MUSICAL INSTRUMENT CLASSIFICATION USING NON-NEGATIVE MATRIX FACTORIZATION ALGORITHMS AND SUBSET FEATURE SELECTION

Emmanouil Benetos, Margarita Kotti, Constantine Kotropoulos[*]

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

ABSTRACT

In this paper, a class of algorithms for automatic classification of individual musical instrument sounds is presented. Several perceptual features used in sound classification applications as well as MPEG-7 descriptors were measured for 300 sound recordings consisting of 6 different musical instrument classes. Subsets of the feature set are selected using branch-and-bound search, obtaining the most suitable features for classification. A class of classifiers is developed based on the non-negative matrix factorization (NMF). The standard NMF method is examined as well as its modifications: the local, the sparse, and the discriminant NMF. The experimental results compare feature subsets of varying sizes alongside the various NMF algorithms. It has been found that a subset containing the mean and the variance of the first mel-frequency cepstral coefficient and the AudioSpectrumFlatness descriptor along with the means of the AudioSpectrumEnvelope and the AudioSpectrumSpread descriptors when is fed to a standard NMF classifier yields an accuracy exceeding 95%.

1. INTRODUCTION

The need for musical content analysis arises in different contexts. It has many practical applications, mainly for automatic music transcription, effective data organization and annotation in multimedia databases, and music retrieval. Automatic musical instrument classification is the first step in the development of such systems. However, despite the massive research which has been carried out on a similar field, namely the automatic speech recognition, limited work has been done on musical content identification.

The experiments carried out so far 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 a practical application, while sound segment classifiers could be effectively used in music retrieval systems. Using sound segments, identifications of 79-84% for 4 instrument classes were reported by employing Bayes decision rules for classification [9]. Cepstral, constant-Q, and autocorrelation coefficients were used as features to recordings extracted from the MIS Database from UIOWA [1] that is used in this paper as well. More recently, Synak et al [10] used MPEG-7 temporal descriptors and various spectral features for sound segments consisting of 18 instrument classes and developed 2 classifiers. The first classifier uses the k-NN algorithm, while the second one uses decision rules based on rough sets theory. They achieve a recognition rate of 68.4% at best.

In this paper, the problem of automatically classifying musical instrument segments is addressed. Recordings from the UIOWA database [1] were used that form 6 instrument classes. 9 features were extracted, covering descriptors used in sound classification experiments as well as spectral descriptors defined by the MPEG-7 audio standard [2]. The first and second moments of the frame-based extracted features were considered, creating a feature set of 41 dimensions. Branch-and-bound selection was applied to the feature set in order to select the most suitable subset that maximizes the classification accuracy [12]. For unsupervised classification, we used the non-negative matrix factorization (NMF) [4], a subspace method for basis decomposition. Several modifications of the NMF were tested and a comparison of their accuracy is made. 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. Feature subsets of varying dimensions were also considered. The results indicate that using the best subset comprising of 6 features and the standard NMF algorithm yields a correct classification rate of 95.2%, which is comparable to the performance of supervised classifiers for the same experiment [11].

The remainder of the paper is organized as follows. The audio features used are discussed in Section 2. Section 3 is devoted to the NMF subspace method and its extensions. Section 4 describes the feature selection strategy, the classification methodology, and the experiments performed to assess its performance. Finally, conclusions are drawn and future directions are indicated in Section 5.

[*] 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.

1-4244-0469-X/06/$20.00 ©2006 IEEE

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

2. FEATURE EXTRACTION

In an audio classification system a careful selection of features that are able to accurately describe the temporal and spectral sound structures is vital. In our approach, a combination of features originating from general audio data classification and the MPEG-7 audio framework is used. The complete list of extracted features is shown 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. 13 MFCCs were used for each audio frame of 10 msec duration. 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 [4] as a novel subspace method in order to obtain a parts-based representation of objects, by imposing non-negative constraints. The problem addressed by NMF is as follows: Given a non-negative $n \times m$ matrix $\mathbf{V}$ (data matrix, consisting of m vectors of dimension n), 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 the weights needed to properly approximate the corresponding column of matrix $\mathbf{V}$, as a linear combination with the columns of $\mathbf{W}$. Usually, 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 posed in (1), a suitable objective function has to be defined and the generalized Kullback-Leibler (KL) divergence between $\mathbf{V}$ and $\mathbf{WH}$ is most frequently used. Various NMF algorithms, differing mainly in the constraints included in their objective function are presented below.

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 $\sum_{i=1}^{n}\sum_{j=1}^{m} v_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = 1$. An NMF optimization is defined as:

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

where $\mathbf{W,H} \geq 0$ means that all elements of matrices $\mathbf{W}$ and $\mathbf{H}$ are non-negative. The optimization problem (3) can be solved by using 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) Different bases should be as orthogonal as possible. 3) Retain components giving most important information. The above constraints are expressed in the following constrained 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 (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 $\mathbf{H}$. The 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 to the above minimization can be found by using update rules [6].

3.4. Discriminant NMF (DNMF)

DNMF keeps the original constraints from the NMF algorithm, enhances the locality of basis vectors imposed in the LNMF algorithm, and attempts to improve classification accuracy by incorporating into the aformentioned constraints information about class discrimination. Two more constraints are introduced: 1) Minimize the within-class scatter matrix $\mathbf{S}_w$. 2) Maximize the between-class scatter matrix $\mathbf{S}_b$. The modified cost function is expressed as:

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

$$+\alpha\sum_{i=1}^{r}\sum_{j=1}^{r}u_{ij} - \beta\sum_{i=1}^{r}\sum_{j=1}^{r}q_{ii} + \gamma\mathbf{S}_w - \delta\mathbf{S}_b \qquad (6)$$

where γ and δ are constants. Information on the form of the class scatter matrices and the update rules that find a local solution to the minimization of (6) can be found on [7].

4. EXPERIMENTAL PROCEDURE AND RESULTS

4.1. Dataset

We used audio files taken from the MIS database developed by the university of Iowa [1]. 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, which is typical for classification experiments. All recordings are discretized at 44.1 kHz and have a duration of about 20 sec.

4.2. Feature selection

For each feature described in Section 2, its mean and its variance were computed. The features are frame-based and their moments were calculated over the entire duration of the recording, resulting in 41 features in total. In order to reduce the feature set 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 $J = \text{tr}(\mathbf{S}_w^{-1}\mathbf{S}_b)$, where $\text{tr}\{\cdot\}$ stands for the trace of a matrix. Because the number of distinct subsets is $\frac{41!}{(41-D)!D!}$, where D is the desired subset size, the branch-and-bound (b&b) 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 at the lowest level each node corresponds to a D-dimensional subset. The b&b algorithm traverses the structure using a depth-first search with backtracking. More information on the subset selection algorithm can be found in [12].

4.3. Classification method

Musical instrument classification in the NMF subspace is performed as follows. 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 file, represented by a feature vector $\mathbf{v}_{test}$, a new test encoding vector is formed as:

$$\mathbf{h}_{test} = \mathbf{W}^{\dagger}\mathbf{v}_{test} \qquad (7)$$

where $\mathbf{W}^{\dagger}$ is defined as the Moore-Penrose generalized inverse matrix of $\mathbf{W}$. Having formed during training 6 classes of encoding vectors $\mathbf{h}_l$, $l = 1, 2, \ldots, 6$, 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 file is defined as:

$$l' = \arg\max_{l=1,2,\ldots,6}\{\frac{\mathbf{h}_{test}^T\mathbf{h}_l}{||\mathbf{h}_{test}||\,||\mathbf{h}_l||}\} \qquad (8)$$

thus maximizing the cosine of the angle between $\mathbf{h}_{test}$ and $\mathbf{h}_l$.

4.4. Performance Evaluation

Three separate experiments on the various NMF algorithms have been performed by using different feature subsets in order to find the feature dimension that maximizes the classification performance. In the first experiment, 6 features were used, in the second experiment 10 features were tested, while in the third experiment the original set of 41 features was utilized. The subset of 6 best features is presented in Table 2.

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

The mean value of the classification accuracy and its standard deviation for the four NMF algorithms and for all the three feature subsets is presented in Figure 1. The highest

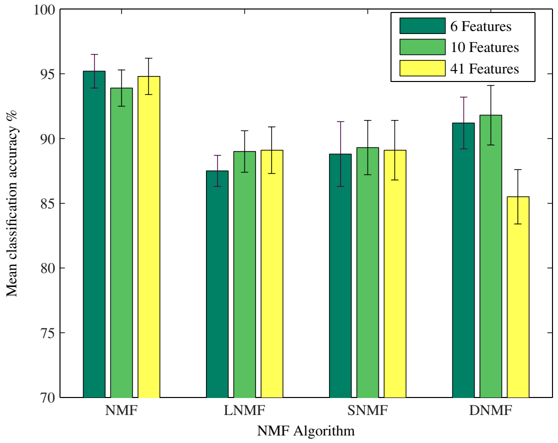

Fig. 1. Classification accuracy for NMF algorithms

Table 3. Confusion matrix for standard NMF, 6 best features.

Instr.	Piano	Bassoon	Cello	Flute	Sax	Violin
Piano	**18**	0	0	0	0	0
Bassoon	1	**8**	0	0	0	0
Cello	0	0	**16**	0	0	0
Flute	2	0	0	**6**	1	0
Sax	0	0	0	0	**9**	0
Violin	0	0	0	0	0	**29**

accuracy is achieved by the standard NMF algorithm with 95.2% when the subset of the best 6 features is used. The achieved performance is comparable to the performance of GMM and HMM classifiers for the same data set, where the achieved performance was 99% and 97%, respectively [11]. It should be noted though, that the present experiments employ unsupervised classification, in contrast to the supervised GMM and HMM classifiers. In addition, the accuracy of the NMF exceeds 93% when either 10 or 41 features are employed. The LNMF is clearly outperformed by all algorithms, which may be explained due to the locality constraints the LNMF imposes when applied to holistic descriptors. The SNMF overall displays better results than the LNMF, but its efficiency depends on the selection of the parameter λ (5). Finally, the DNMF outperforms the LNMF and the SNMF when the subsets of 6 and 10 features are used, but its accuracy drops to 85.5% at the original set, mainly because the algorithm accuracy depends on the values of γ and δ (6).

Additional information about the performance of the standard NMF algorithm using the 6-dimensional set is shown in Table 3 where a confusion matrix is depicted. The columns of the confusion matrix correspond to the predicted musical instrument and the rows to the actual one. Two misclassifications occur for the flute, where it should be noted that the misclassified flute recordings contained the same musical tones (thus, same spectral centroid) with several piano samples.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a method of classifying musical instrument recordings by using NMF algorithms using a variety of features, mainly the MPEG-7 Audio descriptors. The results indicate that the standard NMF algorithm can perform classification with a high accuracy compared to its variants that are more suitable when they are used in conjunction with part-based descriptors. 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 and will be also used in general sound classification experiments. Finally, a supervised NMF classification scheme could be developed, taking into account information about class discrimination.

6. 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] I. Buciu and I. Pitas, "A new sparse image representation algorithm applied to facial expression recognition," in Proc. *17th Int. Conf. Pattern Recognition*, August 2004.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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