Music Genre Classification using On-line Dictionary Learning

M. Srinivas, Debaditya Roy and C. Krishna Mohan
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad - 502205, India
Email: {cs10p002, cs13p1001, ckm}@iith.ac.in

Abstract—**In this paper, an approach for music genre classification based on sparse representation using MARSYAS features is proposed. The MARSYAS feature descriptor consisting of timbral texture, pitch and beat related features is used for the classification of music genre. On-line Dictionary Learning (ODL) is used to achieve sparse representation of the features for developing dictionaries for each musical genre. We demonstrate the efficacy of the proposed framework on the Latin Music Database (LMD) consisting of over 3000 tracks spanning 10 genres namely Axé, Bachata, Bolero, Forró, Gaúcha, Merengue, Pagode, Salsa, Sertaneja and Tango.**

Keywords. *Music Genre Classification, Sparse Representation, Dictionary Learning*

I. Introduction

Over the last decade, there have been many advances made in the field of music genre classification. These efforts echo the need for music genre classification in the thriving music industry which is growing bigger with the use of Internet for digitally storing and retrieving music. Silla et al. [1] introduced the use of segments of music data in order to classify the genre. All the 3227 audio clips belonging to the LMD were divided into three segments - begin(B), middle(M) and end(E). MARSYAS framework [2] devised by Tzanetakis was then used to calculate features for each segment. Feature vectors of size 30×1 were obtained for all three segments - B, M, E which were then applied with different classifiers like Multi Layer Perceptron, 3-Nearest Neighbors, Naïve-Bayes and Support Vector Machine (SVM) among which SVM demonstrated best classification accuracy.

Timbral texture, beat and pitch related features contribute to the MARSYAS feature descriptor which comprise of 30 feature values. Out of the 30 features, 6 features are beat related, 15 are timbral texture related and 5 are pitch related. The MARSYAS features are described in detail in [3] where genetic algorithm (GA) based feature selection (FS) was used which involves finding an ensemble of features including time and space decompositions to best represent the audio track. This ensemble of features when combined with the MARSYAS features and components of the audio signals gave better classification than just the MARSYAS used alone. A comparative analysis of various classifiers using the ensemble of features was performed, in which SVM produced the best results.

Further work on dynamic ensemble of classifiers was done in [4], where 109 features sets encompassing timbral, spectral, mel frequency cepstral coefficients, chroma, spectral centroid, roll off, spectral flux, zero crossings, spectral flatness measure, spectral crest factor, line spectral pair, linear prediction cepstral coefficients and stereo panning spectrum were used for classification. Two variations of the k-Nearest Oracles (KNORA) [5] method known as KNORA Eliminate (KE) and KNORA Union (KU) were applied for dynamic ensemble selection of the above features. KU demonstrated better classification results than KE, when used in combination with SVM.

In [6], Costa et al. employed textural features for classification. This involved representing the audio as spectrograms and then computing Local Binary Patterns (LBP) [7] from the spectrograms. Feature extraction was done both locally (zone-wise) and globally (on the whole spectrogram). The classification performance obtained by using LBP textural features with the SVM classifier was better than the dynamic ensemble approach. Costa et al. further improved upon their method in [8], where LBP features were combined with Mel Scale Zoning (MSZ) technique [9] to obtain better classification on the LMD database using the SVM classifier. Gradually, music genre classification moved in the direction of hybrid content based classifiers with Rhythm Histograms (RH), Inset-Onset Interval Histogram Coefficients (IOIHC) and Statistical Spectrum Descriptors (SSD) being used in [10] along with MARSYAS features. Upon application of SVM for classification, the hybrid classifiers outperformed the methods involving single feature sets.

In [11], Ren et al. proposed the use of time-constrained sequential patterns (TSPs) for identifying music genres. At first, TSP features were extracted from each track and then TSP mining was applied to discover genre-specific TSPs. This was followed by the computation of occurrence frequencies of TSPs in each music piece. These frequency values were then applied to a linear SVM for classification. In [12], a classifier known as the ℓ_1-SVM was applied on the Mel Frequency Cepstral Coefficients (MFCCs) of audio tracks for 1886 tracks spanning nine genres. This classifier worked better than the ℓ_2-SVM and the same ℓ_1-lasso distance is used in the proposed classification technique for comparing the sparsity between two dictionaries.

Yeh et al. [13] proposed a dual-layer bag-of-frames model in which keywords were computed both at the frame-level and the segment-level. This yielded two dictionaries which were then used as features to train and test both the linear and histogram intersection kernel (HIK) SVM. HIK SVM which uses a kernel based on the bag-of-frames gives better classification than the linear SVM. Mairal et al. [14] proposed the concept of Supervised Dictionary Learning (SDL). However, the major drawback of the procedure was deciding the sparsest representation which would be suitable for testing as the dictionary could not adapt to dynamic changes in data as happened in the case of videos. In [15], this problem was overcome with the introduction of On-line Dictionary Learning (ODL) which could be updated on-the-fly based upon the data stream of feature vectors available. This learning technique forms the basis of the proposed classification scheme.

The remainder of this paper is arranged as follows. Section II explains the proposed sparsity based classification method in detail with an insight into ODL. Section III presents our experiments and discusses the performance of the proposed method with other existing approaches. Finally, section IV provides conclusions and possible future directions of this work.

II. Music Genre Classification with Sparsity Based Classifier

This section provides details about the proposed method for music genre classification based on sparse representation. First, a description of the MARSYAS features is presented. Next, a background on On-line Dictionary Learning is provided. Finally, the working of the sparsity based classification framework is explained in detail.

Figure 1 describes the proposed classification scheme. The MARSYAS features for the three segments of each track in the LMD are obtained [17]. The classification scheme works in two distinct phases - training and testing. In the training phase, dictionaries are developed for each class using On-line Dictionary Learning and all the dictionaries are combined to form a single dictionary. Subsequently, in the testing phase, the sparsity of a test clip is computed with the dictionaries of each class using the ℓ_1-lasso distance. The class which exhibits maximum sparsity is then assigned as the class for that test clip.

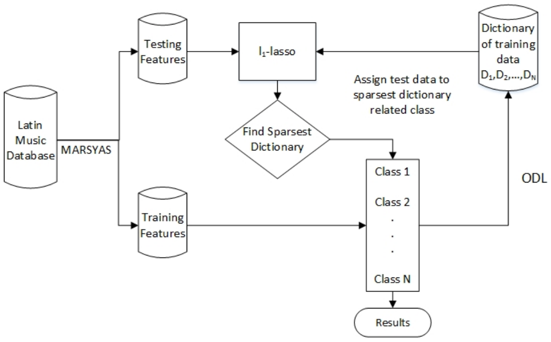

Fig. 1. Block Diagram of the proposed Music Genre Classification Scheme with On-line Dictionary Learning

A. Feature Extraction

The features are collected as three segments from each audio track, where each segment is a 30 second clip equivalent to 1,153 audio samples [1]. The segments are extracted as follows:

1) The first segment or begin segment (B) is computed from audio sample 0 to 1,153.
2) If N is the total number of audio samples in the track, then the middle segment (M) starts from $(N/3 + 500)$ to $(N/3 + 1653)$.
3) The end segment (E) is calculated from the end of the track but to avoid noise, it is overlapped with middle segment, i.e. from $(N/3 + 1453)$ to $(N - 300)$.

Then the MARSYAS framework [16] is used to extract 30 (timbral, pitch and beat related) feature values from each segment in the form of 30-dimensional vector.

B. On-line Dictionary Learning

Consider a signal x in $\mathbb{R}^m$. We say that it admits a sparse approximation over a dictionary D in $\mathbb{R}^{m\times k}$, with k columns referred to as atoms, when one can find a linear combination of a "few" atoms from D that is "close" to the signal x. This dictionary D is computed as follows: 1) Dictionaries are computed for each class namely, $D_1, D_2, \ldots, D_N$. 2) These dictionaries are then combined to form a single dictionary D for the entire training samples of the given dataset. The entire dictionary construction is shown in Algorithm 1.

Algorithm 1 : Dictionary Construction for each training class dataset using online dictionary learning algorithm (ODL)

Inputs: Training class datasets $N\epsilon\mathbb{R}^{m\times n}(\ C_1,\ ...,\ C_N)$, and $T\epsilon R$ (regularization parameter)
Output: N Dictionaries $D\epsilon\mathbb{R}^{m\times k} = [D_1, ..., D_N]$ $(k \ll n)$.

Dictionary construction:
Step 1. For $i = 1$ to N do
Step 2. Construct dictionary D_i for each training class C_i using online dictionary learning algorithm (ODL).

$$(\hat{\mathbf{D}}_\mathbf{i}, \hat{\mathbf{\Phi}}_\mathbf{i}) = \arg\min_{\mathbf{D}_i,\Phi_i} \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\|\mathbf{C}_i - \mathbf{D}_i\Phi_i\|_2^2 + \lambda\|\Phi_i\|_1$$

satisfying $\mathbf{C}_i = \hat{\mathbf{D}}_i\hat{\Phi}_i, \quad i = 1, 2, \ldots, N.$
Step 3. End for
Step 4. Return D_i

C. Sparsity Based Classification

The sparsity based classification using ODL is used to represent the test data as a sparse linear combination of training data acquired from a dictionary. We construct class $C = [C_1, C_2, ..., C_N]$ consisting of training samples for each segment - B, M, E, available for the given N classes. The samples belonging to the same class C_i lie approximately close to each other in a low-dimensional subspace. Let the p^{th} class have K_p training samples and the total number of training samples is denoted by $\{y_i^N\}$ where $i = 1, 2, \ldots, K_i$ and $K_1, K_2, \ldots, K_N$ are training samples corresponding to classes $C_1, C_2, \ldots, C_N$.

Let b be a input vector belonging to the p^{th} class, then it is represented as a linear combination of the training samples belonging to class p.

$$b = D_p \Phi_p \tag{1}$$

where D_p is a $m \times K_p$ dictionary whose columns are the training samples in the p^{th} class and $\mathbf{\Phi}_p$ is a sparse vector for the same class. The two main steps involved in the proposed method are :

1) *Dictionary Construction:* Construct the dictionary for each class of training features using ODL [15]. Then, the dictionaries $D = [D_1, \ldots, D_N]$ are computed using the equation.

$$(\hat{\mathbf{D}}_{\mathbf{i}}, \hat{\mathbf{\Phi}}_{\mathbf{i}}) = \arg\min_{\mathbf{D}_i, \Phi_i} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \|\mathbf{C}_i - \mathbf{D}_i \Phi_i\|_2^2$$
$$+\lambda \|\Phi_i\|_1 \tag{2}$$

where $\mathbf{C}_i = \hat{\mathbf{D}}_i \hat{\mathbf{\Phi}}_i, \quad i = 1, 2, \ldots, N.$

2) *Classification:* In the classification process, the sparse vector Φ for given test feature is found in the test dataset $B = [b_1, \ldots, b_l]$. Using the dictionaries of training samples $D = [D_1, \ldots, D_N]$, the sparse representation Φ satisfying $D\Phi$=B is obtained by solving the following optimization problem:

$$\begin{aligned} \Phi_j &= \arg\min_{\Phi} \tfrac{1}{2} \|\mathbf{b_j} - \mathbf{D}\Phi_j\|_2^2 \\ &\text{subject to} \|\Phi_j\|_1 \leq T_1, \\ \hat{i} &= \arg\min_i \|\mathbf{b_j} - \mathbf{D}\delta_i(\Phi^j)\|_2^2 \;\; j = 1, \cdots, t \end{aligned} \tag{3}$$

where δ_i is a characteristic function that selects the coefficients. Then b_j is assigned to C_i associated with the i^{th} dictionary. It means, finding the sparsest dictionary for a given testing data using *l₁* -lasso algorithm. Then, test data is assigned to the class associated with this sparsest dictionary.

III. Experimental Results and Discussion

The experiments were performed on LMD with the MARSYAS features which are available in [17]. The MARSYAS feature contain 30 timbral texture, pitch and beat related feature values. For each of the 10 classes in the database, there were 297 samples for the begin segment and 300 for the middle and end segments totaling 2970 and 3000 samples, respectively. Each sample is a 30×1 feature vector. The proposed classifier was then trained using 2700 (270 for each class) training samples for each of the three segments. Testing was done using 270 (27 for each class) samples for the begin segment and using 300 (30 for each class) for the middle and end segments.

TABLE I
Confusion Matrix for begin (B) segment

	Axé	Bac	Bol	For	Gaú	Mer	Pag	Sal	Ser	Tan
Axé	252	0	1	2	1	6	1	3	4	0
Bachata	1	262	0	1	1	2	0	3	0	0
Bolero	1	0	240	4	0	0	5	10	8	2
Forró	7	2	3	231	2	8	9	3	5	0
Gaúcha	6	2	0	1	248	3	2	3	5	0
Merengue	2	2	2	1	4	257	2	0	0	0
Pagode	8	1	2	2	2	1	250	1	3	0
Salsa	2	1	2	7	4	6	5	238	5	0
Sartaneja	12	2	4	6	4	0	4	4	234	0
Tango	0	0	4	0	0	0	0	0	1	265

TABLE II
Confusion Matrix for middle (M) segment

	Axé	Bac	Bol	For	Gaú	Mer	Pag	Sal	Ser	Tan
Axé	288	1	0	0	2	0	4	1	4	0
Bachata	0	300	0	0	0	0	0	0	0	0
Bolero	0	0	295	1	0	0	0	0	4	0
Forró	2	1	1	283	0	1	4	5	3	0
Gaúcha	3	0	0	0	293	0	2	1	1	0
Merengue	0	0	0	0	0	300	0	0	0	0
Pagode	0	0	1	1	1	0	294	0	3	0
Salsa	0	1	2	0	0	0	0	296	1	0
Sartaneja	0	0	2	1	0	0	3	1	293	0
Tango	0	0	1	0	0	0	0	0	0	299

TABLE III
Confusion Matrix for end (E) segment

	Axé	Bac	Bol	For	Gaú	Mer	Pag	Sal	Ser	Tan
Axé	292	0	0	0	2	5	0	1	0	0
Bachata	0	299	0	0	0	1	0	0	0	0
Bolero	0	0	296	0	1	0	3	0	0	0
Forró	1	2	0	289	3	1	2	1	1	0
Gaúcha	3	0	0	0	292	2	1	1	1	0
Merengue	0	0	0	0	0	300	0	0	0	0
Pagode	3	0	0	3	2	0	290	0	2	0
Salsa	0	0	0	1	2	1	0	296	0	0
Sarteneja	3	0	2	1	2	1	0	0	291	0
Tango	0	0	1	0	0	0	0	0	0	299

The confusion matrices for all the three segments on dictionary size of 350 are shown in Table I, II and III. It can be observed that in some cases, the proposed classification scheme is not able to distinguish between Forró, Pagode and Sartaneja categories. This can be traced back to the fact that they originate from the same country (Brazil).

The best performance of 98.13% as reported in Table IV was achieved with the end (E) segment. Also, the results are quite close to the best performance in the middle (M) (98.03%) and begin (B) segment (98.1%) as well, which suggest that if the representation is robust all the segments will give reasonable classification accuracy.

TABLE IV
COMPARISON OF RESULTS WITH STATE-OF-THE-ART ON LATIN MUSIC DATABASE

Methods	Features Used	Classifier	Accuracy(%)
Silla Jr. et al. 2007 [1]	MARSYAS	SVM	63.50
Silla Jr. et al. 2008 [3]	normalised MARSYAS + Global space and time decomposition	SVM	65.06
Silla Jr. et al. 2010 [10]	IOIHC	SVM	53.26
	MARSYAS	SVM	70.40
	RH	SVM	57.80
	SSD	SVM	83.93
	ALL- Features	SVM	88.93
	GA- Selected Features	SVM	88.80
Almeida et al. 2012 [4]	MARSYAS + KE	SVM	59.66
	MARSYAS + KU	SVM	70.31
Costa et al. 2012 [6]	GLCM	SVM	70.78
	LBP	SVM	80.33
Costa et al. 2012 [8]	LBP + Global Features	SVM	79.00
	LBP + Linear Zoning	SVM	77.78
	LBP + Bark Scale Zoning	SVM	78.00
	LBP + Mel Scale Zoning	SVM	82.33
	Acoustic Features	SVM	61.00
Costa et al. 2013 [18]	MARSYAS + KU W	SVM	83.00
Proposed	**MARSYAS**	**Sparsity Based**	**98.13**

The performance of the proposed method when compared to other music genre classification methods including the state-of-the-art [18] is presented in Table IV. All the existing methods in literature have variations in the type of features used or in the method of selection of features. SVM classifier was used for classification in all these methods.

The original feature descriptor size for each class during training was training was 2700×30. The proposed method was tested with dictionaries of size 60, 120, 160, 180, 200, 250, 300 and 350. Generally, accuracy improves for larger sized dictionaries. However, after a certain point, increase in dictionary size does not yield better classification accuracy. The dictionary size at this point of time gives the best possible sparse representation of the given feature descriptor. In our case, recognition rate of 98.13% was obtained for dictionary size of 350.

Table IV enlists the various classification schemes in chronological order applied on the LMD including our proposed method. The first genre classification approach on LMD was demonstrated by Silla et al [1]. Recognition rate of 63.50% was achieved by using MARSYAS features in combination with the SVM classifier. This work was further continued in [3] where space and time decomposition features were used alongside MARSYAS to improve the accuracy to 65.06%. In [10], different features, IOIHC, MARSYAS, RH and SSD were used individually and in combination. The combination of all the features proved to be useful as it resulted in increased recognition rate of 88.93%.

Almeida et al. [4] focussed on selection of features, using KE and KU on MARSYAS to obtain classification accuracy of 59.66% and 70.31%. Costa et al. [6] used textural features like GLCM and LBP to classify musical genres. Classification results of 80.33% was obtained with LBP. In [8], LBP was tested with different zoning techniques like linear, bark scale and mel scale zoning(MSZ). MSZ in combination with LBP when applied to an SVM classifier gave state-of-the-art results for textural features at 82.33%. MARSYAS was again used in [18] in combination with KU Weighted to achieve genre recognition rate of 83%.

IV. CONCLUSION

In this paper, we have presented a classification approach based on sparse representation of MARSYAS feature descriptors using On-line Dictionary Learning on the Latin Music Database. Classification accuracy of 98.13% which is achieved using the proposed approach is better than the state-of-the-art [10]. We also showed that classification accuracy for all the segments - B, M and E were comparable, which demonstrates that the sparse representation can classify music genre even with a single segment of music. In future, this method can be extended to other music databases like GTZAN and ISMIR.

ACKNOWLEDGMENT

We would like to thank Mr. Carlos N. Silla Jr. for providing the feature descriptors from the Latin Music Database.

REFERENCES

[1] C.N. Silla, C.A.A. Kaestner, and A.L. Koerich. Automatic music genre classification using ensemble of classifiers. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pages 1687–1692, 2007.

[2] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE Transactions on*, 10(5):293–302, 2002.

[3] C.N. Silla, A.L. Koerich, and C.A.A. Kaestner. Feature selection in automatic music genre classification. In *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, pages 39–44, 2008.

[4] P.R. Lisboa de Almeida, A. de Souza Britto, E.J. da Silva Junior, L.E. Soares de Oliveira, T. Montes Celinski, and A.L. Koerich. Music genre classification using dynamic selection of ensemble of classifiers. In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, pages 2700–2705, 2012.

[5] Albert H.R. Ko, Robert Sabourin, and Alceu Souza Britto Jr. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, 41(5):1718 – 1731, 2008.

[6] Y.M.G. Costa, L. Oliveira, A.L. Koerich, and F. Gouyon. Comparing textural features for music genre classification. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–6, 2012.

[7] Timo Ojala, Matti Pietikainen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51 – 59, 1996.

[8] Y.M.G. Costa, L.S. Oliveira, A.L. Koerich, F. Gouyon, and J.G. Martins. Music genre classification using {LBP} textural features. *Signal Processing*, 92(11):2723 – 2737, 2012.

[9] S. Umesh, L. Cohen, and D. Nelson. Fitting the mel scale. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 1, pages 217–220 vol.1, 1999.

[10] Carlos N. Silla, Jr., Alessandro L. Koerich, and Celso A. A. Kaestner. Improving automatic music genre classification with hybrid content-based feature vectors. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 1702–1707, New York, NY, USA, 2010. ACM.

[11] Jia-Min Ren and J.R. Jang. Discovering time-constrained sequential patterns for music genre classification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(4):1134–1144, 2012.

[12] K. Aryafar, S. Jafarpour, and A. Shokoufandeh. Automatic musical genre classification using sparsity-eager support vector machines. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1526–1529, 2012.

[13] C.-C.M. Yeh, Li Su, and Yi-Hsuan Yang. Dual-layer bag-of-frames model for music genre classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 246–250, 2013.

[14] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Supervised dictionary learning. *CoRR*, abs/0809.3083, 2008.

[15] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 689–696, New York, NY, USA, 2009. ACM.

[16] G. Tzanetakis. Marsyas(music analysis,retreival and synthesis for audio signals). http://marsyas.info/.

[17] C. Silla. The latin music database. http://www.ppgia.pucpr.br/ silla/lmd/, 2007.

[18] Y. Costa, L. Oliveira, A. Koerich, and F. Gouyon. Music genre recognition based on visual features with dynamic ensemble of classifiers selection. In *Systems, Signals and Image Processing (IWSSIP), 2013 20th International Conference on*, pages 55–58, 2013.