Music Genre Recognition Based on Visual Features with Dynamic Ensemble of Classifiers Selection

Yandre Costa*, Luiz Oliveira†, Alessandro Koerich‡, and Fabien Gouyon§
*State University of Maringá (UEM), Maringá, Brazil
Email: yandre@din.uem.br
†Federal University of Paraná (UFPR), Curitiba, Brazil
Email: lesoliveira@inf.ufpr.br
‡Pontifical Catholic University of Paraná (PUCPR), Curitiba, Brazil
Email: alekoe@ppgia.pucpr.br
§Institute fo Systems and Computer Engineering of Porto (INESC), Porto, Portugal
Email: fgouyon@inescporto.pt

***Abstract*—This paper introduces the use of a dynamic ensemble of classifiers selection scheme with a pool of classifiers created to perform automatic music genre classification. The classifiers are based on support vector machine trained with textural features extracted from spectrogram images using Local Binary Patterns. The results obtained on the Latin Music Database showed that local feature extraction and the k-nearest oracle (KNORA) for dynamic ensemble of classifiers selection can reach a recognition rate of 83%, which is a little better than the best result ever reported on this dataset using the restrictions imposed by "artist filter". In addition, the results are compared with those obtained from traditional approaches using acoustic features.**

I. Introduction

In the last ten years, many researchers of music information retrieval community have devoted efforts developing works related to music genre classification. However, there are still some challenges related to automatic music genre classification. Although much research has been devoted to assess this problem, there are still several challenges related to music genre recognition. In [1], McKay and Fujinaga call attention to some hard problems related to musical genres and mention some experiments in which people were not capable of performing correctly music genre classification in more than 76% of the cases. Although not enough to take definitive conclusions, these experiments allow to reach some understanding about the upper limits on automatic classification. In addition, the authors propose that novel approaches should be presented in order to improve the performance in music genre recognition tasks.

By this way, Costa et al. [2] presented a different approach for automatic music genre recognition using features obtained in the visual domain. In this approach, the audio signal was converted into spectrogram images [3] (Short-Time Fourier representation) and textural features were extracted in the visual domain. In that work, the authors have already shown that local feature extraction, by zoning spectrogram images, could help to achieve better results than performing a global feature extraction. More recently, the authors have shown in [5] and [17] that by creating one classifier for each zone created on the spectrogram image helps to improve the recognition rates. In these works, the authors have already demonstrated that good recognition rates can be obtained combining these classifiers outputs with traditional fusion rules presented by Kittler et al. [6], such as sum rule and product rule.

In this work, we aim to explore the complementarity that may happen between base classifiers trained with features extracted by different zoning strategies by using a dynamic selection of classifiers approach, called KNORA [7]. The main contribution of this work is to carry out experiments using a dynamic ensemble of classifiers selection approach in music genre recognition.

Our experiments were performed on a subset of 1,300 music pieces taken from the Latin Music Database (LMD) [8]. This dataset is composed of music pieces from 10 musical genres. The best recognition rate obtained with the dynamic ensemble of classifiers selection scheme is 83%. In spite of being the best recognition rate ever obtained on the LMD with "artist filter" restriction, we have not found statistically significant difference between the results obtained with or without the dynamic ensemble of classifiers selection approach. We have also compared the results presented here with the results obtained using traditional features (i.e. acoustic features). For this, the results obtained with this kind of features are described as well.

II. Feature Extraction

In this work, we have done a segmentation in which three segments of the signal were taken before performing the spectrogram generation, as suggested by Costa et al. [4]. By this way, we have taken n different sub-signals from the signal S as a whole. All sub-signals taken correspond to a projection of S on the interval $[p, q]$ of samples, or $S_{pq} =< s_p, \ldots, s_q >$. In the generic case, one may extract K (overlapping or non-overlapping) sub-signals and obtain a sequence of spectrograms $\overline{\Upsilon}_1, \overline{\Upsilon}_2, \ldots, \overline{\Upsilon}_K$. As done by Silla et al. in [8], we have considered in this work three 10-second segments: beginning ($\overline{\Upsilon}_{beg}$), middle ($\overline{\Upsilon}_{mid}$), and end ($\overline{\Upsilon}_{end}$). Concerned about avoiding segments potentially unproductive in genre classification, we didn't take into account the first and the last ten seconds of the music pieces. The reason for this is that some undesirable effects frequently found in these parts of the signal, such as fade in and fade out, could be avoided.

Once the segmentation step is done, the audio signal is converted into a spectrogram. For this, we used the following parameters: bit rate = 352 kbps, audio sample size = 16 bits, one channel, and audio sample rate = 22.05 kHz. Figure 1 shows the signal decomposition and spectrogram generation steps.

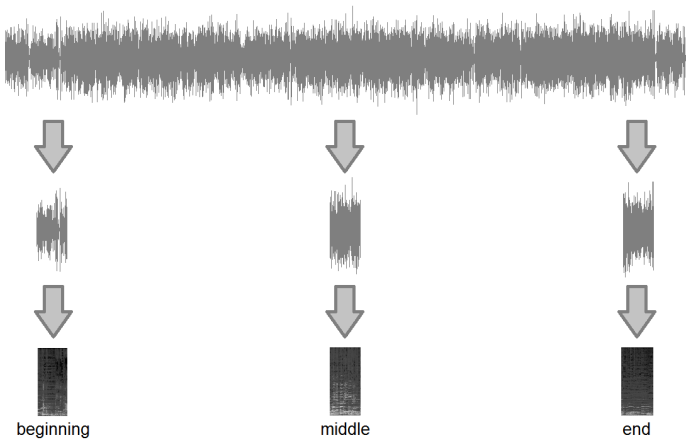

Fig. 1. Signal decomposition and spectrogram generation.

One can notice that texture is the main visual content existent in spectrogram images. Thus, we decided to use the LBP [9] texture operator, a successful texture descriptor frequently used to describe this kind of content.

In [2] we have already demonstrated that a zoning mechanism, in order to preserve some local information rather than a global one, can help to provide better results in terms of recognition rate in this kind of application. In that work, we have used Gray Level Co-occurrence Matrix (GLCM) texture features and only a single classifier was created with feature vectors extracted from all zones. Then, majority voting taking into account each individual zone decision were used in order to get a final decision.

The main aim of this work is to investigate and compare the effects of creating a particular classifier for each created zone and how to combine their outputs, once the oracle between these classifiers has shown very high recognition rates. For this purpose, the classifiers outputs will be combined into two different ways. We have applied two different zoning schemes on the spectrogram images before extracting features.

In the first scheme the spectrogram image is divided into 10 linear zones of equal size, as depicted in Figure 2. Different configurations of linear zoning were evaluated. However, for a number of zones higher than 10, the obtained recognition rates were worse. In addition, this zoning set up is good enough to provide a great number of classifiers, which is desirable to investigate the improvement achieved when using dynamic ensemble of classifiers selection.

In the second scheme, the spectrogram images were zoned according to the Mel scale. The Mel scale is a fundamental result of psychoacoustics, relating real frequency to perceived frequency attempting to represent the frequency bands according to the human perception [10]. With this zoning scheme, each spectrogram image was divided into 15 nonlinear zones, as shown in Figure 3. In both cases, considering that 3

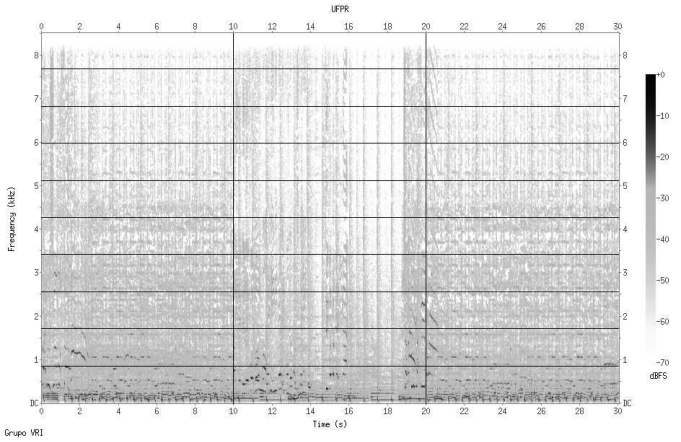

Fig. 2. Linear zoning

spectrogram images are generated from each music piece, since the time decomposition scheme provides 3 segments, the number of total zones, and consequently the number of classifiers is $3n$, where n is the number of created zones per spectrogram.

We believe that the spectrogram image zoning and classifiers combinig scheme is a good way to deal with possible similarities regarding to instruments or rhytmic patterns in the audio signal taken from different music pieces. Using this strategy it is expected to preserve some local information and to capture some particularities of each musical genre.

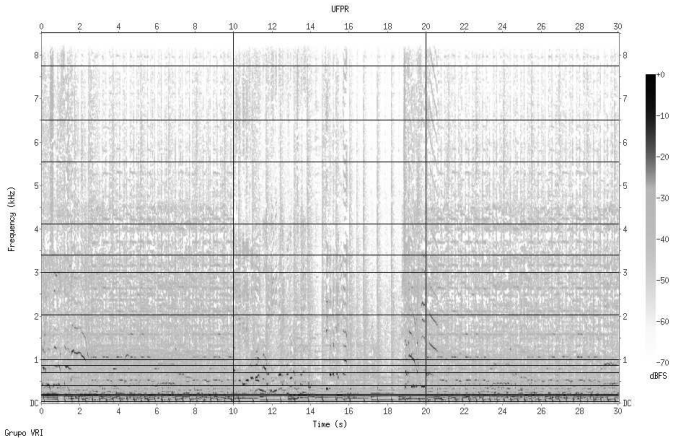

Fig. 3. Mel scale zoning

III. Classifiers and Combination Strategies

Once a pool of classifiers have been created using both linear and Mel scale zoning, their outputs are combined through two different strategies. In the first case we used the well-known sum and product combining rules [6]. In the sum rule, the prediction for each class is given by the summation of the predictions attributed to the class by each individual classifier. In the product rule, the output prediction for each class is given by the multiplication of the predictions attributed to the class by each classifier.

In the second case we evaluated KNORA, a dynamic ensemble of classifiers selection recently presented by Ko et al.

[7]. With KNORA, a group of potentially adequate classifiers (ensemble of classifiers) is selected from a pool of classifiers. For using KNORA, a validation set is necessary.

In KNORA, for any test data point, its nearest K neighbors in the validation set are found. Then, the classifiers which correctly classify those neighbors are figured out and used to compose an ensemble of classifiers for classifying the given pattern [7]. There are some different schemes using KNORA, in this work we evaluated KNORA-ELIMINATE and two variations of KNORA-UNION. In KNORA-ELIMINATE, the classifiers that correctly classify all the K-nearest neighbors of a test pattern are selected to compose the ensemble of classifiers. In KNORA-UNION, the classifiers that correctly classify one of the K-nearest neighbors of a test pattern are selected. Note that, in the last case, a classifier can have more than one vote if it correctly classifies more than one neighbor.

Introduced by Vapink [11], Support Vector Machine (SVM) is the engine choosen to perform classification in this work. The data were normalized according to a linear scaling in which the values range from -1 to +1. We have used the Gaussian kernel, and the parameters C and γ were tuned using a greedy search.

The classification process is carried out as follows: three 10-second segments are extracted from the beginning, middle and end part of the music signal and their spectrograms are computed ($\overline{\Upsilon}_{beg}$, $\overline{\Upsilon}_{mid}$, and $\overline{\Upsilon}_{end}$). Each spectrogram is further split into n zones, according to the values of n described in section II. Then, 59 LBP features are extracted from each spectrogram image zone. Next, one specific classifier is built to each zone created in the images. These classifiers assign a probability to each one of the ten possible classes. Training and classification are performed using the 3-fold cross-validation: 2 folds used for training an N-class SVM classifier, 1 fold for testing, 3 permutations of the training fold (i.e. 1×2+3, 2×1+3, 3×1+2). For each specific zoning scheme, we have created $3n$ classifiers with 600 and 300 feature vectors for training and testing, respectively. When KNORA was used, an additional fold with 400 music pieces was taken as validation set.

Once a lot of classifiers has been created, estimation of probabilities is used to perform the fusion of its outputs aiming to get a final decision. In cases like this, a classifier which produces a posterior probability $P(class|input)$ is required. Thus, the estimation of probabilities is a requirement here, once we are concerned with comparing different fusion strategies. Furthermore, we want to evaluate the KNORA for dynamic ensemble of classifiers selection. In this case, once we have selected an ensemble of classifiers, we can evaluate different ways to combine their outputs. The most common one is to proceed the majority voting between the classifiers outputs. Alternatively, one can use the fusion rules aforementioned. In our experiments, we have evaluated these two ways.

IV. Evaluation of the Results

The three folds described in section III were used to produce the results presented here. For this, the results corresponds to the average recognition rate taken when each one of the three folds was used as testing set. KNORA was used in 8 different ways. Both KNORA-ELIMINATE and KNORA-UNION were tested with majority voting, sum rule and product rule to combine the selected classifiers outputs. In addition, KNORA-UNION was tested with the classifiers outputs weighted by the number of votes assigned to the classifier when the sum and the product fusion rules were used (KNORA UNION W). Aiming to make a comparison between the results obtained here with results obtained with traditional acoustic features, we performed one experiment using 68 features extracted with the framework MARSYAS, more details about these framework can be found in [12]. The recognition rate obtained with these features was about 61%, a recognition rate worse then all the rates obtained with visual features described in this work.

Table I shows the best results achieved when the spectrogram images are split into 10 linear zones with and without using KNORA. In these experiments the value of K ranged from 1 to 20. Table II shows the best results achieved when the spectrogram images are zoned according to the Mel scale with and without using KNORA. The value of K also ranged from 1 to 20.

TABLE I. Results with linear zoning.

Approach	Majority voting	Product rule	Sum rule
Direct Fusion	-	77.78	77.56
KNORA-ELIMINATE	73.78 (K=1)	77.44 (K=5)	77.44 (K=18)
KNORA-UNION	77.56 (K=6)	79.11 (K=2)	78.56 (K=2)
KNORA-UNION W	-	79.11 (K=2)	**79.33** (K=5)

TABLE II. Results with Mel scale zoning.

Approach	Majority voting	Product rule	Sum rule
Direct Fusion	-	82.33	81.11
KNORA-ELIMINATE	74.67 (K=1)	81 (K=19)	80.22 (K=13)
KNORA-UNION	79.67 (K=12)	81 (K=19)	81.89 (K=7)
KNORA-UNION W	-	**83** (K=7)	82.11 (K=13)

One can say that the best result obtained using KNORA is the best result ever obtained on the LMD considering the "artist filter" restriction (Table III). In addition, the confusion matrices showed that the results obtained with KNORA are better for all genres present on the LMD, except for gaúcha and merengue. Nevertheless, the results obtained with or without KNORA are very close. Therefore, we decided to perform a statistical test in order to verify if there is statistically significant difference between the obtained results. The Friedman multi comparison statistical test with Shaffer's procedure with a confidence level of 95% was used, and the results have shown that there is no statistically significant difference between the results obtained with or without KNORA.

TABLE III. Results described on the literature on the LMD.

Work	Recognition rate (%)
Instance selection [13]	59.67
GLCM features [2]	60.11
GLCM + Instance selection [2]	67.20
MIREX 2009 winner [14]	74.66
MIREX 2010 winner [15]	79.86
MIREX 2011 winner [16]	81.90
LBP linear zoning - fusion with product rule [17]	80.33
LBP mel zoning - fusion with product rule [5]	82.33
KNORA UNION W with product rule	**83.00**

V. Conclusion

In this work we have described experiments in music genre recognition. The features used in the classification are extracted

in the visual domain. For this purpose, spectrogram images are generated from the audio signal and LBP textural features are extracted from these images. The spectrogram images are divided into zones. We have evaluated two different zoning schemes, one linear with 10 zones, and one according to the Mel scale, with 15 nonlinear zones.

In order to generate a pool of classifiers, one classifier for each zone was created and their outputs were combined into two ways. In the first way, the outputs were directly combined using conventional fusion rules (i.e. product rule and sum rule). In the second way, a dynamic ensemble of classifiers selection (i.e. KNORA) was used.

The experimental results show that the best results obtained with or without KNORA are very close. However, one can say that the best result described here, about 83% is the best one ever obtained on the LMD dataset when the "artist filter" restriction is used. In future works, we intend to investigate whether it is worth using the dynamic ensemble of classifiers selection approach used here taking into account the overall system complexity increase.

ACKNOWLEDGMENT

This work is partly supported by CAPES (Grants BEX 5779/11-1 and 223/09-FCT595-2009), Araucária Foundation (Grant #16767-424/2009), CNPq (Grants #301653/2011-9 and 402357/2009-4), European Commission, FP7 (Seventh Framework Programme), and ICT-2011.1.5 Networked Media and Search Systems (Grant #287711).

REFERENCES

[1] C. McKay, and I. Fujinaga, "Musical genre classification: Is it worth pursuing and how can it be improved?", in *7th International Conference on Music Information Retrieval*, 2006, pp. 101–106.

[2] Y. M. G. Costa, L. S. Oliveira, A. L. Koerich, and F. Gouyon, "Music genre recognition using spectrograms", in *18th International Conference on Systems, Signals and Image Processing*, 2011, pp. 151–154.

[3] M. R. French, and R. G. Handy, "Sectrograms: turning signals into pictures", in *Journal of engineering technology*, 2007, pp. 25–31.

[4] C. Costa, J. Valle-Jr, and A. L. Koerich, "Automatic classification of audio data", in *International Conference on Systems, Man, and Cybernetics*, 2004, pp. 562–567.

[5] Y. M. G. Costa, L. S. Oliveira, A. L. Koerich, F. Gouyon, and J. Martins, "Music genre classification using LBP textural features", in *Signal Processing*, vol. 92, pp. 2723–2737, 2012.

[6] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, 1998.

[7] A. Ko, R. Sabourin, A. Britto, "From dynamic classifier selection to dynamic ensemble selection", in *Pattern Recognition*, vol. 41, pp. 1718–1731, 2008.

[8] C. N. Silla, A. L. Koerich, and C. A. A. Kaestner, "The latin music database", in *9th International Conference on Music Information Retrieval*, 2008, pp. 451-456.

[9] T. Ojala, M. Pietikinen, and T. Menp, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971–987, 2002.

[10] S. Umesh, L. Cohen, and D. Nelson, "Fitting the mel scale", in *International Conference on Acoustics, Speech, and Signal Processing*, 1999, pp. 217–220.

[11] V. Vapnik, "The Nature of Statistical Learning Theory", in *Springer-Verlag*, New York, 1995.

[12] G. Tzanetakis, and P. Cook, "Musical genre classification of audio signals", in *IEEE Transactions on speech and audio processing*, vol. 10, pp. 293–302, 2002.

[13] M. Lopes, F. Gouyon, A. L. Koerich, and L. S. Oliveira, "Selection of training instances for Music Genre Classification", in *International Conference on Pattern Recognition*, 2010, pp. 4569–4572.

[14] C. Cao, and M. Li, "Thinkit's submission for MIREX 2009 audio music classification and similarity tasks", in *Audio train/test task of MIREX 2009*, 2009.

[15] K. Seyerlehner, M. Schedl, T. Pohle, and P. Kness, "Using block-level features for genre classification, tag classification and music similarity estimation", in *Audio train/test task of MIREX 2010*, 2010.

[16] P. Hamel, "Pooled features classification", in *Audio train/test task of MIREX 2011*, 2011.

[17] Y. M. G. Costa, L. S. Oliveira, A. L. Koerich, and F. Gouyon, "Comparing textural features for music genre classification", in *International Joint Conference on Neural Networks*, 2012, pp. 1867–1872.