SPARSE MULTI-LABEL LINEAR EMBEDDING WITHIN NONNEGATIVE TENSOR FACTORIZATION APPLIED TO MUSIC TAGGING

Yannis Panagakis* **Constantine Kotropoulos***
*Dept. of Informatics
Aristotle University of Thessaloniki
Box 451 Thessaloniki, GR-54124, Greece
{panagakis,costas}@aiia.csd.auth.gr

Gonzalo R. Arce†
†Dept. of Electrical & Computer Engineering
University of Delaware
Newark, DE 19716-3130, U.S.A.
arce@ece.udel.edu

ABSTRACT

A novel framework for music tagging is proposed. First, each music recording is represented by bio-inspired auditory temporal modulations. Then, a multilinear subspace learning algorithm based on sparse label coding is developed 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*, whose convergence to a stationary point is guaranteed. Finally, a recently proposed method is employed to propagate the multiple labels of training auditory temporal modulations to auditory temporal modulations extracted from a test music recording by means of the sparse ℓ_1 reconstruction coefficients. The overall framework, that is described here, 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.). They result into a semantic representation of music, which can be used as input to collaborative filtering systems assisting users to search for music content. However, a drawback of such approach is that a newly added music recording must be tagged manually first, before it can be retrieved [18, 19], which is a time consuming and expensive process. Therefore, an emerging problem in Music Information Retrieval (MIR) aims to automate the process of music tagging. 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 [11–13] and emotion [14]. These classification systems effectively annotate music with class labels, such as "rock", "happy", etc., by assuming a predefined taxonomy and an explicit mapping of a music recording onto mutually exclusive classes. However, such assumptions are unrealistic and result into a number of problems, since music perception is inherently subjective [19]. The latter problems can be overcome by the less restrictive approach of annotating the audio content with more than one labels in order to reflect more aspects of music. Relatively little work has been made on multi-label automatic music annotation compared to the work made on multi-label automatic image annotation (cf. [3, 20] and the references therein). However, various automatic music tagging algorithms have been proposed [2, 6, 8, 17, 19]. For instance, audio tag prediction is treated as a set of binary classification problems where standard classifiers, such as the Support Vector Machines [17] or Ada-Boost [2] can be applied. Furthermore, methods that resort to probabilistic modeling have been proposed [6, 19]. These methods attempt to infer the correlations or joint probabilities between the tags and the low-level acoustic features extracted from audio.

In this paper, the problem of automatic music tagging is addressed as a 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 music genre classification [11–13], 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 semantic gap, it is unclear how to exploit the semantic similarity between the label sets associated to two music recordings for efficient feature extraction within multi-label music tagging. Motivated by 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 [2, 6, 17, 19]. There is substantial evidence in the literature that the multilinear subspace learning algorithms are more appropriate for reducing the dimensionality of tensor objects [13, 16]. To this end, a novel multilinear subspace learning algorithm is developed here 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] by formulating an optimization problem, which is then solved by the Projected Gradient method [1, 9]. The proposed method is referred to as *Sparse Multi-label Linear Embedding Nonnegative Tensor Factorization* (SMLENTF). The SMLENTF reduces the high-dimensional feature space, where the high-order data (i.e. the auditory temporal modulations) lie, into 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 be computed efficiently via ℓ_1 optimization. Finally, tags are propagated from the training atoms to a test music recording with the coefficients of sparse ℓ_1 representation.

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 demonstrate 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 CAL500 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 derived by a computational auditory model is briefly described. The SMLENTF is introduced in Section 4. The multi-label annotation framework, that is based on the sparse representations, 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) [7]. Throughout the paper, tensors are denoted by boldface Euler script calligraphic letters (e.g. $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{A}}$), 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 $\boldsymbol{\mathcal{X}}$ 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 $\boldsymbol{\mathcal{X}}$ is addressed by N indices, i.e., $x_{i_1 i_2 i_3 \ldots i_N}$. Mode-$n$ unfolding of tensor $\boldsymbol{\mathcal{X}}$ 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 [7].

An N-order tensor $\boldsymbol{\mathcal{X}}$ 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. $\boldsymbol{\mathcal{X}} = \mathbf{u}^{(1)} \circ \mathbf{u}^{(2)} \circ \ldots \circ \mathbf{u}^{(N)}$. 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 $\boldsymbol{\mathcal{X}}$ is the minimal number of rank-1 tensors that yield $\boldsymbol{\mathcal{X}}$ when linearly combined. Next, several products between matrices will be used, such as the Khatri-Rao product denoted by $\odot$ and the Hadamard product (i.e. element-wise product) denoted by $*$, whose definitions can be found in [7] for example.

3. AUDITORY REPRESENTATION OF TEMPORAL MODULATIONS

A key step for representing music signals in a psychophysiologically consistent manner is to resort on how the 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 *auditory representation of temporal modulations* is employed [13]. The auditory representation is a joint acoustic and modulation frequency representation that discards much of the spectro-temporal details and focuses on the underlying slow temporal modulations of the music signal [15]. 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. It 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 channel of the auditory spectrogram. Psychophysiological evidence justifies the discrete rate $r \in \{2, 4, 8, 16, 32, 64, 128, 256\}$ (Hz) in order to represent the temporal modulation content of sound [13]. 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 modulations of a music recording are 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

Multilinear subspace learning algorithms are required in order to map the high-dimensional original tensor space

onto a lower-dimensional semantic space defined by the labels. In conventional supervised multilinear subspace learning algorithms, such as the General Tensor Discrimininant Analysis [16], it is assumed that data points annotated by the same label should be close to each other in the feature space, while data bearing different labels should be far away. However, this assumption is not valid in a multi-label task, as discussed in [20]. Accordingly, such subspace learning algorithms will fail to derive a lower-dimensional semantic space based on multiple labels.

Let $\{\mathcal{X}_i|_{i=1}^I\}$ be a set of I training nonnegative tensors $\mathcal{X}_i \in \mathbb{R}_+^{I_1 \times I_2 \times \ldots \times I_N}$ of order N. We can represent such a set by a $(N+1)$-order tensor $\mathcal{Y} \in \mathbb{R}_+^{I_1 \times I_2 \times \ldots \times I_N \times I_{N+1}}$ with $I_{N+1} = I$. Furthermore, let us assume that the multi-labels of the training tensor $\mathcal{Y}$ are represented by the matrix $\mathbf{C} \in \mathbb{R}_+^{V \times I}$, where V indicates the cardinality of the tag vocabulary. Obviously, $c_{ki} = 1$ if the ith tensor is labeled with the kth tag in the vocabulary and 0 otherwise. Since, every tensor object (music recording here) 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 exploit the multi-label information in order to drive semantically oriented feature extraction from tensor objects. 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 $w_{ij}^1 = w_{ji}^1 = 1$ and 0 otherwise. However, in real-world datasets, data samples with exactly the same label set are rare. In such a case, the semantic relationship between the 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 $\hat{\mathbf{C}}_i = [\mathbf{c}_{:1}|\mathbf{c}_{:2}|\ldots|\mathbf{c}_{:i-1}|\mathbf{c}_{:i+1}|\ldots|\mathbf{c}_{:I}]$. If $V \ll I$ 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, 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 the optimization problem (1) is NP-hard due to the nature of the underlying combinational optimization. In [5], it has been proved that if the solution is sparse enough, then the solution of (1) is equivalent to the solution of the following 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 [4].

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

Let $d_{ii}^1 = \sum_{i \neq j} w_{ij}^1$ be the diagonal elements of the diagonal matrix $\mathbf{D}^1$. Given $\{\mathcal{X}_i|_{i=1}^I\}$, one can model the semantic relationships between the tensor objects by constructing the multi-label linear embedding matrix, which exploits $\mathbf{W}^1$ and $\mathbf{W}^2$ as in [20]: $\mathbf{M} = \mathbf{D}^1 - \mathbf{W}^1 + \frac{\beta}{2}(\mathbf{I} - \mathbf{W}^2)^T(\mathbf{I} - \mathbf{W}^2)$, where $\beta > 0$ is a parameter, which adjusts the contribution of the ℓ_1 graph in the multi-label linear embedding [20]. Let $\{\mathbf{U}^{(n)}|_{n=1}^{N+1}\}$ be the mode-n factor matrices derived by the NTF applied to $\mathcal{Y}$ [11]. We define $\mathbf{Z}^{(n)} \triangleq \mathbf{U}^{(N+1)} \odot \ldots \odot \mathbf{U}^{(n+1)} \odot \mathbf{U}^{(n-1)} \odot \ldots \odot \mathbf{U}^{(1)}$. One can incorporate the semantic information of tensor objects into the NTF by minimizing the following objective function for the SMLENTF in matricized form:

$$f\big(\mathbf{U}^{(n)}|_{n=1}^{N+1}\big) = \frac{1}{2}\|\mathbf{Y}_{(n)} - \mathbf{U}^{(n)}\big[\mathbf{Z}^{(n)}\big]^T\|_F^2 + \lambda\,\mathrm{tr}\Big\{\big[\mathbf{U}^{(N+1)}\big]^T\,\mathbf{M}\,\mathbf{U}^{(N+1)}\Big\}, \tag{3}$$

where $\lambda > 0$ is a parameter, which controls the trade off between the goodness of fit to the training data tensor $\mathcal{Y}$ and the multi-label linear embedding and $\|.\|_F$ denotes the Frobenious norm. Consequently, we propose to minimize (3) subject to the nonnegative 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 $\mathcal{Y}$ when linearly combined.

Let $\nabla_{\mathbf{U}^{(n)}} f = \frac{\partial f}{\partial \mathbf{U}^{(n)}}$ be the partial derivative of the objective function $f(\mathbf{U}^{(n)}|_{n=1}^{N+1})$ with respect to $\mathbf{U}^{(n)}$. It can be shown that for $n = 1, 2, \ldots, N$ we have

$$\nabla_{\mathbf{U}^{(n)}} f = \mathbf{U}^{(n)}\big[\mathbf{Z}^{(n)}\big]^T\mathbf{Z}^{(n)} - \mathbf{Y}_{(n)}\mathbf{Z}^{(n)}, \tag{4}$$

while for $n = N+1$ we obtain

$$\nabla_{\mathbf{U}^{(N+1)}} f = \mathbf{U}^{(N+1)}\big[\mathbf{Z}^{(N+1)}\big]^T\mathbf{Z}^{(N+1)} + \lambda\, M\mathbf{U}^{(N+1)} - \mathbf{Y}_{(N+1)}\mathbf{Z}^{(N+1)}. \tag{5}$$

Following the strategy employed in the derivation of the Projected Gradient Nonnegative Matrix Factorization [9], we obtain an iterative alternating algorithm for the 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 stationary point of (3) can be found by the update rule:

$$\mathbf{U}_{[t+1]}^{(n)} = [\,\mathbf{U}_{[t]}^{(n)} - n_{[t]}\nabla_{\mathbf{U}_{[t]}^{(n)}} f\,]^+, \tag{6}$$

where t denotes the iteration index and $[.]^+$ is the projection operator, which is defined element-wise as $[.]^+ \triangleq \max(., 0)$. The projection operator ensures that $\mathbf{U}_{[t+1]}^{(n)}$ contains only nonnegative elements after each iteration. The learning rate $n_{[t]}$ can be determined by the Armijo rule along the projection arc [1] or more effectively by the Algorithm 4 in [9] in order to ensure the convergence of the algorithm to a stationary point. The update rule (6) is executed iteratively in an alternating fashion for $n = 1, 2,$

$\ldots, N+1$ until the global convergence criterion is met:

$$\sum_{n=1}^{N+1} \|\nabla^P_{\mathbf{U}^{(n)}_{[t]}} f\|_F \leq \epsilon \sum_{n=1}^{N+1} \|\nabla_{\mathbf{U}^{(n)}_{[t]}} f\|_F, \qquad (7)$$

where $[\nabla^P_{\mathbf{U}^{(n)}_{[t]}} f]_{ij} = \min\left(0, [\nabla_{\mathbf{U}^{(n)}_{[t]}} f]_{ij}\right)$ if $[\mathbf{U}^{(n)}_{[t]}]_{ij} = 0$; and $[\nabla^P_{\mathbf{U}^{(n)}_{[t]}} f]_{ij} = [\nabla_{\mathbf{U}^{(n)}_{[t]}} f]_{ij}$ if $[\mathbf{U}^{(n)}_{[t]}]_{ij} \geq 0$. The parameter ϵ is a predefined small positive number, typically 10^{-5} [9]. The convergence criterion (7) is employed in order to check the stationarity of the solution set $\{\mathbf{U}^{(n)}_{[t]}|^{N+1}_{n=1}\}$ since it is equivalent to the Karush-Kuhn-Tucker optimality condition [1, 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 a 2D auditory representation of temporal modulations is extracted as is 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 $\mathcal{Y} \in \mathbb{R}_+^{I_1 \times I_2 \times I_3}$, where $I_1 = I_f = 96$, $I_2 = I_r = 8$, and $I_3 = I_{samples}$ is obtained. Let $\mathcal{Y}_{train} \in \mathbb{R}_+^{I_1 \times I_2 \times I}$, $I < I_{samples}$, be the tensor where the training auditory temporal modulations representations are stored. By applying the SMLENTF onto the $\mathcal{Y}_{train}$ 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 $\mathcal{Y}_{train}$, respectively. Next, the projection matrix $\mathbf{P} = \mathbf{U}^{(2)} \odot \mathbf{U}^{(1)} \in \mathbb{R}_+^{768 \times k}$, with $k \ll \min(768, I)$, is obtained. The columns of $\mathbf{P}$ span a reduced dimension feature space, where the semantic relations between the vectorized auditory temporal modulations are retained. Consequently, by projecting all the training auditory temporal modulations onto this reduced dimension space an overcomplete dictionary $\mathbf{D} = \mathbf{P}^T \mathbf{Y}^T_{train(3)} \in \mathbb{R}_+^{k \times I}$ is obtained. Alternatively, the dictionary can be obtained by $\mathbf{D} = \mathbf{P}^\dagger \mathbf{Y}^T_{train(3)}$, 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 is projected onto the reduced dimension space and a new feature vector is obtained i.e. $\bar{\mathbf{x}} = \mathbf{P}^T \mathbf{x} \in \mathbb{R}_+^k$ or $\bar{\mathbf{x}} = \mathbf{P}^\dagger \mathbf{x} \in \mathbb{R}^k$. Now, $\bar{\mathbf{x}}$ can be represented as a compact linear combination of the semantically related atoms of $\mathbf{D}$. That is, the test auditory representation of temporal modulations is considered semantically related to the few training auditory representations of 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:

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

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

$$\bar{\mathbf{a}} = \mathbf{C}\,\mathbf{b}. \qquad (9)$$

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]. The CAL500 is a corpus of 500 tracks of Western popular music, each of which has been manually annotated by three human annotators at least, who employ 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 [2, 6, 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 $\mathcal{Y}_{train} \in \mathbb{R}_+^{96 \times 8 \times 450}$ was constructed by stacking the auditory temporal modulations representations. The projection matrix $\mathbf{P}$ was derived from the training tensor $\mathcal{Y}_{train}$ by employing either the SMLENTF or the MLE [20]. The length of the tag vector returned by our system was 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 used in forming the matrix $\mathbf{M}$ was set to 0.5 for both the SMLENTF and the MLE.

Three metrics, the mean per-word precision and the mean per-word recall and the F_1 score 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 annotated by the system with label w that are actually labeled with word w. As in [6], if no test music recordings are labeled with the word w, then the per-word precision is undefined, accordingly these words are omitted during the evaluation procedure. The F_1

score is the harmonic mean of precision and recall, that is $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$.

In Table 1, quantitative results on automatic music tagging are presented. In particular, CBA refers to the probabilistic model proposed in [6]. MixHier is Turnbull *et al.* system based on a Gaussian mixture model [19], while Autotag refers to Bertin-Mahieux *et al.* system proposed in [2]. 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 less 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) inside parentheses computed from 10-fold cross-validation on the CAL500 dataset.

System	Precision	Recall
CBA [6]	0.286 (0.005)	0.162 (0.004)
MixHier [19]	0.265 (0.007)	0.158 (0.006)
Autotag [2]	0.281	0.131
UpperBnd [19]	0.712 (0.007)	0.375 (0.006)
Random [19]	0.144 (0.004)	0.064 (0.002)
Human [19]	0.296 (0.008)	0.145 (0.003)

Table 1. Mean annotation results on the CAL500 Dataset.

In Figure 1, the mean precision, the mean recall, and the F_1 score is plotted as a function of the feature space dimensionality derived by the MLE and the SMLENTF. Clearly, the SMLENTF outperforms the MLE for all the dimensions of the feature space. The best music annotation performance with respect to the mean per-word precision and the mean per-word recall is summarized in Table 2. The numbers inside parentheses are the standards errors estimated thanks to the 10-fold cross-validation.

System	Dimension (k)	Precision	Recall	F_1 Score
MLE [20]	150	0.346 (0.004)	0.154 (0.002)	0.2128
SMLENTF	150	0.371 (0.003)	0.165 (0.002)	0.2291

Table 2. Best mean annotation results obtained by MLE and SMLENTF on the CAL500 Dataset.

By inspecting Table 1, Table 2, and Figure 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, no matter what the feature space dimensionality is. Furthermore, MLE outperforms the CBA, the MixHier, and the Autotag system with respect to the per-word precision, while in terms of the per-word recall its performance is comparable to that achieved by the MixHier. In addition both the SMLENTF and the MLE perform better than humans with respect to the per-word precision and the per-word recall in the task under study. These results make our framework the top performing system in music tagging motivating further research. 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 an one-to-one way as in [2, 6, 17, 19] by applying the multi-label linear embedding and the sparse representations both in the features extraction and the 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 as well as sparse representations have been employed for multi-label music annotation. A multilinear subspace learning technique, the SMLENTF, has been developed, which incorporates the semantic information of 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.

In many real commercial applications, the number of available tags is large. Usually most of the tags are associated to a small number of audio recordings. Thus, it is desirable the automatic music tagging systems to perform well in such small sets. Future research will address the performance of the proposed framework under such conditions.

8. REFERENCES

[1] D. P. Bertsekas: *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.

[2] 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.

[3] 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, 2007.

[4] 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.

[5] D. L. Donoho, and X. Huo: "Uncertainty Principles and Ideal Atomic Decomposition," *IEEE Trans. Information Theory*, Vol. 47, No. 7. pp. 2845–2862, 2001.

[6] M. Hoffman, D. Blei, and P. Cook: "Easy as CBA: A Simple Probabilistic Model for Tagging Music," *Proceedings of the 10th Int. Symp. Music Information Retrieval*, Kobe, Japan, 2009.

[7] T. Kolda and B. W. Bader: "Tensor Decompositions and Applications," *SIAM Review*, Vol. 51, No. 3, pp. 455–500, 2009.

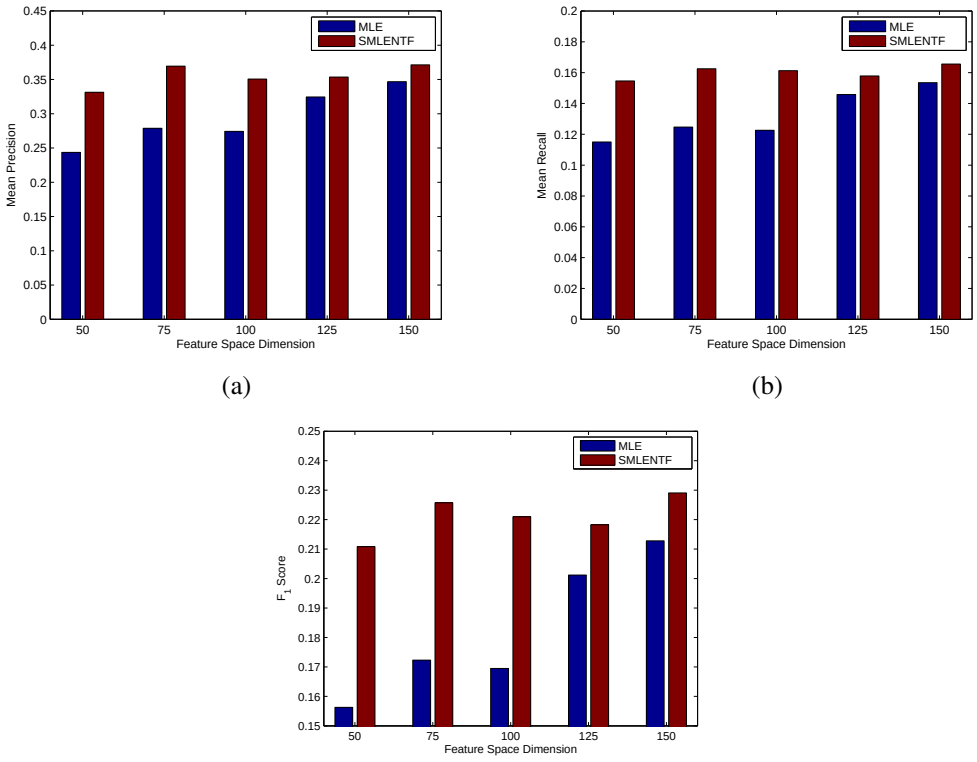

Figure 1. Mean annotation results for the MLE and the SMLENTF with respect to (a) the mean precision, (b) the mean recall, and (c) the F_1 score on the CAL500 dataset.

[8] E. Law, K. West, M. Mandel, M. Bay, and J. S. Downie: "Evaluation of Algorithms Using Games: The Case of Music Tagging," *Proceedings of 10th Int. Symp. Music Information Retrieval*, Kobe, Japan, pp. 387–392, 2009.

[9] C. J. Lin: "Projected Gradient Methods for Nonnegative Matrix Factorization," *Neural Computation*, Vol. 19, No. 10, pp. 2756–2779, 2007.

[10] R. Munkong and J. Biing-Hwang: "Auditory Perception and Cognition," *IEEE Signal Processing Magazine*, Vol. 25, No. 3, pp. 98–117, 2008.

[11] Y. Panagakis, C. Kotropoulos, and G. R. Arce: "Music Genre Classification Using Locality Preserving Non-Negative Tensor Factorization and Sparse Representations," *Proceedings of 10th Int. Symp. Music Information Retrieval*, Kobe, Japan, pp. 249–254, 2009.

[12] Y. Panagakis, C. Kotropoulos, and G. R. Arce: "Music Genre Classification via Sparse Representation of Auditory Temporal Modulations," *Proceedings of EUSIPCO 2009*, Glasgow, Scotland, 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, 2010.

[14] S. Rho, B. Han, and E. Hwang: "SVR-based Music Mood Classification and Context-based Music Recommendation," *Proceedings of 17th ACM Int. Conf. Multimedia*, pp. 713–716, Beijing, China, 2009.

[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, 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," *Proceedings of 9th Int. Symp. Music Information Retrieval*, Philadelphia, USA, pp. 325–330, 2008.

[18] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet: "Towards Musical Query-By-Semantic-Description Using the CAL500 Data Set," *Proceedings of 30th ACM Int. Conf. Research and Development in Information Retrieval*, Amsterdam, The Netherlands, pp. 439-446, 2007.

[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. 467–476, 2008.

[20] C. Wang, S. Yan, L. Zhang, and H.-J. Zhang: "Multi-label Sparse Coding for Automatic Image Annotation," *Proceedings of IEEE Int. Conf. Computer Vision and Pattern Recognition*, Florida, USA, pp. 1643-1650, 2009.