AUTOMATIC MUSIC TAGGING VIA PARAFAC2

Yannis Panagakis and Constantine Kotropoulos

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

ABSTRACT

Automatic music tagging is addressed by resorting to auditory temporal modulations and Parallel Factor Analysis 2 (PARAFAC2). The starting point is to represent each music recording by its auditory temporal modulations. Then, an irregular third order tensor is formed. The first slice contains the vectorized training temporal modulations, while the second slice contains the corresponding multi-label vectors. The PARAFAC2 is employed to effectively harness the multi-label information for dimensionality reduction. Any vectorized test auditory representation of temporal modulations is first projected onto the semantic space derived via the PARAFAC2 and the coefficient vector is obtained. Then, the annotation vector is obtained by multiplying this coefficient vector by the left singular vectors of the second slice (i.e., the slice associated to the label vector). The proposed framework, outperforms the state-of-the-art auto-tagging systems, when applied to the CAL500 dataset in a 10-fold cross-validation experimental protocol.

***Index Terms*—** Automatic Music Tagging, Multi-label Classification, PARAFAC2, Tensor Decompositions.

1. INTRODUCTION

The emergence of Web 2.0 has revealed the importance of the automatic prediction of tags for music in large music database management and music recommendation [1]. Tags are text-based labels encoding semantic information related to music (e.g., instrumentation, genres, emotions) [1, 2]. They result into a semantic representation of music, which can be exploited by music oriented recommendation systems, such as *last.fm* [1] and *Pandora* [2], assisting users to search for music content. In particular, the users of the aforementioned systems can browse large music collections by employing tags provided by other users. However, such an approach suffers from two drawbacks. First, a newly added music recording must be tagged manually, before it can be retrieved [2], which is a time consuming and expensive process. Second, unpopular music recordings may not be tagged at all [1]. Automating music tagging may rectify the just mentioned drawbacks and complement the set of tags provided by humans.

Music information retrieval research has mainly focused on content-based classification of music in terms of genre [3, 4] and emotion [5], that effectively annotates music with class labels, such as “rock”, “happy”, etc. However, one should assume that a predefined taxonomy and an explicit mapping of a music recording onto mutually exclusive classes exists. The latter assumptions are unrealistic since the notion of music similarity is inherently subjective [6, 2]. A less restrictive approach is to annotate the audio content with more than one labels in order to capture more aspects of music. Various content-based automatic music tagging systems have been proposed [1, 2, 6, 7, 8, 9, 10, 11, 12]. Most of the aforementioned systems resort to the so-called *bag-of-features* approach [1], which models the audio signals by the long-term statistical distribution of their short-time spectral features. These features are then fed into machine learning algorithms that associate tags with audio features. For instance, audio tag prediction may be treated as a set of binary classification problems, where standard classifiers, such as the Support Vector Machines [8, 10] or Ada-Boost [7] can be applied. Furthermore, methods have been proposed that resort to probabilistic modeling [2, 12, 9]. These methods attempt to infer the correlations or joint probabilities between the tags and the low-level acoustic features extracted from audio. Recently, we proposed an alternative framework to the aforementioned automatic music tagging systems that resorts to auditory temporal modulations for music representation, while Sparse Multi-label Linear Embedding Nonnegative Tensor Factorization (SMLENTF) was used to efficiently harness the multi-label information for feature extraction as well as multi-label music annotation by means of sparse representations [11].

In this paper, a novel framework for automatic multi-label music annotation is proposed. Following [11], each audio recording is represented by its slow temporal modulations [4]. Such a representation emphasizes the temporal dynamics of

[1] http://www.last.fm/
[2] http://www.pandora.com/

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

the music signal and has been proved to be very robust for both music genre classification [3, 4] and automatic music tagging [11, 12]. However, the auditory temporal modulations do not explicitly utilize the label set (i.e., the tags) of the music recordings. Due to the semantic gap, it is unclear how the semantic similarity between the label sets associated to two music recordings can be exploited for efficient feature extraction within multi-label music tagging. To this end, an irregular third order tensor is formed. The first slice contains the vectorized training temporal modulations, while the second slice contains the corresponding multi-label vectors. The goal is to infer the semantic relationships between the temporal modulations and the label set by computing the SVD for each slice such that the matrix of right singular vectors is the same across both slices. The just mentioned problem is solved effectively via Parallel Factor Analysis 2 (PARAFAC2) [13]. This approach makes sense, since auditory temporal modulations along with its corresponding multi-label vector are represented as linear combinations of basis vectors with coefficients taken from the same vector space. The left singular vectors of the first slice span a lower-dimensional semantic space dominated by the label information. Any vectorized test auditory representation of temporal modulations is first projected onto this semantic space and a coefficient vector is obtained. Then, the annotation vector is obtained by multiplying the coefficient vector by the left singular vectors of the second slice (i.e., the slice associated to multiple labels).

The performance of the proposed automatic music tagging framework is assessed by conducting experiments on the CAL500 dataset [2]. The reported experimental results demonstrate the superiority of the proposed framework over the state-of-the-art auto-tagging systems on the CAL500 dataset, when 10-fold cross-validation is applied.

The paper is organized as follows. In Section 2, basic multilinear algebra concepts and notations are defined. The multi-label annotation framework, that is based on the PARAFAC2 is detailed in Section 3. Experimental results are demonstrated in Section 4 and conclusions are drawn in Section 5.

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) [14]. Throughout the paper, tensors are denoted by boldface Euler script calligraphic letters (e.g. $\boldsymbol{\mathcal{X}}$), 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 $\mathbf{u}_{:j}$. $\|.\|_F$ denotes the Frobenius matrix norm, while $\mathbf{B}^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbf{B}$. Hereafter, 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 [14].

3. MULTI-LABEL ANNOTATION VIA PARAFAC2

Following [11], a two dimensional (2D) representation of its slow temporal modulations is extracted with the same parameters as in [4, 11] for each audio recording. Thus, the ensemble of I training 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 $\boldsymbol{\mathcal{Y}} \in \mathbb{R}_+^{I_1 \times I_2 \times I}$ where $I_1 = I_{frequencies} = 96$ and $I_1 = I_{rates} = 8$, is obtained. Let us also assume that the multi-labels of the training tensor $\boldsymbol{\mathcal{Y}}$ 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., $\mathbf{c}_{:i}$).

Subspace learning algorithms are required in order to map the high-dimensional original feature space onto a lower-dimensional semantic space defined by the labels. In conventional supervised subspace learning algorithms (e.g., Linear Discriminant Analysis) 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. Accordingly, such subspace learning algorithms will fail to derive a lower-dimensional semantic space based on multiple labels.

To overcome the limitation of conventional subspace learning algorithms, a novel application of PARAFAC2 [13] to semantically oriented feature extraction and multi-label multi-class classification problem is proposed here. Let $\mathbf{X}^{(1)} = \mathbf{Y}_{(3)}^T \in \mathbb{R}_+^{768 \times I}$ be the training matrix whose columns are the vectorized auditory temporal modulations and $\mathbf{X}^{(2)} = \mathbf{C} \in \mathbb{R}_+^{V \times I}$ be the corresponding training tag-recording matrix. Our goal is to derive a lower-dimensional semantic space based on multiple labels using the PARAFAC2. The latter is a variant of PARAFAC, a multi-way generalization of the SVD.

Formally, an irregular third order tensor $\boldsymbol{\mathcal{X}}$ is formed. Its first slice contains the vectorized training temporal modulations (i.e., $\mathbf{X}^{(1)}$), while the second slice contains the corresponding multi-label vectors (i.e., $\mathbf{X}^{(2)}$). Since $\boldsymbol{\mathcal{X}}$ has two slices, the PARAFAC2 seeks a decomposition of the form:

$$\mathbf{X}^{(n)} = \mathbf{U}^{(n)} \mathbf{H} \mathbf{S}^{(n)} \mathbf{W}^T, \quad n = 1, 2, \tag{1}$$

where $\mathbf{U}^{(n)} \in \mathbb{R}^{J_n \times k}$, $n = 1, 2$ is an orthogonal matrix for each slice, $\mathbf{H} \in \mathbb{R}^{k \times k}$, $\mathbf{S}^{(n)} \in \mathbb{R}^{k \times k}$ is a diagonal matrix of weights for the nth slice of $\mathfrak{X}$, and $\mathbf{W} \in \mathbb{R}^{I \times k}$ is the coefficient matrix, obviously $J_1 = I_1 \cdot I_2 = 768$ and $J_2 = V$. The decomposition (1) can be obtained by solving the optimization problem:

$$\underset{\mathbf{U}^{(n)},\mathbf{H},\mathbf{S}^{(n)},\mathbf{W}}{\operatorname{argmin}} \sum_{n=1}^{2} \|\mathbf{X}^{(n)} - \mathbf{U}^{(n)}\mathbf{H}\mathbf{S}^{(n)}\mathbf{W}^T\|_F^2. \qquad (2)$$

An effective algorithm for solving (2) can be found in [15].

Having found the decomposition (1), one can form $\mathbf{B} \triangleq \mathbf{U}^{(1)}\mathbf{H}\mathbf{S}^{(1)} \in \mathbb{R}_+^{768 \times k}$ that spans a reduced dimension feature space, where the semantic relations between the vectorized tensor samples are retained. Given a vectorized test auditory temporal modulations representation $\mathbf{x} \in \mathbb{R}_+^{768}$ the reduced dimension feature vector $\tilde{\mathbf{x}} = \mathbf{B}^\dagger \mathbf{x} \in \mathbb{R}^k$ is derived.

By applying the PARAFAC2 on the training tensor, the semantic relations between the label vectors are propagated to the feature space through the common right singular vectors. In music tagging, the semantic relations are expected to propagate from the feature space to the label vector space. Let us denote by $\mathbf{a} \in \mathbb{R}_+^{V \times k}$ the label vector of the test music recording, $\mathbf{a}$ is obtained by

$$\mathbf{a} = \mathbf{U}^{(2)}\mathbf{H}\mathbf{S}^{(2)}\tilde{\mathbf{x}}. \qquad (3)$$

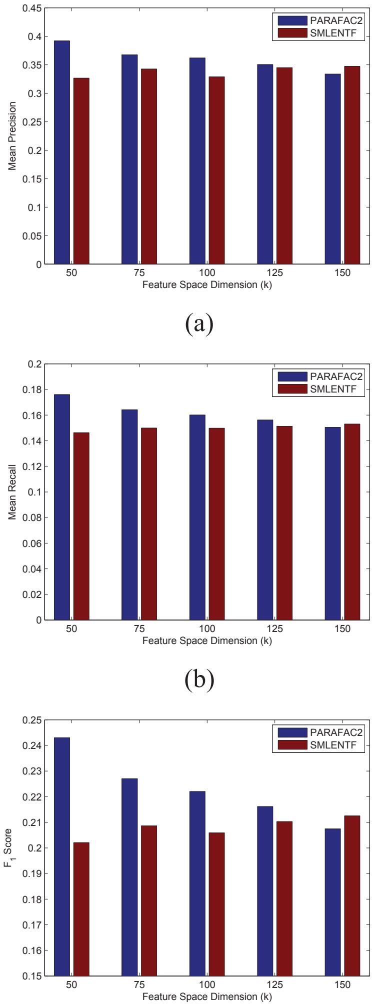
The labels associated with the largest values in $\mathbf{a}$ form the tag vector recommended for the test music recording.

4. EXPERIMENTAL EVALUATION

In order to assess the performance of the proposed framework in automatic music tagging, experiments were conducted on the CAL500 dataset [2]. 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 $V = 174$ tags. The tags used in CAL500 dataset annotation span six semantic categories. All the recordings were preprocessed as in [11].

Following the experimental set-up used in [2, 7, 9, 11], 10-fold cross-validation was employed during the experimental evaluation process. Thus, each training set consists of 450 recordings. All the test music recordings are annotated by using (3). The length of the tag vector returned by the system under study was 10. That is, each test music recording was annotated with 10 tags. 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 whose definitions can be found in [9, 11].

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 PARAFAC2 and the previous state-of-the-art auto-tagging system (i.e., the SMLENTF [11]). Clearly, the PARAFAC2 outperforms the SMLENTF for most of the reduced feature space dimension and especially the small one.



(a)

(b)

(c)

Fig. 1. Mean annotation results for the PARAFAC2 and the SMLENTF with respect to (a) the mean precision, (b) the mean recall, and (c) the F_1 score on the CAL500 dataset.

In Table 1, quantitative results on automatic music tagging based on audio features only are summarized. 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. Thus, UpperBnd indicates the best possible performance under each metric. Random and UpperBnd were computed in [2] and give a sense of the actual range for each metric. Human indicates the performance of humans in assigning tags to the recordings of the CAL500 dataset. 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 with vocabulary size $V = 174$ on the CAL500 dataset except for the auto-tagging systems HEM-GMM and HEM-DTM, which have been evaluated using a smaller vocabulary (i.e., 74 tags) and 5-fold cross-validation.

Table 1. Mean annotation results on the CAL500 Dataset.

System	**Protocol**	**Precision**	**Recall**	F_1 **Score**
PARAFAC2	10FCV, V =174	**0.392 (0.003)**	**0.176 (0.002)**	**0.243**
SMLENTF [11]	10FCV, V =174	0.371 (0.003)	0.165 (0.002)	0.229
CBA [9]	10FCV, V =174	0.286 (0.005)	0.162 (0.004)	0.207
MixHier [2]	10FCV, V =174	0.265 (0.007)	0.158 (0.006)	0.198
ModelAvg [2]	10FCV, V =174	0.189 (0.007)	0.108 (0.009)	0.137
Autotag1 [7]	10FCV, V =174	0.281	0.131	0.179
Autotag2 [7]	10FCV, V =174	0.312	0.153	0.205
HEM-GMM [12]	5FCV, V =74	0.490	0.230	0.260
HEM-DTM [12]	5FCV, V =74	0.470	0.250	0.300
UpperBnd [2]	10FCV, V =174	0.712 (0.007)	0.375 (0.006)	0.491
Random [2]	10FCV, V =174	0.144 (0.004)	0.064 (0.002)	0.089
Human [2]	10FCV, V =174	0.296 (0.008)	0.145 (0.003)	0.194

By inspecting Table 1 and Figure 1, PARAFAC2 clearly exhibits the best performance with respect to the per-word precision and per-word recall, and F_1 score among the state-of-the-art auto-tagging systems, that is compared to, with respect to 10-fold cross-validation. Better performance may obtained by preserving the nonnegativity of the auditory temporal modulations or by adding more slices to the training tensor, which can capture, for example, the lyrics content of music recording or the contextual information of the tags, or social networks indices.

5. CONCLUSIONS

An appealing automatic music tagging framework has been proposed. This framework resorts to auditory temporal modulations for music representation, while the PARAFAC2 has been employed for semantically oriented feature extraction and multi-label music annotation. The results reported in the paper outperform humans' performance as well as any other result obtained by the state-of-the-art auto-tagging systems in the CAL500 dataset when 10-fold cross-validation is employed.

ACKNOWLEDGMENT

This work has been supported by HRAKLEITOS II research project, co-funded by the European Social Fund and National Resources.

6. REFERENCES

[1] T. Bertin-Mahieux, D. Eck, and M. Mandel, "Automatic tagging of audio: The state-of-the-art," in *Machine Audition: Principles, Algorithms and Systems*, Wenwu Wang, Ed. IGI Publishing, 2010, In press.

[2] 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.

[3] C. H. Lee, J. L. Shih, K. M. Yu, and H. S. Lin, "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 670–682, 2009.

[4] 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.

[5] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. 11th Int. Symp. Music Information Retrieval*, Utrecht, The Netherlands, 2010, pp. 255–266.

[6] R. Miotto and N. Orio, "A probabilistic approach to merge context and content information for music retrieval," in *Proc. 11th Int. Symp. Music Information Retrieval*, Utrecht, The Netherlands, 2010, pp. 15–20.

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

[8] M. I. Mandel and D. P. W. Ellis, "Multiple-instance learning for music information retrieval," in *Proc. 9th Int. Symp. Music Information Retrieval*, Philadelphia, USA, 2008, pp. 577–582.

[9] 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, 2009, pp. 369–374.

[10] S. R. Ness, A. Theocharis, G. Tzanetakis, and L. G. Martins, "Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs," in *Proc. 17th ACM Int. Conf. Multimedia*, Beijing, China, 2009, pp. 705–708.

[11] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Sparse multi-label linear embedding within nonnegative tensor factorization applied to music tagging," in *Proc. of 11th Int. Symp. Music Information Retrieval*, Utrecht, The Netherlands, 2010, pp. 393–398.

[12] E. Coviello, L. Barrington, A. B. Chan, and G. R. G. Lanckriet, "Automatic music tagging with time series models," in *Proc. 11th Int. Symp. Music Information Retrieval*, Utrecht, The Netherlands, 2010, pp. 81–86.

[13] R. A. Harshman, "PARAFAC2: Mathematical and technical notes," *UCLA Working Papers in Phonetics*, vol. 22, pp. 30–47, 1972.

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

[15] P. Chew, B. Bader, T. Kolda, and A. Abdelali, "Cross-language information retrieval using PARAFAC2," in *Proc. 13th ACM Int. Conf. Knowledge Discovery and Data Mining*, San Jose, CA, USA, 2007, pp. 143–152.