Large Scale Musical Instrument Identification

Emmanouil Benetos, Margarita Kotti, and 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, automatic musical instrument identification using a variety of classifiers is addressed. Experiments are performed on a large set of recordings that stem from 20 instrument classes. Several features from general audio data classification applications as well as MPEG-7 descriptors are measured for 1000 recordings. Branch-and-bound feature selection is applied in order to select the most discriminating features for instrument classification. The first classifier is based on non-negative matrix factorization (NMF) techniques, where training is performed for each audio class individually. A novel NMF testing method is proposed, where each recording is projected onto several training matrices, which have been Gram-Schmidt orthogonalized. Several NMF variants are utilized besides the standard NMF method, such as the local NMF and the sparse NMF. In addition, 3-layered multilayer perceptrons, normalized Gaussian radial basis function networks, and support vector machines employing a polynomial kernel have also been tested as classifiers. The classification accuracy is high, ranging between 88.7% to 95.3%, outperforming the state-of-the-art techniques tested in the aforementioned experiment.**

Keywords— **Musical instrument identification, Non-negative matrix factorization, MPEG-7 audio descriptors.**

I. Introduction

Automatic instrument recognition is a subtask of musical content identification. It could be treated as the first step in developing automatic musical transcription systems and multimedia database annotation. The use of real world recordings should be encouraged in instrument recognition experiments, instead of using synthesized instrument sounds, where noise is absent.

The problems addressed so far in musical instrument 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, an identification accuracy between 42% and 57% was reported for 20 classes of instruments using Fast Artificial Neural Networks (FANN), a variant of multilayer perceptrons for classification [6]. Time encoded signal processing and recognition, which is a time-domain specific feature extraction, was employed to recordings extracted from the Musical Instrument Samples Database of UIOWA [1]. The same database is used in this paper as well. A Gaussian mixture model (GMM) classifier for instrument recognition in monophonic and polyphonic audio was proposed [7]. The reported average instrument identification accuracy for 5 classes of the UIOWA samples is 62%.

In this paper, the problem of automatic identification of musical instrument segments is addressed. Recordings from the UIOWA database are used that form 20 instrument classes covering almost all types of orchestral instruments. We extend our previously reported results that were limited to six instruments [8]. A total number of 13 audio features are extracted, including sound description features used in general audio data (GAD) classification experiments as well as descriptors defined by the MPEG-7 audio standard. The first-order and second-order statistics of the features are considered, creating a feature set of 187 dimensions, as explained in Section II. Feature selection using branch-and-bound search strategy is employed in order to select the subset of the most discriminative features [11]. 70% of the available data are used for training and the remaining 30% for testing. Several classifiers have been assessed for musical instrument identification.

The first classifier is based on non-negative matrix factorization (NMF), a subspace method for basis decomposition [2]. The proposed novel NMF classifier trains each class individually and performs Gram-Schmidt orthogonalization to each trained class basis matrix. Orthogonalization has not been employed in the context of NMF, although it is essential because the basis vectors extracted by NMF are not orthogonal. Afterwards, the test data are projected onto each trained class matrix. The class label of each test vector is determined by using the cosine similarity measure (CSM). Several extensions of the NMF method are also tested, such as the standard NMF, the local NMF (LNMF), and the sparse NMF (SNMF), enabling thus a comparative study of algorithms' efficiency. Moreover, multilayer perceptrons (MLPs), radial basis functions (RBF) networks and support vector machines (SVMs) have also been employed for classification and their performance is evaluated. The results indicate that identification accuracy is high for all classifiers, ranging from 88.7% for the LNMF classifier to 95.3% for the SNMF classifier. Performance differences of the several classifiers are examined if they are statistically significant. Experimental results indicate that the employed classifiers outperform traditional unsupervised NMF classifiers and FANN classifiers in the same experiment [6] [8].

The outline of the paper is organized as follows. The set of extracted features is discussed in Section II. Section III is devoted to the NMF method, its variants, and the proposed supervised NMF classifier. The various neural network classifiers employed are presented in Section

IV. Section V describes the data set used, the feature selection strategy, the experimental procedure, and the results. Finally, conclusions are drawn in Section VI.

II. Feature Extraction

A careful selection of sound description features is essential in classification experiments. In our approach, a combination of features originating from GAD classification and the MPEG-7 audio framework is used [5]. The complete list of extracted features is presented in Table I.

TABLE I
Feature set.

no.	Feature	# values/frame
1	MPEG-7 AudioPower	1
2	MPEG-7 AudioFundamentalFrequency	1
3	Total Loudness	1
4	Specific Loudness Sensation	8
5	MPEG-7 AudioSpectrumCentroid	1
6	Spectrum Rolloff Frequency	1
7	MPEG-7 AudioSpectrumSpread	1
8	AudioSpectrumFlattness	4
9	Mel-frequency Cepstral Coefficients	24
10	AutoCorrelation Values	13
11	MPEG-7 Log Attack Time	1
12	MPEG-7 Temporal Centroid	1
13	Zero Crossing Rate	1

Apart from features 10-12, the 1st and 2nd moments of features computed on a frame basis, as well as their derivatives, are computed. This results in 187 features in total. The 1st feature describes the energy of the audio signal. Feature 2 is computed using maximum likelihood harmonic matching. Features 3 and 4 refer to a perceptual modeling of the human auditory system [10]. A description of the spectral shape of the signal is offered by features 5-9. Temporal properties of the signals are extracted in features 10-13. It should be noted that for each audio frame of 10 msec duration, 24 Mel-frequency cepstral coefficients and 8 specific loudness sensation (SONE) coefficients are used. All features are linearly normalized into the [0,1] range.

III. Non-negative matrix factorization (NMF)

NMF is a subspace method able to obtain a parts-based representation of objects by imposing non-negative constraints [2]. 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 $n \times 1$), find the non-negative matrix factors $\mathbf{W}$ and $\mathbf{H}$ in order to approximate the original matrix as:

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

where the $n \times r$ matrix $\mathbf{W}$ contains the basis vectors and the columns of the $r \times m$ matrix $\mathbf{H}$ contain 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 number of components 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 divergence between $\mathbf{V}$ and $\mathbf{WH}$ is the most frequently used objective function:

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

where $\mathbf{WH} = \mathbf{Y} = [y_{ij}]$. Frequently, additional constraints are incorporated to the objective function (2).

A. NMF Variants

The standard NMF method enforces non-negativity constraints on matrices $\mathbf{W}$ and $\mathbf{H}$. The standard NMF factorization is defined as the solution of the optimization problem:

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

The optimization problem (2) can be solved by using the iterative multiplicative rules [2].

Aiming to impose spatial locality in the solution and consequently to reveal 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) Make the different bases as orthogonal as possible. 3) Retain the components giving the most important information. A local solution can be found by using 3 update rules [3].

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

$$D(\mathbf{V}||\mathbf{WH}) = \sum_{i=1}^{n}\sum_{j=1}^{m}[v_{ij}\log\frac{v_{ij}}{y_{ij}} - v_{ij} + y_{ij}] + \lambda\sum_{j=1}^{m}||\mathbf{h}_j||_l \tag{4}$$

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 $||\mathbf{w}_i||_l = 1, \forall i$. In SNMF, the sparseness is measured by the minimum l-norm of the column of $\mathbf{H}$. A local solution of the minimization problem (4) can be obtained by the update rules proposed in [4].

B. Supervised NMF Classification

The major drawback of unsupervised NMF classification is the 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 values of matrices $\mathbf{W}$ and $\mathbf{H}$ can affect the convergence of the algorithm, as the value of NMF objective function defined in (2) may be trapped in a local minimum.

The creation of a supervised classifier where the NMF training procedure is performed for each data class individually is proposed in [8]. It results in a pair of matrices $\mathbf{W}$ and $\mathbf{H}$ for each class:

$$\mathbf{V}_i = \mathbf{W}_i\mathbf{H}_i, \qquad i = 1, 2, \cdots, N \tag{5}$$

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

$$r_i = \left\lfloor \frac{n_i m_i}{n_i + m_i} \right\rfloor \tag{6}$$

where n_i and m_i are the dimensions of matrix $\mathbf{V}_i$. However, the basis defined by the columns of matrix $\mathbf{W}_i$ is not orthogonal. Thus the proposed classifier performs Gram-Schmidt orthogonalization on $\mathbf{W}_i$ by utilizing QR decomposition:

$$\mathbf{W}_i = \mathbf{Q}_i \mathbf{R}_i, \qquad i = 1, 2, \cdots, N \tag{7}$$

where the $n \times r$ matrix $\mathbf{Q}_i$ is orthogonal and the $r \times r$ matrix $\mathbf{R}_i$ is upper triangular. Consequently, the orthogonal basis matrix for each class is now $\mathbf{Q}_i$ and the new encoding matrix is:

$$\mathbf{H}'_i = \mathbf{R}_i \mathbf{H}_i, \qquad i = 1, 2, \cdots, N \tag{8}$$

During test procedure, each test recording is represented by the feature vector $\mathbf{v}_{test}$. Afterwards, $\mathbf{v}_{test}$ is projected onto each class basis matrix $\mathbf{Q}_i$, yielding:

$$\mathbf{h}^{'(i)}_{test} = \mathbf{Q}^{\dagger}_i \cdot \mathbf{v}_{test}. \tag{9}$$

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

$$CSM_i = \max_{j=1,2,\ldots,r_i} \left\{ \frac{\mathbf{h}^{'(i)T}_{test} \mathbf{h}^{'(i)}_j}{||\mathbf{h}^{'(i)}_{test}|| ||\mathbf{h}^{'(i)}_j||} \right\} \tag{10}$$

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

$$l' = \arg \max_{i=1,2,\ldots,N} \{CSM_i\}. \tag{11}$$

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

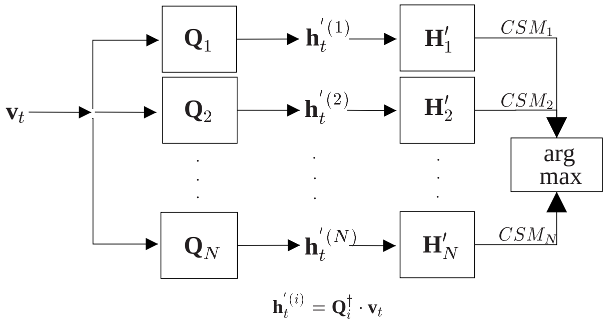

Fig. 1. Testing using the proposed NMF classifier ($\mathbf{h}'_t$ and $\mathbf{v}_t$ stand for $\mathbf{h}'_{test}$ and $\mathbf{v}_{test}$ respectively).

IV. Neural Network Classifiers

Several types of artificial neural networks (ANNs) have been employed for classification. Firstly, a 3-layered MLP with a logistic activation function is utilized. The learning technique used is the back-propagation algorithm, with learning rate equal to 0.3. Moreover, a normalized Gaussian RBF network is considered. The k-means clustering algorithm provides the basis functions and the logistic regression model is employed for learning. Finally, an SVM classifier with a 1st order polynomial kernel is used. The multi-class problem described in Section V is solved using pairwise classification.

V. Experimental Results

A. Dataset

Audio files extracted from the Musical Instrument Samples database collected by the university of Iowa [1] were used. Overall 1000 audio files were extracted that belong to 20 different instrument classes: alto flute, alto saxophone, double bass, bass clarinet, bass flute, bass trombone, bassoon, B♭ clarinet, cello, E♭ clarinet, horn, piano, soprano saxophone, tenor trombone, trombone, tuba, viola, violin, flute, and oboe. The recordings are partitioned into a training set of 700 recordings and a test set of 300 recordings, preserving a 70%/30% proportion 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.

B. Feature selection

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:

$$J = \mathrm{tr}(\mathbf{S}^{-1}_w \mathbf{S}_b) \tag{12}$$

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 $\frac{187!}{(187-D)!D!}$, 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 $(187 - D + 1)$ 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 search strategy traverses the structure using depth-first search with backtracking [11]. After selecting subsets of sizes 10, 20, 40, and 80, the set of 40 best features is considered most suitable for musical instrument identification.

C. Performance Evaluation

Experiments are carried out using 3-fold cross validation. About 200 iterations for the NMF classifiers are needed for convergence during training, while the MLP classifier requires about 500 epochs. Testing for all classifiers is an almost instantaneous procedure. The mean value of the classification accuracy and its standard deviation for the three NMF algorithms and the 3 ANN classifiers is shown in Figure 2. The SNMF algorithm is tested using two different values for the parameter λ (0.1 and 0.001). The highest mean classification accuracy of 95.3% is achieved by the SNMF algorithm when $\lambda = 0.001$. The achieved result by far outperforms the recognition accuracy for the 20-class recognition experiment in [6]. In addition, this application outperforms the results from the 6-class recognition experiment in [8], which used unsupervised and supervised NMF classifiers without basis orthogonalization. It should be noted, however, that the performance of the SNMF algorithm depends on the selection of λ. The lowest accuracy of 88.7% is achieved by the LNMF classifier, while the 3 ANN classifiers display very good recognition rates ranging from 93.21% to 93.88%.

Consequently, the statistical significance of the recognition rates between the SNMF classifier and the remaining classifiers is addressed. The method described in [9] is employed, where an assumption is made that the errors of all classifiers are distributed according to the binomial law. It is shown that the performance gains of the SNMF 2 classifier are statistically significant up against the performance of the NMF, LNMF and SNMF 1 classifiers with 95% confidence ($\alpha = 0.05$). Conversely, the difference in the performance of the SNMF 2 classifier and that of SVM, RBF, and MLP classifiers is found to be statistically insignificant. Insight to the performance of the SNMF algorithm with $\lambda = 0.001$ is provided in Table II, where the confusion matrix for one run of the experiment is detailed. The columns of the confusion matrix correspond to the predicted musical instrument and the rows to the actual one. The indices of the instruments correspond to their order of presentation in Section V-A. It is worth mentioning that 4 instances of bass clarinet were misclassified as B♭ clarinet. In addition, 3 instances of bass flute were misclassified as alto flute. As far as the SVM classifier is concerned, misclassifications occur between the Alto Saxophone and the Soprano Saxophone.

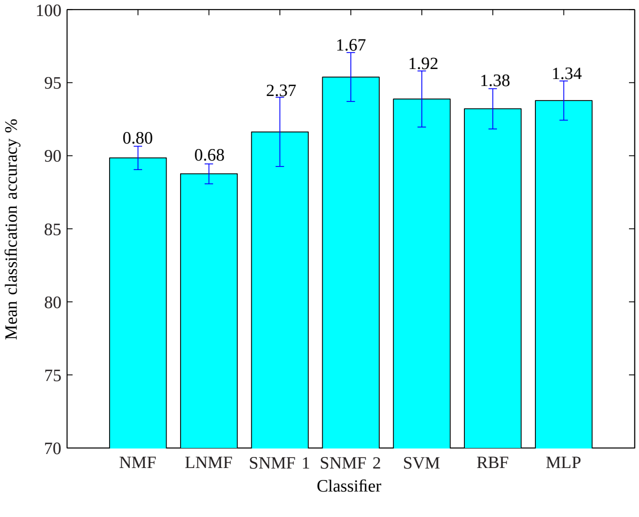

Fig. 2. Mean and standard deviation of the classification accuracy for the supervised NMF classifiers and the neural network ones.

Inst.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	**11**	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	**9**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	**12**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	**10**	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
5	3	0	0	0	**7**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	**8**	0	0	0	0	0	0	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	1	0	0	0	**8**	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	**13**	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	**3**	0	0	0	0	0	0	0	0	1	0
11	0	0	0	0	0	0	0	0	0	0	**3**	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	**18**	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	**9**	0	0	0	0	0	1	0
14	0	0	0	0	0	1	0	0	0	0	0	0	0	**13**	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**12**	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**18**	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	**53**	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**39**	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**14**	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**15**

TABLE II

CONFUSION MATRIX FOR 1 RUN OF SNMF CLASSIFIER (λ = 0.001).

VI. CONCLUSIONS

In this paper, musical instrument recognition experiments have been performed on a large set of recordings with a variety of sound description features. A novel classifier using non-negative matrix factorization with basis orthogonalization was employed and tested against several machine learning classifiers. Future work will focus on classification using several sound collections.

ACKNOWLEDGMENT

This work has been supported by the FP6 European Union Network of Excellence MUSCLE ”Multimedia Understanding through Semantics, Computation and Learning” (FP6-507752). E. Benetos was supported by the “Alexander S. Onassis” Public Benefit Foundation and M. Kotti by the Propondis Public Welfare Foundation, through scholarships.

REFERENCES

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

[2] 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.

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

[4] 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.

[5] G. Peeters, “A large set of audio features for sound description (similarity and classification) in the cuidado project,” Technical report, IRCAM, 2004.

[6] G. Mazarakis, P. Tzevelekos and G. Kouroupetroglou, “Musical Instrument Recognition and Classification Using Time Encoded Signal Processing and Fast Artificial Neural Networks,” *Lecture Notes in Artificial Intelligence (LNAI)*, vol. 3955, pp. 246-255, 2006.

[7] J. Eggink and G. J. Brown, “Application of missing feature theory to the recognition of musical instruments in polyphonic audio,” in Proc. *4th Int. Conf. Music Information Retrieval*, October 2003.

[8] E. Benetos, M. Kotti, and C. Kotropoulos, “Applying supervised classifiers based on non-negative matrix factorization to musical instrument classification,” in Proc. *IEEE Int. Conf. Multimedia & Expo*, July 2006.

[9] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, “What size test set gives good error rate estimates?,”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 52-64, January 1998.

[10] E. Pampalk, “A Matlab Toolbox to Compute Similarity from Audio,” in Proc. *5th Int. Conf. Music Information Retrieval*, October 2004.

[11] 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.