_0$ is the ℓ_0 quasi-norm returning the number of the non-zero entries of a vector. Finding the solution to optimization problem (1) is NP-hard due to the nature of the underlying combinational optimization. In [4], it has been proved that if the solution is sparse enough, then the solution of (1) is equivalent to the solution of the optimization problem:

$$\arg\min_{\mathbf{a}} \|\mathbf{a}\|_1 \quad \text{subject to } \hat{\mathbf{C}}_i\,\mathbf{a} = \mathbf{c}_{:i}, \tag{2}$$

where $\|.\|_1$ denotes the ℓ_1 norm of a vector. (2) can be solved in polynomial time by standard linear programming methods [3].

Therefore, the ℓ_1 semantic graph $\mathbf{W}^2$ is constructed as follows. For each label vector, $\hat{\mathbf{C}}_i$ is constructed and then it is normalized so as to have unit length column vectors. Then, (2) is solved, by replacing $\hat{\mathbf{C}}_i$ with the normalized one, and the sparse representation vector $\mathbf{a}$ is obtained. Next, $w^2_{ij} = a_j$ for $1 \le j \le i-1$; $w^2_{ij} = a_{j-1}$ for $i+1 \le j \le Q$. Clearly, the diagonal elements of $\mathbf{W}^2$ are equal to zero.

Given $\{\mathcal{X}_q|_{q=1}^{Q}\}$, one can model the semantic relationships between these tensor objects by constructing the multi-label linear embedding matrix, using $\mathbf{W}^1$ and $\mathbf{W}^2$ as in [20]. The multi-label linear embedding matrix is defined as $\mathbf{M} = \mathbf{D}^1 - \mathbf{W}^1 + \frac{\beta}{2}(\mathbf{I}-\mathbf{W}^2)^T(\mathbf{I}-\mathbf{W}^2)$, where $\mathbf{D}^1$ is a diagonal matrix with elements $d^1_{ii} = \sum_{i\neq j} w^1_{ij}$ and $\beta > 0$ is a parameter for balancing the contribution of each graph in the multi-label linear embedding [20]. Let $\mathbf{Z}^{(n)} = \mathbf{U}^{(N+1)} \odot \ldots \odot \mathbf{U}^{(n+1)} \odot \mathbf{U}^{(n-1)} \odot \ldots \odot \mathbf{U}^{(1)}$. One can incorporate the semantic information of tensor objects into the NTF by constructing the following objective function for the SMLENTF in matricized form:

$$f_{SMLENTF}\left(\mathbf{U}^{(n)}|_{n=1}^{N+1}\right) = \|\mathbf{A}_{(n)} - \mathbf{U}^{(n)}\left[\mathbf{Z}^{(n)}\right]^T\|_F^2 + \lambda\,\mathrm{tr}\left\{\left[\mathbf{U}^{(N+1)}\right]^T\mathbf{M}\,\mathbf{U}^{(N+1)}\right\}, \tag{3}$$

where $\lambda > 0$ is a parameter, which controls the trade off between goodness of fit to the data tensor $\mathcal{A}$ and the multi-label linear embedding and $\|.\|_F$ denotes the Frobenius norm. Consequently, we propose to minimize (3) subject to the nonnegativity constraint on factor matrices $\mathbf{U}^{(n)} \in \mathbb{R}_+^{I_n\times k}$, $n = 1,2,\ldots,N+1$, where k is the desirable number of rank-1 tensors approximating $\mathcal{A}$ when linearly combined.

Let $\nabla_{\mathbf{U}^{(n)}} f_{SMLENTF} = \frac{\partial f_{SMLENTF}}{\partial \mathbf{U}^{(n)}}$ be the partial derivative of the objective function $f_{SMLENTF}(\mathbf{U}^{(n)}|_{n=1}^{N+1})$ with respect to $\mathbf{U}^{(n)}$. Now, let us define the nonnegative matrices $\mathbf{M}^+$ (with elements $m^+_{ij} = m_{ij}$ if $m_{ij} > 0$ and 0 otherwise) and $\mathbf{M}^-$ (with elements $m^-_{ij} = -m_{ij}$ if $m_{ij} < 0$ and 0 otherwise). Since $\mathbf{U}^{(n)}$, $n = 1,2,\ldots,N+1$, $\mathbf{M}^+$, and $\mathbf{M}^-$ are nonnegative, the partial derivatives of the objective function can be decomposed as differences of two nonnegative components denoted by $\nabla^+_{\mathbf{U}^{(n)}} f_{SMLENTF}$ and $\nabla^-_{\mathbf{U}^{(n)}} f_{SMLENTF}$, respectively. It can be shown that for $n = 1,2,\ldots,N$ we have

$$\nabla_{\mathbf{U}^{(n)}} f_{SMLENTF} = \underbrace{\mathbf{U}^{(n)}\left[\mathbf{Z}^{(n)}\right]^T\mathbf{Z}^{(n)}}_{\nabla^+_{\mathbf{U}^{(n)}} f_{SMLENTF}} - \underbrace{\mathbf{A}_{(n)}\mathbf{Z}^{(n)}}_{\nabla^-_{\mathbf{U}^{(n)}} f_{SMLENTF}}, \tag{4}$$

while for $n = N+1$ and since $\mathbf{M} = \mathbf{M}^+ - \mathbf{M}^-$ we obtain

$$\nabla_{\mathbf{U}^{(N+1)}} f_{SMLENTF} = \underbrace{\mathbf{U}^{(N+1)}\left[\mathbf{Z}^{(N+1)}\right]^T\mathbf{Z}^{(N+1)} + \lambda\, M^+\mathbf{U}^{(N+1)}}_{\nabla^+_{\mathbf{U}^{(N+1)}} f_{SMLENTF}} - \underbrace{\left(\mathbf{A}_{(N+1)}\mathbf{Z}^{(N+1)} + \lambda\,\mathbf{M}^-\,\mathbf{U}^{(N+1)}\right)}_{\nabla^-_{\mathbf{U}^{(N+1)}} f_{SMLENTF}}. \tag{5}$$

Following the strategy employed in the derivation of Nonnegative Matrix Factorization [8], we obtain an iterative alternating algorithm for SMLENTF as follows. Given $N+1$ randomly initialized nonnegative matrices $\mathbf{U}^{(n)}|_{n=1}^{N+1} \in \mathbb{R}_+^{I_n\times k}$, a local minimum of (3) subject to the nonnegativity constraints can be found by the multiplicative update rule:

$$\mathbf{U}^{(n)}_{[t+1]} = \mathbf{U}^{(n)}_{[t]} * \frac{\nabla^-_{\mathbf{U}^{(n)}_{[t]}} f_{SMLENTF}}{\nabla^+_{\mathbf{U}^{(n)}_{[t]}} f_{SMLENTF}}, \tag{6}$$

where the division in (6) is elementwise and t denotes the iteration index. The multiplicative update rule (6) suffers from two drawbacks: 1) The denominator may be zero; 2) $\mathbf{U}^{(n)}_{[t+1]}$ does not change when $\mathbf{U}^{(n)}_{[t]} = \mathbf{0}$ and $\nabla_{\mathbf{U}^{(n)}[t]} f_{SMLENTF} < \mathbf{0}$. In order to overcome these drawbacks, we can modify (6) as in [9]. A robust multiplicative update rule for SMLENTF is then

$$\mathbf{U}^{(n)}_{[t+1]} = \mathbf{U}^{(n)}_{[t]} - \frac{\hat{\mathbf{U}}^{(n)}_{[t]}}{\nabla^+_{\mathbf{U}^{(n)}_{[t]}} f_{SMLENTF} + \delta} * \nabla_{\mathbf{U}^{(n)}_{[t]}} f_{SMLENTF}, \tag{7}$$

where $\hat{\mathbf{U}}^{(n)}_{[t]} = \mathbf{U}^{(n)}_{[t]}$ if $\nabla_{\mathbf{U}^{(n)}_{[t]}} f_{SMLENTF} \ge \mathbf{0}$ and σ otherwise. The parameters σ, δ are predefined small positive numbers, typically 10^{-8} [9].

5. MULTI-LABEL ANNOTATION VIA SPARSE REPRESENTATIONS

In this section, the task of automatic music tagging is addressed by sparse representations of auditory temporal modulations projected onto a reduced dimension feature space, where the semantic relations between them are retained.

For each music recording the 2D auditory temporal modulations are extracted as briefly described in Section 3 and

detailed in [13]. Thus, each ensemble of recordings is represented by a third-order data tensor, which is created by stacking the second-order feature tensors associated to the recordings. Consequently, the data tensor $\mathcal{A} \in \mathbb{R}_+^{I_1 \times I_2 \times I_3}$, where $I_1 = I_f = 96$, $I_2 = I_r = 8$, and $I_3 = I_{samples}$ is obtained. Let $\mathcal{A}_{training} \in \mathbb{R}_+^{I_1 \times I_2 \times Q}$, $Q < I_{samples}$, be the tensor where the training auditory temporal modulations are stored. By applying the SMLENTF onto the $\mathcal{A}_{training}$ three factor matrices are derived, namely $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$, $\mathbf{U}^{(3)}$, associated to the frequency, rate, and samples modes of the training tensor $\mathcal{A}_{training}$, respectively. Consequently, the projection matrix $\mathbf{P} = \mathbf{U}^{(2)} \odot \mathbf{U}^{(1)} \in \mathbb{R}_+^{768 \times k}$, with $k \ll \min(768, Q)$, is obtained. The columns of $\mathbf{P}$ span a reduced dimension feature space, where the semantic relations between the vectorized auditory temporal modulation are retained. Consequently, by projecting all the training auditory temporal modulations onto this reduced dimension space an overcomplete dictionary $\mathbf{D} = \mathbf{P}^T \mathbf{A}_{training(3)}^T \in \mathbb{R}_+^{k \times Q}$ is obtained. Alternatively, the dictionary can be obtained by $\mathbf{D} = \mathbf{P}^\dagger \mathbf{A}_{training(3)}^T$, where $(.)^\dagger$ denotes the Moore-Penrose pseudoinverse

Given a vectorized representation of auditory temporal modulations $\mathbf{x} \in \mathbb{R}_+^{768}$ associated to a test music recording, first it is projected onto the reduced dimension space and a new feature vector is obtained as $\bar{\mathbf{x}} = \mathbf{P}^T \mathbf{x} \in \mathbb{R}_+^k$. Now, $\bar{\mathbf{x}}$ can be represented as a compact linear combination of the semantically related atoms of $\mathbf{D}$. That is the test representation of auditory temporal modulations is considered semantically related to a few training representation of auditory temporal modulations with non-zero approximation coefficients. This implies that the corresponding music recordings are semantically related, as well. Again, since $\mathbf{D}$ is overcomplete, the sparse coefficient vector $\mathbf{b}$ can be obtained by solving the following optimization problem:

$$\arg\min_{\mathbf{b}} \|\mathbf{b}\|_1 \quad \text{subject to } \mathbf{D}\,\mathbf{b} = \bar{\mathbf{x}}. \tag{8}$$

By applying the SMLENTF, the semantic relations between the label vectors are propagated to the feature space. In music tagging, the semantic relations are expected to propagate from the feature space to the label vector space. Let us denote by $\bar{\mathbf{a}}$ the label vector of the test music recording. Then $\bar{\mathbf{a}}$ is obtained by

$$\bar{\mathbf{a}} = \mathbf{C}\,\mathbf{b}. \tag{9}$$

The labels with the largest values in $\bar{\mathbf{a}}$ yield the final tag vector of the test music recording.

6. EXPERIMENTAL EVALUATION

In order to assess the performance of the proposed framework in automatic music tagging, experiments were conducted on the CAL500 dataset [18, 19]. CAL500 is a corpus of 500 tracks of Western popular music, each of which has been manually annotated by three human annotators, at least, using a vocabulary of 174 tags. The tags used in CAL500 dataset annotation span six semantic categories, namely instrumentation, vocal characteristics, genres, emotions, acoustic quality of the song, and usage terms (e.g. "I would like to listen this song while *driving*, *sleeping* etc") [19]. All the recordings were converted to monaural wave format at a sampling frequency of 16 kHz and quantized with 16 bits. Moreover, the music signals have been normalized, so that they have zero mean amplitude with unit variance in order to remove any factors related to the recording conditions.

Following the experimental set-up used in [1, 5, 19], 10-fold cross-validation was employed during the experimental evaluation process. Thus each training set consists of 450 audio files. Accordingly, the training tensor $\mathcal{A}_{CAL500} \in \mathbb{R}_+^{96 \times 8 \times 450}$ was constructed by stacking the auditory temporal modulations representations. The projection matrix $\mathbf{P}$ is derived from the training tensor $\mathcal{A}_{CAL500}$ by employing either the SMLENTF or the MLE [20]. The length of the tag vector produced by our system is 10, that is each test music recording was annotated with 10 tags. Throughout the experiments, the value of λ in SMLENTF was empirically set to 0.5, while the value of β in the matrix $\mathbf{M}$ process was set to 0.5 for both the SMLENTF and the MLE. Furthermore, both the SMLENTF and the MLE produce a semantic space of reduced dimensions, i.e. $k = 150$.

Following [19], two metrics, the mean per-word precision and the mean per-word recall are employed in order to assess the annotation performance of the proposed automatic music tagging system. Per-word recall is defined as the fraction of songs actually labeled with word *w* that the system annotates with label *w*. Per-word precision is defined as the fraction of songs that the system annotates with label *w* that are actually labeled with word *w*. As in [5], if no test music recordings are labeled with the word *w*, then the per-word precision is undefined, and accordingly these words are omitted during the evaluation procedure.

In Table 1 quantitative results on automatic music tagging are presented. For comparison purposes, the best performance of state-of-the-art computer audition systems evaluated on the same dataset is included. In particular, CBA refers to the probabilistic model proposed in [5]. MixHier is Turnbull *et al.* system based on a Gaussian mixture model [19], while Autotag refers to Bertin-Mahieux *et al.* system proposed in [1]. Random refers to a baseline system that annotates songs randomly based on tags' empirical frequencies. Even though the range of precision and recall is $[0, 1]$, the aforementioned metrics may be upper-bounded by a value less than 1 if the number of tags appearing in the ground truth annotation is either greater or lesser than the number of tags that are returned by the automatic music annotation system. Consequently, UpperBnd indicates the best possible performance under each metric. Random and UpperBnd were computed by Turnbull *et al.* [19], and give a sense of the actual range for each metric. Finally, Human indicates the performance of humans in assigning tags to the recordings of the CAL500 dataset. All the reported performance metrics are means and standard errors (i.e. the sample standard deviation divided by the sample size) computed from 10-fold cross-validation on the CAL500 dataset. By inspecting Table 1, SMLENTF clearly exhibits the best performance with respect to the per-word precision and per-word recall among the state-of-the-art computer audition systems that is compared to. Furthermore, MLE outperforms the CBA, the MixHier, and the Autotag systems with respect to per-word precision, while in terms of per-word precision its performance is comparable to that achieved by the CBA and the MixHier. In addition both the SMLENTF and the MLE perform better than the humans with respect to per-word precision and per-word recall in the task under study. These results make our framework the top performance computer

Table 1: Mean Annotation Results on CAL500 Dataset.

System	Precision	Recall
Human [19]	0.296 (0.008)	0.145 (0.003)
UpperBnd [19]	0.712 (0.007)	0.375 (0.006)
Random [19]	0.144 (0.004)	0.064 (0.002)
SMLENTF	**0.387 (0.004)**	**0.173 (0.0015)**
MLE [20]	0.345 (0.004)	0.162 (0.002)
CBA [5]	0.286 (0.005)	0.162 (0.004)
MixHier [19]	0.265 (0.007)	0.158 (0.006)
Autotag [1]	0.281	0.131

audition system that outperforms humans in the music tagging motivating applications to real-world automatic music tagging tasks. The success of the proposed system can be attributed to the fact that the semantic similarities between two music signals with overlapped labels that are measured in a sparse representation-based way rather than in one-to-one way as in [17, 1] by applying multi-label linear embedding into and sparse representations both in the features extraction and classification process.

7. CONCLUSIONS

In this paper, an appealing automatic music tagging framework has been proposed. This framework resorts to auditory temporal modulations for music representation, while multi-label linear embedding and sparse representation-based classification has been employed for multi-label music annotation. A multilinear subspace learning technique (i.e. SMLENTF) has been developed, which incorporates the semantic information of tensor objects (i.e., the auditory temporal modulations) with respect to the music tags into the NTF. The results reported in the paper outperform humans' performance as well as any other result obtained by the state-of-the-art computer audition systems in music tagging applied to the CAL500 dataset.

REFERENCES

[1] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere, "Autotagger: A model for predicting social tags from acoustic features on large music databases," *J. New Music Research*, vol. 37, no. 2, pp. 115-135, 2008.

[2] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, March 2007.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no.1, pp. 33–61, 1998.

[4] D. L. Donoho, and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Information Theory*, vol. 47, no. 7. pp. 2845–2862, 2001.

[5] M. Hoffman, D. Blei, and P. Cook, "Easy as CBA: A simple probabilistic model for tagging music," in *Proc. 10th Int. Symp. Music Information Retrieval*, Kobe, Japan, October 26-30. 2009.

[6] T. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, Sept. 2009.

[7] E. Law, K. West, M. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging," in *Proc. 10th Int. Symp. Music Information Retrieval*, Kobe, Japan, October 26-30. 2009.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization", in *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[9] C. J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Trans. Neural Networks*, vol. 18, no. 6, pp. 1589–1596, 2007.

[10] R. Munkong and J. Biing-Hwang, "Auditory perception and cognition," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 98–117, May 2008.

[11] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Music genre classification using locality preserving non-negative tensor factorization and sparse representations," in *Proc. 10th Int. Symp. Music Information Retrieval*, Kobe, Japan, October 26-30. 2009.

[12] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Music genre classification via sparse representation of auditory temporal modulations," in *Proc. EUSIPCO 2009*, Glasgow, Scotland, August 24-28. 2009.

[13] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Non-Negative multilinear principal component analysis of auditory temporal modulations for music genre classification," *IEEE Trans. Audio Speech and Language Technology*, vol. 18, no. 3, pp. 576–588, March 2010.

[14] R. Seungmin, H. Byeong-jun, and H., Eenjun, "SVR-based music mood classification and context-based music recommendation," in *Proc. 17th ACM Int. Conf. Multimedia*, Beijing, China, October 19-24. 2009, pp. 713–716.

[15] S. Sukittanon, L. E. Atlas, and J. W Pitton, "Modulation-scale analysis for content identification," *IEEE Trans. Signal Processing*, vol. 52, no. 10, pp. 3023–3035, Oct. 2004.

[16] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1700–1715, 2007.

[17] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multilabel classification of music into emotions," in *Proc. 9th Int. Symp. Music Information Retrieval*, Philadelphia, USA, September 14-18. 2008.

[18] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Towards musical query-by-semantic-description using the CAL500 data set," in *Proc. 30th ACM Int. Conf. Research and Development in Information Retrieval*, Amsterdam, The Netherlands, July 23-27. 2007, pp. 439-446.

[19] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio Speech and Language Processing*, vol. 16, no 2, pp. 2008.

[20] C. Wang, S. Yan, L. Zhang, H.-J. Zhang, "Multi-label sparse coding for automatic image annotation," in *Proc IEEE Int. Conf. Computer Vision and Pattern Recognition*, Florida, USA, June 20-25. 2009, pp. 1643-1650.