APPLYING SUPERVISED CLASSIFIERS BASED ON NON-NEGATIVE MATRIX FACTORIZATION TO MUSICAL INSTRUMENT CLASSIFICATION

Emmanouil Benetos, Margarita Kotti, Constantine Kotropoulos[*]

Department of Informatics, Aristotle Univ. of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
E-mail: {empeneto, mkotti, costas}@aiia.csd.auth.gr

ABSTRACT

In this paper, a new approach for automatic audio classification using non-negative matrix factorization (NMF) is presented. Training is performed onto each audio class individually, whilst during the test phase each test recording is projected onto the several training matrices. Experiments demonstrating the efficiency of the proposed approach were performed for musical instrument classification. Several perceptual features as well as MPEG-7 descriptors were measured for 300 sound recordings consisting of 6 different musical instrument classes. Subsets of the feature set were selected using branch-and-bound search, in order to obtain the most discriminating features for classification. Several NMF techniques were utilized, namely the standard NMF method, the local NMF, and the sparse NMF. The experiments demonstrate an almost perfect classification (classification error 1.0%), outperforming the state-of-the-art techniques tested for the aforementioned experiment.

1. INTRODUCTION

The need for musical content analysis arises in different contexts and has many practical applications, mainly for automatic music transcription, effective data organization, annotation in multimedia databases, and internet search. Automatic musical instrument classification is the first step in developing such applications. It is a research area which can also be applied to general sound recognition tasks. However, despite the massive research which has been carried out in the automatic speech recognition, limited work has been done on musical content identification.

The problems addressed so far in musical content identification can be broadly classified into two categories: classification of isolated instrument tones and classification of sound segments. Classifiers using isolated tones have a limited use in practical applications, while sound segment classifiers could be effectively used in music information retrieval (MIR) systems. Using sound segments, identifications of 79-84% for 4 classes of instruments were reported in [8] using Bayes decision rules for classification. Cepstral coefficients, constant-Q coefficients and autocorrelation coefficients were used as features to recordings extracted from the MIS Database from UIOWA [1], which is used in this paper as well. More recently, MPEG-7 temporal descriptors and spectral features were used in conjunction with the k-NN algorithm and decision rules based on rough set theory [9]. A recognition rate of 68.4% at best was reported for classifying sounds coming from 18 instrument classes.

Non-negative matrix factorization (NMF) is a subspace method for basis decomposition [4]. Its various modifications have been used in several classification experiments, where the training procedure is performed by applying an NMF algorithm to a data matrix containing the training vectors of all the available classes. This technique results to an unsupervised training approach. NMF classification experiments report encouraging results compared to other unsupervised classifiers, but also indicate that a supervised NMF classification approach is needed to obtain comparable results with other supervised classifiers.

In this work, the problem of automatically classifying musical instrument segments is addressed. Recordings from the UIOWA database were used that form 6 instrument classes. A total number of 9 features were extracted, covering perceptual descriptors as well as spectral descriptors defined by the MPEG-7 audio standard [2]. The first and second moments of the features were considered, creating a feature set of 41 dimensions as explained in Section 5.2. Branch-and-bound selection was applied to the feature set in order to select the subset that maximizes the classification accuracy [12]. The audio files were split into a training set and a test set using 70% of the available data for training and the remaining 30% for testing. For classification, NMF is used by training individually a classifier for each class and projecting the test data onto each trained class matrix. The class label of each test recording is determined by using the cosine similarity measure (CSM). Several variants of the NMF algorithm were employed, such as the standard NMF method, the local, and the sparse NMF enabling a comparative study of the algorithms' efficiency. The results indicate that using the subset comprising of 6 best features and the standard NMF algorithm yields a correct classification rate of 99.0%, outperforming the traditional NMF classification methods and other statistical model-based classifiers employed for the aforementioned experiment [10].

The remainder of the paper is organized as follows. The audio features extracted are discussed in Section 2. Section 3 is devoted to the NMF method and its extensions. Section 4 presents the standard unsupervised NMF classification approach and the proposed supervised classifier. Section 5 describes the data set used, the feature selection strategy, and the experiments performed to assess the performance of the proposed classifier. Finally, conclusions are drawn in Section 6.

2. FEATURE EXTRACTION

In an audio classification system a careful selection of features that are able to accurately describe the temporal and the spectral properties of the sound is vital. In our approach, a combination of features originating from general audio data classification and the MPEG-7

[*] This work has been supported by the project PYTHAGORAS-II, "Efficient techniques for information organization, browsing, and retrieval in multimedia", funded by the Greek Ministry of Education and the European Union.

Table 1. Set of extracted features.

1	Zero-Crossing Rate
2	Delta Spectrum (Spectrum Flux)
3	Spectral Rolloff Frequency
4	Mel-Frequency Cepstral Coefficients
5	MPEG-7 AudioSpectrumCentroid
6	MPEG-7 AudioSpectrumEnvelope
7	MPEG-7 AudioSpectrumSpread
8	MPEG-7 AudioSpectrumFlatness
9	MPEG-7 AudioSpectrumProjection Coefficients

audio framework is used. The complete list of extracted features is presented in Table 1. The scalar features 1-3 are proposed in systems concerning general audio data (GAD) classification and speech recognition. They can be treated as a short-term description of the textural shape of the audio segments. The mel-frequency cepstral coefficients (MFCCs) form a feature vector. They are widely used in audio processing applications providing a description of the spectral shape of the audio signal. For each audio frame of 10 msec duration, 13 MFCCs were used. The features 5-8 are proposed by the MPEG-7 audio standard [2]. They belong to the basic spectral descriptors category. As 9th feature we used the projection coefficient to a single basis. AudioSpectrumProjection coefficients are part of the MPEG-7 spectral basis descriptors.

3. NON-NEGATIVE MATRIX FACTORIZATION

Non-negative matrix factorization (NMF) has been proposed as a novel subspace method in order to obtain a parts-based representation of objects by imposing non-negative constraints [4]. The problem addressed by NMF is as follows. Given a non-negative $n \times m$ data matrix $\mathbf{V}$ (consisting of m vectors of dimensions $n \times 1$), it is possible to find non-negative matrix factors $\mathbf{W}$ and $\mathbf{H}$ in order to approximate the original matrix:

$$\mathbf{V} \approx \mathbf{WH} \tag{1}$$

where the $n \times r$ matrix $\mathbf{W}$ contains the basis vectors and the $r \times m$ matrix $\mathbf{H}$ contains in its columns the weights needed to properly approximate the corresponding column of matrix $\mathbf{V}$ as a linear combination of the columns of $\mathbf{W}$. Usually, the component number r is chosen so that $(n+m)r < nm$, thus resulting in a compressed version of the original data matrix.

To find an approximate factorization in (1), a suitable objective function has to be defined. The generalized Kullback-Leibler (KL) divergence between $\mathbf{V}$ and $\mathbf{WH}$ is the most frequently used objective function. Various algorithms that incorporate additional constraints in deriving (1) have been proposed and are briefly reviewed subsequently.

3.1. Standard NMF

The standard NMF enforces the non-negativity constraints on matrices $\mathbf{W}$ and $\mathbf{H}$. Thus, a data vector can be formed by an additive combination of basis vectors. The proposed cost function is the generalized KL divergence:

$$D(\mathbf{V}||\mathbf{WH}) = \sum_{i=1}^{n}\sum_{j=1}^{m}[v_{ij}\log\frac{v_{ij}}{y_{ij}} - v_{ij} + y_{ij}] \tag{2}$$

where $\mathbf{WH} = \mathbf{Y} = [y_{ij}]$. $D(\mathbf{V}||\mathbf{WH})$ reduces to KL divergence when $\sum_{i=1}^{n}\sum_{j=1}^{m} v_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = 1$. NMF factorization is defined then as the solution of the optimization problem:

$$\min_{\mathbf{W},\mathbf{H}} \; D(\mathbf{V}||\mathbf{WH}) \quad \textit{subject to} \;\; \mathbf{W},\mathbf{H} \geq 0, \sum_{i=1}^{n} w_{ij} = 1 \; \forall j \tag{3}$$

where $\mathbf{W},\mathbf{H} \geq 0$ means that all elements of matrices $\mathbf{W}$ and $\mathbf{H}$ are non-negative. The above optimization problem can be solved by using the iterative multiplicative rules [4].

3.2. Local NMF (LNMF)

Aiming to impose constraints concerning spatial locality and consequently revealing local features in the data matrix $\mathbf{V}$, LNMF incorporates 3 additional constraints into the standard NMF problem: 1) Minimize the number of basis components representing $\mathbf{V}$. 2) The different bases should be as orthogonal as possible. 3) Retain the components giving most important information. The above constraints are expressed in the following LNMF cost function:

$$\begin{aligned} D(\mathbf{V}||\mathbf{WH}) &= \sum_{i=1}^{n}\sum_{j=1}^{m}[v_{ij}\log\frac{v_{ij}}{y_{ij}} - v_{ij} + y_{ij}] \\ &+ \alpha\sum_{i=1}^{r}\sum_{j=1}^{r} u_{ij} - \beta\sum_{i=1}^{r}\sum_{j=1}^{r} q_{ii} \end{aligned} \tag{4}$$

where α, β are constants, $\mathbf{W}^T\mathbf{W} = \mathbf{U} = [u_{ij}]$, and $\mathbf{HH}^T = \mathbf{Q} = [q_{ij}]$. The minimization is similar to the one used in NMF (3) and a local solution can be found by using 3 update rules [5].

3.3. Sparse NMF (SNMF)

Inspired by NMF and sparse coding, the aim of SNMF is to impose constraints that can reveal local sparse features on data matrix $\mathbf{V}$. The following cost function is optimized for SNMF:

$$D(\mathbf{V}||\mathbf{WH}) = \sum_{i=1}^{n}\sum_{j=1}^{m}[v_{ij}\log\frac{v_{ij}}{y_{ij}} - v_{ij} + y_{ij}] + \lambda\sum_{j=1}^{m}||\mathbf{h}_j||_l \tag{5}$$

where λ is a positive constant and $||\mathbf{h}_j||_l$ the l-norm of the j-th column of $\mathbf{H}$. An SNMF factorization is defined as in (3), including also that $\forall i ||\mathbf{w}_i||_l = 1$. In SNMF, the sparseness is measured by a linear activation penalty: the minimum l-norm of the column of $\mathbf{H}$. A local solution of the minimization problem (5) can be obtained by the update rules proposed in [6].

4. CLASSIFICATION BASED ON NMF

4.1. Unsupervised NMF classification

The standard approach to audio classification in the NMF subspace is performed as follows [10]. Using data from the training set, the data matrix $\mathbf{V}$ is created (each column $\mathbf{v}_j$ contains a feature vector computed from an audio file). The training procedure is performed by applying an NMF algorithm to the data matrix yielding the basis matrix $\mathbf{W}$ and the encoding matrix $\mathbf{H}$.

In the test phase, for each test audio recording, represented by a feature vector $\mathbf{v}_{test}$, a new test encoding vector is obtained by:

$$\mathbf{h}_{test} = \mathbf{W}^{\dagger}\mathbf{v}_{test} \tag{6}$$

where $\mathbf{W}^{\dagger}$ is defined as the Moore-Penrose generalized inverse matrix of $\mathbf{W}$. Having formed during training N classes of encoding vectors $\mathbf{h}_l$, $l = 1, 2, \ldots, N$ (by applying an NMF algorithm on $\mathbf{V}$, yields matrices $\mathbf{W}$ and $\mathbf{H}$ as in (1)), a nearest neighbor classifier is employed to classify the new test sample by using the cosine similarity measure (CSM). The class label l' of the test sound is:

$$l' = \arg \max_{l=1,2,\ldots,N} \left\{ \frac{\mathbf{h}_{test}^T \mathbf{h}_l}{\|\mathbf{h}_{test}\| \|\mathbf{h}_l\|} \right\} \tag{7}$$

thus maximizing the cosine of the angle between $\mathbf{h}_{test}$ and $\mathbf{h}_l$. An alternative measure is also used, where the class label of each test file is determined by examining each row of $\mathbf{h}_{test}$:

$$l' = \arg \max_i h_{i,test} \tag{8}$$

where $h_{i,test}$ is the i-th element of $\mathbf{h}_{test}$.

4.2. The proposed approach

The major drawback of the NMF classifier presented in Section 4.1 is the unsupervised manner of learning parts-based patterns from the data, since no information about the class discrimination is incorporated into the NMF training procedure. In addition, the initial random values of matrices $\mathbf{W}$ and $\mathbf{H}$ can affect greatly the convergence of the algorithm, as the value of NMF objective function defined in (2) may result in a local minimum, thus not yielding in an appropriate factorization.

The creation of a supervised classifier where the NMF training procedure is performed for each data class individually is proposed. This results in a pair of matrices $\mathbf{W}$ and $\mathbf{H}$ for each class:

$$\mathbf{V}_i = \mathbf{W}_i \mathbf{H}_i, \quad i = 1, 2, \cdots, N \tag{9}$$

where $\mathbf{N}$ is the number of different classes, $\mathbf{V}_i$ the data matrix of class i. The number of components used for training each class is given by:

$$r_i = \left\lfloor \frac{n_i m_i}{n_i + m_i} \right\rfloor \tag{10}$$

where n_i and m_i are the dimensions of matrix $\mathbf{V}_i$. In a sense, this approach is an application of one-class classification, where the training of each class is performed individually, by using a set of training data representing the respective class in the absence of counter-examples [11].

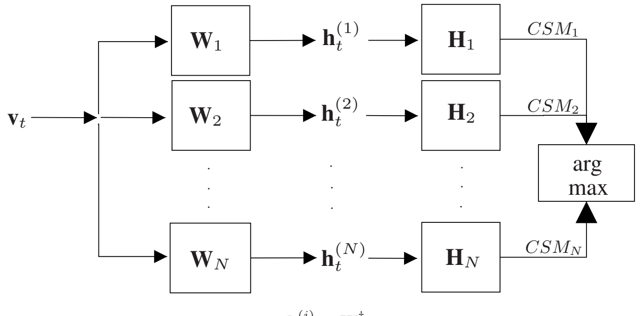

Fig. 1. Testing using the proposed NMF classifier ($\mathbf{h}_t$ and $\mathbf{v}_t$ stand for $\mathbf{h}_{test}$ and $\mathbf{v}_{test}$ respectively).

During test procedure, each test sound is represented by the feature vector $\mathbf{v}_{test}$, as in the approach described in Section 4.1. Afterwards, $\mathbf{v}_{test}$ is projected onto each class basis matrix $\mathbf{W}_i$, yielding:

$$\mathbf{h}_{test}^{(i)} = \mathbf{W}_i^{\dagger} \cdot \mathbf{v}_{test} \tag{11}$$

For each class, the vector $\mathbf{h}_{test}^{(i)}$ is compared to each column vector of matrix $\mathbf{H}_i$ using the CSM. The vector that maximizes the CSM for the matrix $\mathbf{H}_i$ is calculated as a measure of similarity for this class:

$$CSM_i = \max_{j=1,2,\ldots,r_i} \left\{ \frac{\mathbf{h}_{test}^{(i)T} \mathbf{h}_j^{(i)}}{\|\mathbf{h}_{test}^{(i)}\| \|\mathbf{h}_j^{(i)}\|} \right\} \tag{12}$$

where $\mathbf{h}_j^{(i)}$ represents the j-th column of matrix $\mathbf{H}_i$. Finally, the class label of the recording is determined by the the maximum CSM_i, i.e.:

$$l' = \arg \max_{i=1,2,\ldots,N} \{CSM_i\} \tag{13}$$

A block diagram of the testing procedure using the proposed NMF classification method is plotted in Figure 1.

5. EXPERIMENTAL RESULTS

5.1. Dataset

Audio files extracted from the Musical Instrument Samples database collected by the university of Iowa [1] were used. Overall 300 audio files were extracted that belong to 6 different instrument classes: piano, violin, cello, flute, bassoon, and soprano saxophone. In detail, 58 piano recordings, 101 violin recordings, 52 cello recordings, 31 saxophone recordings, 29 flute recordings, and 29 bassoon ones were used. The 300 sounds are partitioned into a training set of 210 audio files and a test set of 90 audio files, preserving a 70%/30% analogy between the two sets, which is typical for classification experiments. All recordings have a duration of about 20 sec and are sampled at 44.1 kHz sampling rate.

5.2. Feature selection

For each feature described in Section 2, its mean and its variance were computed, resulting in 41 features in total. In order to reduce the feature vector dimension, a suitable feature subset for classification has to be selected. The optimal feature subset should maximize the ratio of the inter-class dispersion over the intra-class dispersion:

$$J = \mathrm{tr}(\mathbf{S}_w^{-1} \mathbf{S}_b) \tag{14}$$

where $\mathrm{tr}(\cdot)$ stands for the trace of a matrix, $\mathbf{S}_w$ is the within-class scatter matrix, and $\mathbf{S}_b$ is the between-class scatter matrix. Because the number of distinct subsets is $\frac{41!}{(41-D)!D!}$, where D is the desired subset size, the branch-and-bound search strategy is considered for complexity reduction. In this strategy, a tree structure of $(41 - D + 1)$ levels is created, where every node corresponds to a subset. The highest level corresponds to the full set, while each node corresponds to a D-dimensional subset at the lowest level. The branch-and-bound algorithm traverses the structure using a depth-first search with backtracking [12].

Table 2. Subset of the 6 best features.

1	Mean of the 1st MFCC
2	Variance of the 1st MFCC
3	Mean of the AudioSpectrumFlatness
4	Variance of the AudioSpectrumFlatness
5	Mean of the AudioSpectrumEnvelope
6	Mean of the AudioSpectrumSpread

Two separate experiments using the various NMF algorithms have been performed by employing different feature subsets in order to find the feature dimension that maximizes the classification performance. In the first experiment, 6 features were used, while in the second experiment a set of 20 features was utilized. The subset of 6 best features is summarized in Table 2.

5.3. Performance Evaluation

Experiments were carried out using 7-fold cross validation and the mean value of the classification accuracy and its standard deviation for the three NMF algorithms and for all the two feature subsets is shown in Figure 2. The SNMF algorithm was tested using two different values for the parameter λ (0.001 and 0.1). The highest mean accuracy of 99.0% is achieved by the standard NMF algorithm when the subset of 6 features is used. The achieved results outperforms the classification accuracy for the aforementioned experiment in [10] which used the standard NMF classifier, as well as supervised GMM and continuous HMM classifiers. In addition, the accuracy of NMF exceeds 97% when the 20-feature subset is employed. The LNMF is clearly outperformed by all algorithms, which may be explained due to the locality constraints LNMF imposes when applied to holistic descriptors. The SNMF overall displays better results than the LNMF, but its efficiency depends on the selection of parameter λ (performance is slightly better when $\lambda = 0.001$).

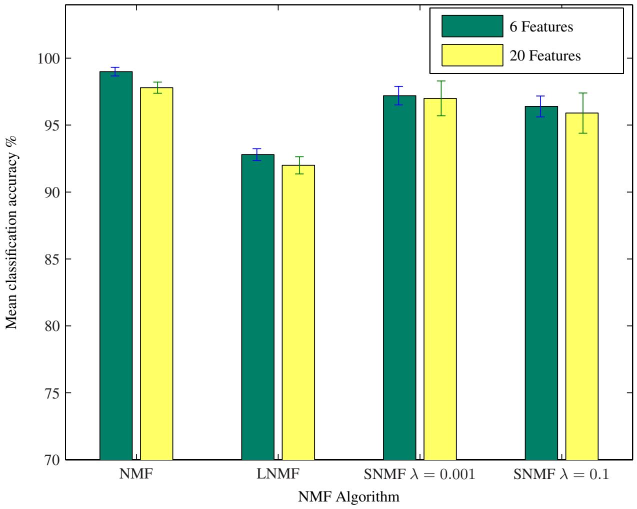

Fig. 2. Mean classification accuracy for NMF algorithms.

Additional information about the performance of the standard NMF algorithm using the 6-dimensional set is depicted in Table 3 where a confusion matrix for one run of the experiment is depicted. The columns of the confusion matrix correspond to the predicted musical instrument and the rows to the actual one. Only one misclassification occurs for the flute, that it is wrongly classified as piano. It is worth mentioning that the flute samples displayed similar dynamical and spectral shape with some piano samples.

6. CONCLUSIONS

In this paper, we have proposed a new method of classifying audio signals using non-negative matrix factorization which is trained individually for each class. Experiments applied to musical instrument classification indicate that the standard NMF algorithm can classify the musical instrument recordings with a high accuracy compared to its variants. It has also been shown that a feature subset selection can increase the classification accuracy. In the future, NMF techniques will be applied to discriminate the whole spectrum of orchestral instruments. Finally, for musical instrument classification experiments, advanced timbral features could also be extracted, such as the timbral descriptors proposed by the MPEG-7 standard.

Table 3. Confusion matrix for standard NMF, 6 Features.

Instr.	Piano	Bassoon	Cello	Flute	Sax	Violin
Piano	**18**	0	0	0	0	0
Bassoon	0	**9**	0	0	0	0
Cello	0	0	**16**	0	0	0
Flute	1	0	0	**8**	0	0
Sax	0	0	0	0	**9**	0
Violin	0	0	0	0	0	**29**

7. REFERENCES

[1] Univ. of Iowa Musical Instrument Sample Database, http://theremin.music.uiowa.edu/index.html.

[2] MPEG-7 overview (version 9), *ISO/IEC JTC1/SC29/WG11 N5525*, March 2003.

[3] H. G. Kim, N. Moreau, and T. Sikora, "Audio classification based on MPEG-7 spectral basis representations," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 716-725, May 2004.

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

[5] S. Z. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in Proc. *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-6, 2001.

[6] C. Hu, B. Zhang, S. Yan, Q. Yang, J. Yan, Z. Chen, and W. Ma, "Mining ratio rules via principal sparse non-negative matrix factorization," in Proc. *IEEE Int. Conf. Data Mining*, 2004.

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

[8] J. C. Brown, O. Houix, and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instruments," *J. Acoustical Society of America*, vol. 109, no. 3, pp. 1064-1072, March 2001.

[9] A. Wieczorkowska, J. Wroblewski, P. Synak, and D. Slezak, "Application of temporal descriptors to musical instrument sound recognition," *J. Intelligent Information Systems*, vol. 21, no. 1, pp. 71-93, July 2003.

[10] E. Benetos, M. Kotti, C. Kotropoulos, J. J. Burred, G. Eisenberg, M. Haller, and T. Sikora, "Comparison of subspace analysis-based and statistical model-based algorithms for musical instrument classification," *2nd Workshop On Immersive Communication And Broadcast Systems*, October 2005.

[11] D. M. J. Tax, *One-Class Classification*, PhD thesis, Delft University of Technology, The Netherlands, 2001.

[12] F. van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*, London UK: Wiley, 2004.