Signal Processing 92 (2012) 2723–2737

Contents lists available at SciVerse ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Music genre classification using LBP textural features

Y.M.G. Costa [a,b,*], L.S. Oliveira [b], A.L. Koerich [b,c], F. Gouyon [d], J.G. Martins [b,e]

[a] State University of Maringá (UEM), Av. Colombo, 5790 - Bloco C56, Maringá, PR 87020-900, Brazil
[b] Federal University of Paraná (UFPR), Rua Cel. Francisco H. dos Santos, 100, Curitiba, PR 81531-990, Brazil
[c] Pontifical Catholic University of Paraná (PUCPR), R. Imaculada Conceição, 1155, Curitiba, PR 80215-901, Brazil
[d] Institute for Systems and Computer Engineering of Porto (INESC), R. Dr. Roberto Frias, 378, Porto 4200-465, Portugal
[e] Federal Technological University of Paraná (UTFPR), R. Cristo Rei, 19, Toledo, PR 85902-490, Brazil

ARTICLE INFO

Article history:
Received 31 October 2011
Received in revised form
25 February 2012
Accepted 23 April 2012
Available online 17 May 2012

Keywords:
Music genre
Texture
Image processing
Pattern recognition

ABSTRACT

In this paper we present an approach to music genre classification which converts an audio signal into spectrograms and extracts texture features from these time-frequency images which are then used for modeling music genres in a classification system. The texture features are based on Local Binary Pattern, a structural texture operator that has been successful in recent image classification research. Experiments are performed with two well-known datasets: the Latin Music Database (LMD), and the ISMIR 2004 dataset. The proposed approach takes into account some different zoning mechanisms to perform local feature extraction. Results obtained with and without local feature extraction are compared. We compare the performance of texture features with that of commonly used audio content based features (i.e. from the MARSYAS framework), and show that texture features always outperforms the audio content based features. We also compare our results with results from the literature. On the LMD, the performance of our approach reaches about 82.33%, above the best result obtained in the MIREX 2010 competition on that dataset. On the ISMIR 2004 database, the best result obtained is about 80.65%, i.e. below the best result on that dataset found in the literature.

1. Introduction

With the rapid expansion of the Internet, a huge amount of data from different sources has become available online. Studies indicate that in 2007 the amount of digital data scattered around the world consumed about 281 exabytes. In 2011, the amount of digital information produced in the year should be equal nearly 1800 exabytes, or 10 times that produced in 2006 [1].

Among all the different sources of information, music certainly is the one that can benefit from this impressive growing since it can be shared by users with different background and education, easily crossing cultural and language barriers [2]. In general, indexing and retrieving music is based on meta information tags such as ID3 tags. This metadata includes information such as song title, artist, album, year, musical genre, etc. [3]. Among all these descriptors, musical genre is probably the most obvious descriptor which comes to mind, and it is probably the most widely used to organize and manage large digital music databases [4].

Taking into account previous works, we can find different reasons which motivate research on automatic

* Corresponding author at: State University of Maringá (UEM), Av. Colombo, 5790 - Bloco C56, Maringá, PR 87020-900, Brazil. Tel.: +55 44 30114059; fax: +55 44 30115035.
E-mail addresses: yandre@din.uem.br (Y.M.G. Costa), lesoliveira@inf.ufpr.br (L.S. Oliveira), alekoe@ppgia.pucpr.br (A.L. Koerich), fgouyon@inescporto.pt (F. Gouyon), martins@utfpr.edu.br (J.G. Martins).

0165-1684/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2012.04.023

music genre classification. McKay and Fujinaga [5] pointed out that individuals differ on how they classify a given recording, but they can also differ in terms of the pool of genre labels from which they choose. On the other hand, Gjerdingen and Perrot [6] claimed that people are consistent in their genre categorization, even when these categorizations are wrong, or for very short segments. Pachet and Cazaly [7] showed that some traditional music taxonomies, like taxonomy of music industry, and internet taxonomy, are very inconsistent. Finally, Pampalk [8] says that genre classification-based evaluations can be used as proxy for listening tests of music similarity.

According to Lidy et al. [9] there are different approaches to describe the contents of a given piece of music. The most commonly used is the content-based approach which extracts representative features from the digital audio signal. Other approaches such as semantic analysis and community metadata have proved to perform well for traditional Western music, however, their use for other kinds of music is compromised because both community meta-data and lyrics-based approaches are dependent of natural language processing (NLP) tools, which are typically more developed for English than other languages.

In the case of the content-based approach, one of the earlier works was introduced by Tzanetakis and Cook [10] where they represented a music piece using timbral texture, beat-related, and pitch-related features. The employed feature set has become of public use, as part of the MARSYAS framework (Music Analysis, Retrieval and SYnthesis for Audio Signals), and it has been widely used for music genre recognition [3,9,11]. Other characteristics such as Inter-Onset Interval Histogram Coefficients, Rhythm Patterns and its derivatives Statistical Spectrum Descriptors, and Rhythm Histograms have been proposed in the literature recently [12–14].

In spite of all efforts done during the last years, automatic music genre classification still remains an open problem. McKay and Fujinaga [5] pointed out some problematic aspects of genre and refer to some experiments where human beings were not able to classify correctly more than 76% of the music pieces [15]. In spite of the fact that more experimental evidence is needed, these experiments give some insights about the upper bounds on software performance. McKay and Fujinaga also suggest that different approaches should be proposed to achieve further improvements.

In light of this, in this paper we propose an alternative approach for music genre classification which converts the audio signal into a spectrogram [16] (short-time Fourier representation) and then extract features from this visual representation. The rationale behind this is that treating the time-frequency representation as a texture image we can extract features which are expected to be suitable to build a robust music genre classification system even if there is not a straight relation between the musical dimension and the extracted features. Furthermore, these image-based features may capture different information from the approaches that work directly with the audio signal. Fig. 1 illustrates two examples of spectrograms taken from music pieces of different genres.

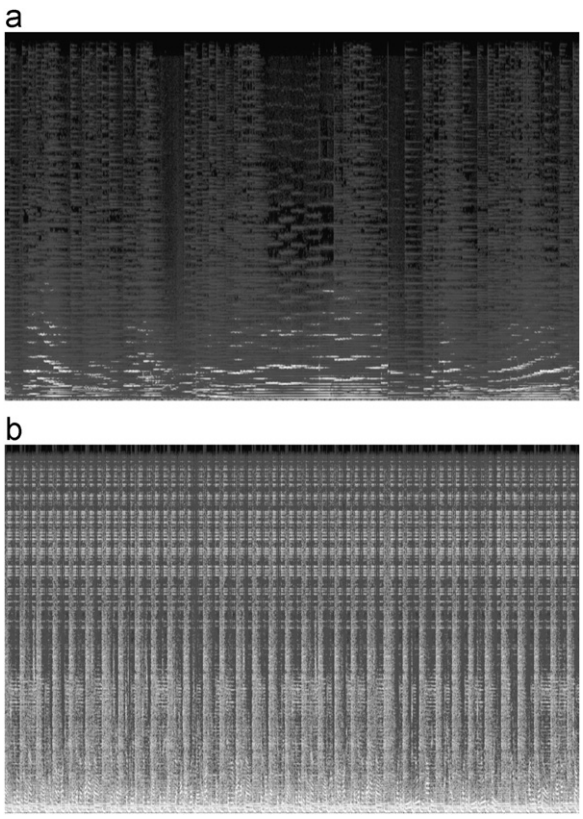

Fig. 1. Spectrogram examples. (a) Classical and (b) Electronic.

Fig. 1(a) shows a spectrogram taken from a classical music piece. In this case there is a very clear presence of almost horizontal lines, related to harmonic structures, while in Fig. 1(b) one can observe the intensive beats, typical of electronic music, depicted as clear vertical lines. The features used in this work are provided by Local Binary Pattern (LBP), a structural texture operator presented by Ojala et al. [17].

By analyzing the spectrogram images, we have noticed that the textures are not uniform, so we decided to consider a local feature extraction beyond the global feature extraction. Furthermore, our previous results [18] have shown that using Gray Level Co-occurrence Matrix (GLCM) descriptors, local feature extraction can help to improve outcomes in music genre classification using spectrograms. With this in mind, we have studied different zoning techniques to obtain local information of the given pattern beyond the global feature extraction. We also demonstrate through experimentation that certain zones of the spectrogram perform better than others.

The use of spectrograms in music genre classification has already been proposed in other works [18–20]. However, some important issues remain overlooked. Thus, some innovations are presented here, such as: the use of LBP structural approach in order to get texture descriptors from the spectrogram; zoning mechanism taking into account human perception in setting up frequency bands; creation of one individual classifier for each created zone, combining their outputs in order to get

the final decision; and comparison of results with and without zoning mechanism with a structural texture descriptor.

Through a set of comprehensive experiments on the Latin Music Database [21] and on the ISMIR 2004 database [22], we demonstrate that in most cases the proposed approach compares favorably to the traditional approaches reported in the literature. The results obtained with LMD in this work can be directly compared with those obtained by Lopes et al. [23] and Costa et al. [18], since all of them used artist filter restriction and folds with exactly the same music pieces to perform the classifier training and testing. The overall recognition rate improvement was about 22.66% when comparing with [23], and about 15.13% when comparing with the best result obtained in [18]. Taking into account the best results obtained with LMD in Music Information Retrieval Evaluation eXchange (MIREX) 2009 and MIREX 2010 [24] competitions, the improvement was about 7.67% and 2.47%, respectively. Concerning ISMIR 2004 database, obtained results are comparable to results described in the literature. In addition, these results can corroborate the versatility of the proposed approach.

The remaining of this paper is organized as follows: Section 2 describes the music databases used in the experiments. Section 3 presents the LBP texture operator used to extract features in this work. Section 4 introduces the methodology used for classification while Section 5 reports all the experiments that have been carried out on music genre classification. Finally the last section presents the conclusions of this work as well as opens up some perspectives for future work.

2. Music databases

The LMD and the ISMIR 2004 database are among the most used music database for researching in Music Information Retrieval. These two databases were chosen because, taking into account the signal segmentation strategy described in Section 3, these are among those databases that could be used.

2.1. Latin Music Database

The Latin Music Database (LMD) contains 3227 full-length music samples in MP3 format originated from music pieces of 501 artists [21]. The database is uniformly distributed along 10 music genres: Axé, Bachata, Bolero, Forró, Gaúcha, Merengue, Pagode, Salsa, Sertaneja, and Tango. One of the main characteristics of the LMD is the fact of bringing together many genres with a significant similarity among themselves with regard to instrumentation, rhythmic structure, and harmonic content. This happens because many genres present in the database are from the same country or countries with strong similarities regarding cultural aspects. Hence, the attempt to discriminate these genres automatically is particularly challenging.

In this database, music genre assignment was manually made by a group of human experts, based on the human perception on how each music is danced. The genre labeling was performed by two professional teachers with over ten years of experience in teaching ballroom Latin and Brazilian dances. The project team did a second verification in order to avoid mistakes.

In our experiments we have used 900 music pieces from the LMD, which are split into 3 folds of equal size (30 music pieces per class). The splitting is done using an artist filter [25], which places the music pieces of an specific artist exclusively in one, and only one, fold of the dataset. The use of the artist filter does not allow us to employ the whole dataset since the distribution of music pieces per artist is far from uniform. Furthermore, in our particular implementation of the artist filter we added the constraint of the same number of artists per fold. In order to compare the results obtained with other, the folds splitting taken was exactly the same used by Lopes et al. [23] and by Costa et al. [18]. It is worth of mention that the artist filter makes the classification task much more difficult. This database and experimental protocol has been used in the audio genre classification competition organized by the MIREX [24].

2.2. ISMIR 2004

The ISMIR 2004 database [22] was created by the Music Technology Group to support some tasks in Music Information Retrieval (MIR). This database became very popular in the MIR research community. It is composed of music pieces assigned to six different genres: classical, electronic, jazz/blues, metal/punk, rock/pop, and world. The distribution of music pieces per genre is not uniform, and the training and test sets are predefined. Thus, it is not possible to use artist filter with this dataset. Both training and test sets are originally composed of 729 music pieces, resulting in a total of 1458 music pieces in the dataset throughout.

Taking into account the signal segmentation strategy used in the experiments described here, it was not possible to use all the musics of this dataset. Instead of 729 music pieces originally assigned to the training set, it was possible to use only 711 pieces. Regarding to test set, instead of 729 music pieces, 713 were used.

3. Feature extraction

Since our approach is based on the visual representation of the audio signal, the first step of the feature extraction process consists in converting the audio signal to a spectrogram. In the LMD, the spectrograms were created using audio files with the following technical features: bit rate of 352 kbps, audio sample size of 16 bits, one channel, and audio sample rate of 22.05 kHz. In the ISMIR 2004 database, the audio files used had the following technical features: bit rate of 706 kbps, audio sample size of 16 bits, one channel, and audio sample rate of 44.1 kHz. In both cases, Discrete Fourier Transform was computed with a window size of 1024 samples using the Hanning window function which has good all-round frequency-resolution and dynamic-range properties.

In this work we have used the idea of time decomposition presented by Costa et al. [26] in which an audio signal S is decomposed into n different sub-signals.

Each sub-signal is simply a projection of S on the interval $[p,q]$ of samples, or $S_{pq} = \langle s_p, \ldots, s_q \rangle$. In the generic case, one may extract K (overlapping or non-overlapping) sub-signals and obtain a sequence of spectrograms $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_K$. We have used the strategy proposed by Silla et al. [21] which considers three 10-s segments from the beginning ($\overline{Y}_{beg}$), middle ($\overline{Y}_{mid}$), and end ($\overline{Y}_{end}$) parts of the original music. This process is depicted in Fig. 2.

After generating the spectrograms, the next step consists in extracting the features from the images. As stated before, the proposed approach considers the spectrogram as a texture and it uses the LBP operator to get features. Among the several structural techniques of texture representation, the LBP has been recently one of the most successful. Presented by Ojala et al. [17], LBP is a model that describes the texture taking into account for each pixel C, P neighbors equally spaced at a distance of R, as shown in Fig. 3.

An histogram h of LBPs found in the image is defined by the texture intensity differences of C and its P neighbors. As stated by Mäenpää and Pietikäinen [27], much of the information about the textural characteristics is preserved in the joint difference distribution (T) which is defined in Eq. (1)

$$T \approx (g_0 - g_C, \ldots, g_{P-1} - g_C) \tag{1}$$

where g_C is the gray level intensity of pixel C (the central pixel), and g_0 to g_{P-1} corresponds to the gray level intensities of the P neighbors. When the neighbors do not correspond to an image pixel integer value, its value is obtained by interpolation. An important characteristic of this descriptor is its invariance to changes in the value of the central pixels, when comparing with its neighbors.

Considering the resulting sign of the difference between C and each neighbor, as denoted in Eq. (2), it is defined that: if the sign is positive the result is 1, otherwise 0 as denoted in Eq. (3). Thus, it is possible to obtain this invariance of the intensity value of pixels in gray-scale format

$$T \approx (s(g_0 - g_C), \ldots, s(g_{P-1} - g_C)) \tag{2}$$

where

$$s(g_i - g_C) = \begin{cases} 1 & \text{if } g_i - g_C \geq 0 \\ 0 & \text{if } g_i - g_C < 0 \end{cases} \tag{3}$$

where $i = [0,P]$ is the index of the neighbors of C.

With this, the LBP value can be obtained by multiplying the binary elements for a binomial coefficient. Assigning a binomial weight 2^P to each sign $s(g_P - g_C)$, the differences in a neighborhood are transformed into a unique LBP code, a value $0 \leq C' \leq 2^P$. Eq. (4) describe how this code is obtained

$$\mathrm{LBP}_{P,R}(x_C, y_C) = \sum_{P=0}^{P-1} s(g_P - g_C) 2^P \tag{4}$$

assuming that $x_C \in \{0, \ldots, N-1\}$, $y_C \in \{0, \ldots, M-1\}$ for a $N \times M$ image sample.

Observing the non-uniformity of the vector obtained, Ojala et al. [17] introduced a concept based on the transition between 0's and 1's in the LBP image. A binary LBP code is considered uniform if the number of transitions is less than or equal to 2, also considering that the code is seen as a circular list. That is, the code 00100100 is not considered uniform, because it contains four transitions. But the code 00100000 is characterized as uniform because it has only two transitions. Fig. 4 illustrates this idea.

Therefore, instead of using the whole histogram, which size is 2^P, it is possible to use only the uniform values, constituting a low-dimensional feature vector, with only 59 features. Ojala et al. [17] stated that, beyond the

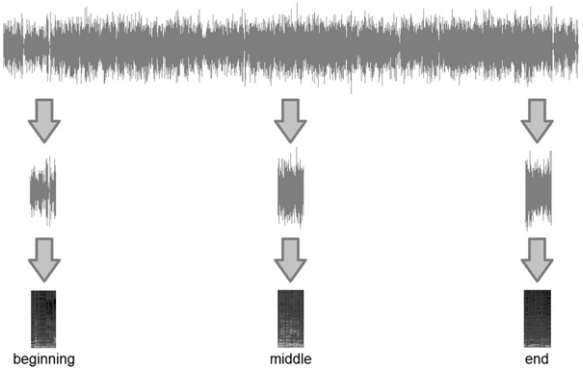

Fig. 2. Creating spectrograms using time decomposition.

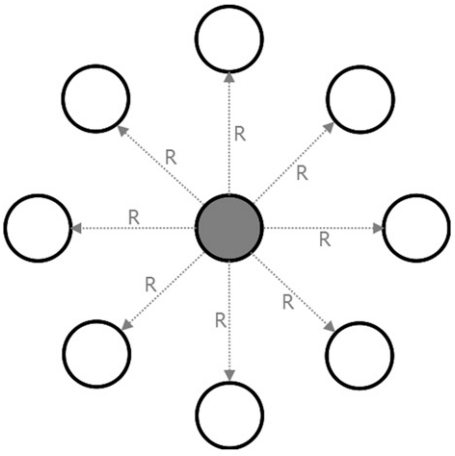

Fig. 3. The LBP operator. A pixel C, dark circle in the middle, and its P neighbors, lighter circles.

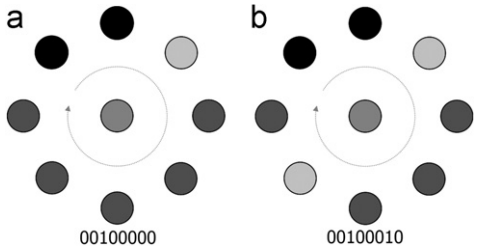

Fig. 4. LBP uniform pattern. (a) The two transitions showed identifies the pattern as uniform. (b) With four transitions, it is not considered a uniform pattern.

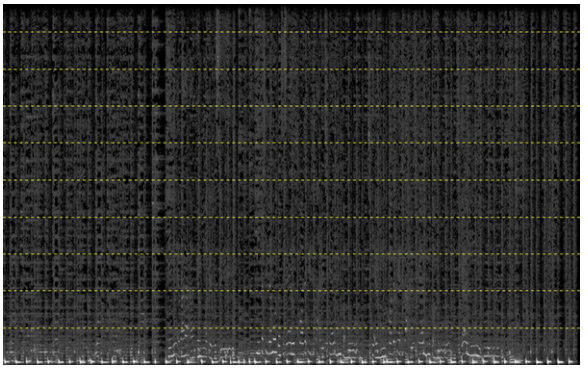

Fig. 5. A zoning mechanism used to extract local information.

58 possible uniform combinations, all the non-uniform patterns should be placed into an extra bin of the histogram. This version of the descriptor was called "u2", a label accompanying the values of the radius R and the neighborhood size P, making the LBP definition as follows: $\mathrm{LBP}_{P,R}^{label}$.

Furthermore, we observed during the experiments that the feature extraction with $\mathrm{LBP}_{8,2}^{u2}$ is fast and accurate enough for the proposed application. The R value is related to the spatial image resolution. Once changing the R value could turn the feature extraction more expensive in terms of time-consuming, we performed some preliminary experiments changing the spectrogram images resolution in which we could observe that $R=2$ with $P=8$ would produce a good cost-effective.

At this point we could have a piece of music represented by three 59-dimensional feature vectors. However, by analyzing the texture images, we have noticed that the texture produced by the spectrograms are not uniform. Furthermore, previous results described in [18] suggest that spectrogram image zoning, in order to preserve local feature, could help to achieve good results. Therefore, it is important to consider a local feature extraction beyond a global one.

With this in mind, we have used some different zoning techniques which are a simple but efficient way of obtaining local information of a given pattern. The idea consists in dividing the spectrogram into n parts as depicted in Fig. 5. In this example, for $n=10$, each spectrogram image was linearly divided and would be represented by ten 59-dimensional feature vectors, summing up 30 vectors for a music piece.

In the next section we show some details about the different zoning schemes that have been used and how the feature vectors are used for training and classification.

4. Methodology used for classification

The classifier used in this work was the Support Vector Machine (SVM) introduced by Vapnik in [28]. Normalization was performed by linearly scaling each attribute to the range $[-1,+1]$. Different parameters and kernels for the SVM were tried out but the best results were achieved using a Gaussian kernel. Parameters cost and gamma were tuned using a grid search.

The classification process is done as follows: the three 10-s segments of the music are converted into the spectrograms ($\overline{Y}_{beg}$, $\overline{Y}_{mid}$, and $\overline{Y}_{end}$). Each of them is divided into n zones, according to the zoning mechanism used, and one feature vector is extracted from each zone. One classifier for each spectrogram zone is created. Then, the 59-dimensional feature vector extracted from each zone is sent to a specific classifier which assigns a prediction to each one of the possible classes. In experiments with the LMD, training and classification are carried out using a threefold cross-validation procedure: two folds used for training a N-class SVM classifier, one fold for testing, three permutations of the training fold (i.e. $1+2$, $1+3$, $2+3$). In each case, $3n$ classifiers are created with 600 and 300 feature vectors for training and testing, respectively, where n is the number of zones. On the other hand, in the ISMIR 2004 database, there are predefined training and test sets. In both cases, three different zoning schemes are used, beyond the global feature extraction. For global feature extraction, n is 1.

With this amount of classifiers to deal with the classification of a single music piece, we used estimation of probabilities to combine the outputs of such classifiers and reach a final decision (Fig. 6). In this situation, it is very useful to have a classifier producing *a posterior* probability $P(class|input)$. Here, we are interested in the estimation of probabilities because different fusion strategies like *Max*, *Min*, *Product*, and *Sum* will be tried out. The following equations, presented by Kittler et al. [29], describe on how the classifier outputs are combined with these four decision rules to reach a final decision

$$Max\ Rule\ (v) = \arg\max_{k=1}^{c} \max_{i=1}^{n} P(\omega_k|l_i(v)) \tag{5}$$

$$Min\ Rule\ (v) = \arg\max_{k=1}^{c} \min_{i=1}^{n} P(\omega_k|l_i(v)) \tag{6}$$

$$Product\ Rule\ (v) = \arg\max_{k=1}^{c} \prod_{i=1}^{n} P(\omega_k|l_i(v)) \tag{7}$$

$$Sum\ Rule\ (v) = \arg\max_{k=1}^{c} \sum_{i=1}^{n} P(\omega_k|l_i(v)) \tag{8}$$

where v represents the pattern to be classified, n is the number of classifiers, l_i represents the output label of the ith classifier in a problem in which the possible class labels are $\Omega=\omega_1,\omega_2,\ldots,\omega_c$, c is the number of classes, and $P(\omega_k|y_i(v))$ is the estimation of probability of pattern v belonging to class ω_k according to the ith classifier.

The rationale behind the zoning and combining scheme is that music signals may include similar instruments and

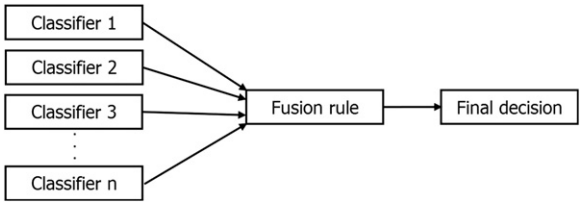

Fig. 6. Combination of classifier outputs to obtain a final decision.

similar rhythmic patterns which leads to similar areas with similar intensities in the spectrogram images. By zoning the images we can extract local information and try to highlight the specificities of each music genre. In Fig. 7 we can notice that at low frequencies the textures are quite similar but they start to become different as the frequency increases. The opposite can happen as well, and for this reason, the zoning mechanism becomes an interesting alternative. As stated before, in this work we have investigated three different zoning strategies, which are based on different scales, beyond the global feature extraction. The following subsections describe these zoning strategies.

4.1. Linear zoning

In this zoning scheme, the spectrogram image is divided into 10 linear zones of equal size. Fig. 8 shows the division used with the LMD. It is important to mention that we have tried out different configurations of linear zoning. Results showed that after ten zones there is no improvement in terms of classification rate. The frequency upper limit of the spectrograms generated for the LMD and the ISMIR 2004 database were not the same because some music pieces of the LMD, specially of Tango genre, refer to very old recordings with no relevant information above 8.5 kHz. Thus, for this database only information up to 8.5 kHz was considered. On the other hand, the music pieces from the ISMIR 2004 database present relevant information for all music pieces from all genres up to 14 kHz. Thus, this was the frequency upper limit used for this database.

Table 1 shows the limits of each one of the 10 stated frequency bands in linear division for the LMD, while Table 2 shows the limits used with the ISMIR 2004 database. In this case, 10 different classifiers are created for each spectrogram generated from a segment of a music piece. Thereby, there are 30 classifiers whose outputs are combined to get a final decision about the music genre.

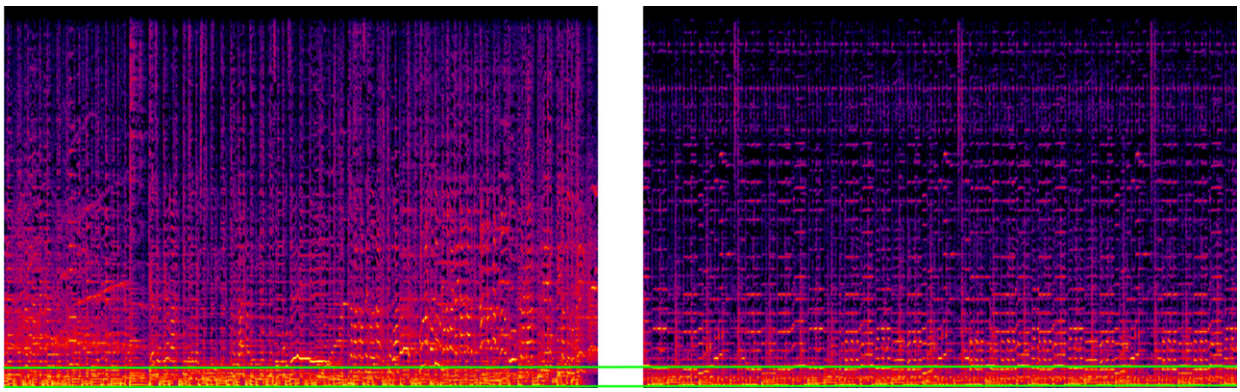

Fig. 7. Spectrograms of different music genres with some areas of similarity at low frequencies.

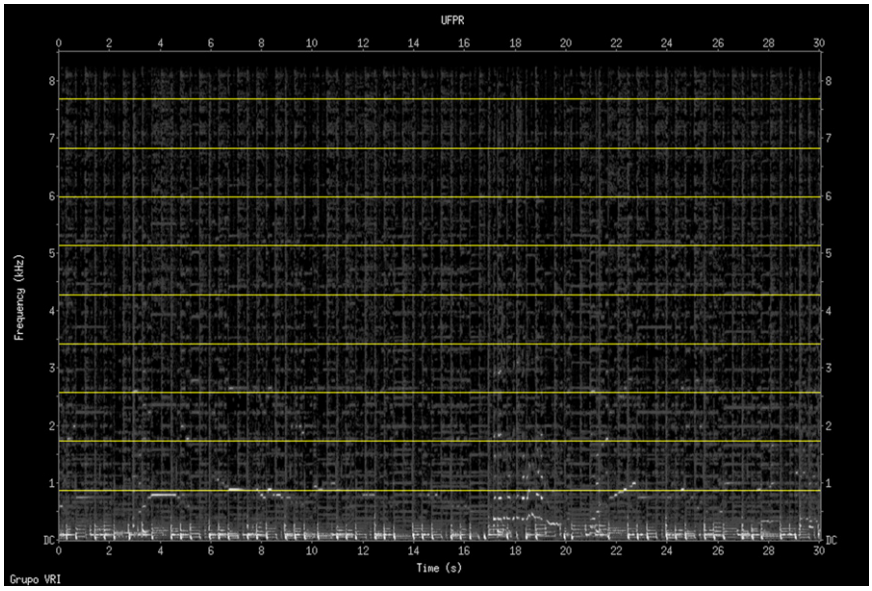

Fig. 8. Linear zoning used to extract local information.

Table 1
Frequency bands limits, in Hz, for linear zoning with LMD.

1	2	3	4	5	6	7	8	9	10
0–850	850–1700	1700–2550	2550–3400	3400–4250	4250–5100	5100–5950	5950–6800	6800–7650	7650–8500

Table 2
Frequency bands limits, in Hz, for linear zoning with ISMIR 2004.

1	2	3	4	5	6	7	8	9	10
0–1400	1400–2800	2800–4200	4200–5600	5600–7000	7000–8400	8400–9800	9800–11,200	11,200–12,600	12,600–14,000

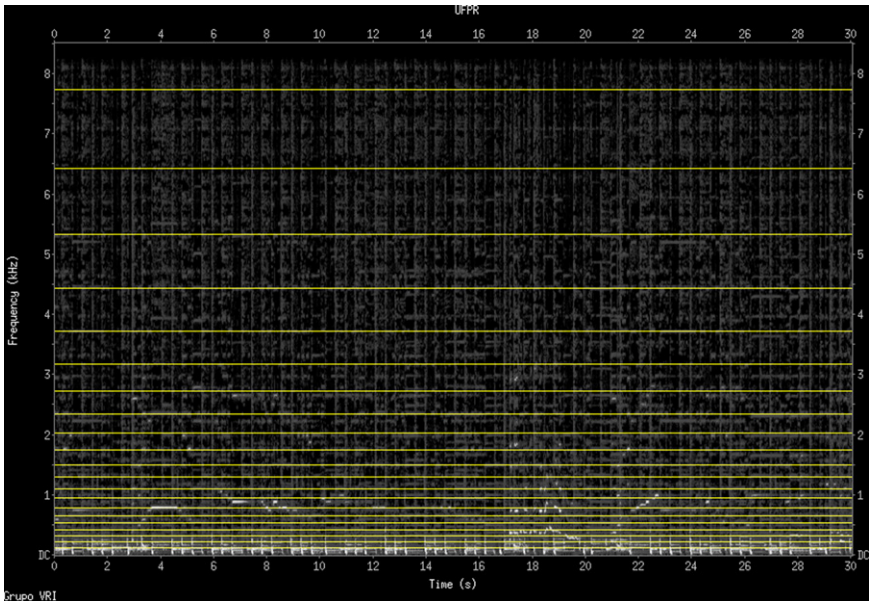

Fig. 9. Bark scale zoning used to extract local information.

Table 3
Frequency bands limits, in Hz, for Bark scale zoning.

1	2	3	4	5	6	7	8
0–100	100–200	200–300	300–400	400–510	510–630	630–770	770–920
9	10	11	12	13	14	15	16
920–1080	1080–1270	1270–1480	1480–1720	1720–2000	2000–2320	2320–2700	2700–3150
17	18	19	20	21	22	23	24
3150–3700	3700–4400	4400–5300	5300–6400	6400–7700	7700–9500	9500–12,000	12,000–14,000

4.2. Bark scale zoning

The Bark scale is a subdivision of the audible frequency range into critical bands [30]. It was created in an attempt of representing the subdivision of the frequency range over which the human ear is able to perceive tones and noises. This scale is not linear, therefore the size of the frequency bands may be different. Fig. 9 shows this division superimposed over a spectrogram image generated from a music piece of the LMD.

Table 3 shows the limits of each of the 24 stated frequency bands in Bark scale division used in the experiments describe here. Note that for the LMD only 22 zones are created for each spectrogram, since the upper frequency limit of these spectrograms was defined as 8.5 kHz. For the ISMIR 2004 database, 24 zones per spectrogram are created. Since we have 3 spectrograms for a music piece, there are 66 classifiers for the LMD and 72 classifiers for the ISMIR 2004 database whose outputs are combined to get a final decision about the music genre.

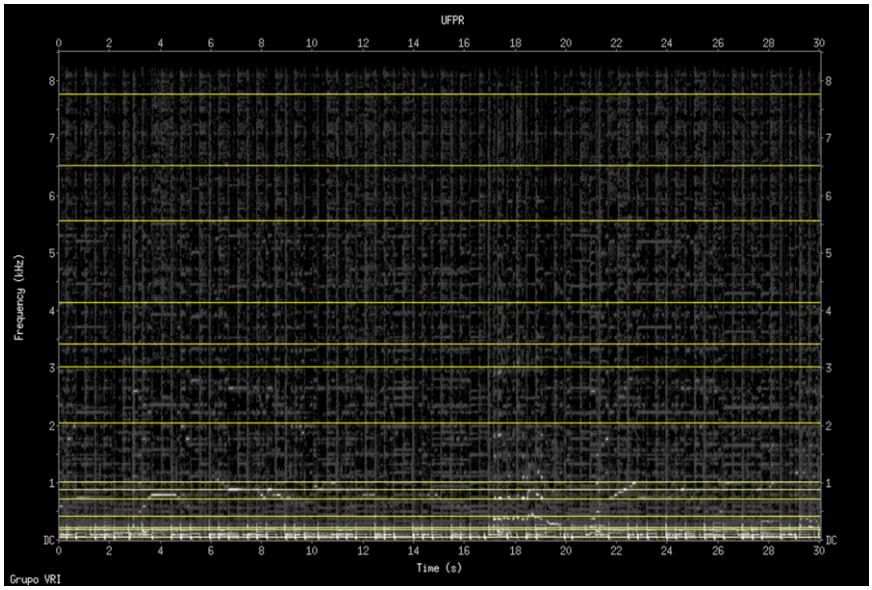

Fig. 10. Mel scale zoning used to extract local information.

Table 4
Frequency bands limits, in Hz, for Mel scale zoning.

1	2	3	4	5	6	7	8
0–40	40–161	161–200	200–404	404–693	693–867	867–1000	1000–2022
9	10	11	12	13	14	15	
2022–3000	3000–3393	3393–4109	4109–5526	5526–6500	6500–7743	7743–14,000	

4.3. Mel scale zoning

According to Umesh et al. [31], the Mel scale is a fundamental result of psychoacoustics, relating real frequency to perceived frequency. Like the Bark scale, the Mel scale attempts to represent the frequency bands according to the human perception. Fig. 10 shows this division superimposed over a spectrogram image taken from a music piece of the LMD. As we can see, this scale is not linear as well.

Table 4 shows the limits of each one of the 15 stated frequency bands in Mel scale division. For the LMD, the frequency upper limit in the 15th band is 8.5 kHz. With this zoning scheme, 15 different classifiers are created for each spectrogram generated from a segment of a music piece. Thereby, there are 45 classifiers whose outputs are combined to get a final decision about the music genre.

5. Experimental results and discussion

The following subsections present the experiments carried out with the global feature extraction and with the three different zoning mechanisms proposed in the previous section. Additional experiments carried out using acoustic features are also presented. The experimental results reported on the LMD refers to the average classification rates and standard deviations considering the three folds aforementioned.

5.1. Results with global feature extraction

In this subsection we describe the results obtained without performing a zoning mechanism. Our previous work have shown that zoning is very important when used with GLCM features [18]. Therefore, here we want to investigate if the same holds for LBP features. In this case, three different classifiers are created, since we are dealing with three sub-signals taken from the original signal, as depicted in Fig. 2. Then, the outputs of these classifiers were combined with those four different combination rules described in Section 4. Table 5 shows results obtained with LMD, while results obtained with ISMIR 2004 are shown in Table 6.

The results obtained with the LMD show that LBP features performed much better than GLCM features when no zoning strategy is used. In the best case, with sum combination rule, the performance surpassed the best result obtained in [18] about 12 percentage points. In addition, the best result is very close to the best result obtained in MIREX 2010 competition.

Taking into account that the result presented in this subsection is the best one achieved in our experiments

with the ISMIR 2004 dataset, Table 7 shows the confusion matrix produced with the product rule in order to get a better understanding about the results achieved so far.

Table 5
Average recognition rates with global feature extraction in LMD.

Genre	Max rule Rec. rate (%)	Min rule Rec. rate (%)	Product rule Rec. rate (%)	Sum rule Rec. rate (%)
Axé	72.22	71.11	75.56	**76.67**
Bachata	**93.33**	91.11	92.22	92.22
Bolero	78.89	81.11	**83.33**	81.11
Forró	**77.78**	63.33	66.67	73.33
Gaúcha	57.78	**64.44**	**64.44**	**64.44**
Merengue	93.33	90.00	93.33	**94.44**
Pagode	66.67	**76.67**	75.56	73.33
Salsa	83.33	82.22	83.33	**85.56**
Sertaneja	54.44	60.00	**62.22**	58.89
Tango	90.00	**91.11**	90.00	90.00
Overall	76.78 ± 0.38	77.11 ± 1.71	78.67 ± 0.67	**79.00** ± 1.00

Table 6
Recognition rates with global feature extraction in ISMIR 2004.

Genre	Max rule Rec. rate (%)	Min rule Rec. rate (%)	Product rule Rec. rate (%)	Sum rule Rec. rate (%)
Classical	97.06	96.73	**97.39**	**97.39**
Electronic	85.84	84.07	**89.38**	88.50
Jazz/blues	**38.46**	30.77	**38.46**	**38.46**
Metal/punk	**44.44**	35.56	35.56	37.78
Rock/pop	60.78	63.73	**69.61**	68.63
World	54.55	**68.60**	65.29	61.16
Overall	77.42	78.96	**80.65**	79.80

Table 7
Confusion matrix (%) with global feature extraction in ISMIR 2004.

Genre	(0)	(1)	(2)	(3)	(4)	(5)
(0) Classical	**97.39**	0.33	0.00	0.00	0.33	1.96
(1) Electronic	0.88	**89.38**	0.00	1.77	3.54	4.42
(2) Jazz/blues	15.38	7.69	**38.46**	0.00	3.85	34.62
(3) Metal/punk	0.00	11.11	0.00	**35.56**	46.67	6.67
(4) Rock/pop	2.94	12.75	0.00	0.98	**69.61**	13.73
(5) World	20.66	8.26	0.00	0.00	5.79	**65.29**

Although the best overall recognition rate obtained with the ISMIR 2004 dataset is slightly better than that obtained in the LMD, one can note that the variation in the recognition rates among the classes is very high for the ISMIR 2004 dataset. It is important to recall that the number of music pieces per genre in the ISMIR 2004 dataset is far from uniform.

5.2. Results with linear zoning

Table 8 shows the recognition rates obtained individually for each one of the 30 zones created with the linear zoning described in Section 4.1.

After training one classifier for each zone, their outputs are combined with max, min, product, and sum rules. The results in terms of recognition rates obtained on the LMD are shown in Table 9, while the results obtained in the ISMIR 2004 database are shown in Table 10.

In terms of recognition rate, the linear zoning scheme did not provide better results than those obtained with global feature extraction. However, in order to verify the potential of the created pool of classifiers in future work, using dynamic classifier selection, we decided to check the upper limit between the created classifiers. The possible upper limit of classification accuracy is defined as the ratio of samples which are classified correctly by at least one classifier in a pool for all samples. In the LMD,

Table 9
Average recognition rates (%) combining all zones of linear zoning with different rules in LMD.

Genre	Max rule Rec. rate (%)	Min rule Rec. rate (%)	Product rule Rec. rate (%)	Sum rule Rec. rate (%)
Axé	70.00	67.78	76.67	**80.00**
Bachata	**96.67**	87.78	92.22	94.44
Bolero	75.56	75.56	**87.78**	**87.78**
Forró	72.22	54.44	75.56	**76.67**
Gaúcha	43.33	**66.67**	57.78	54.44
Merengue	**94.44**	87.78	93.33	93.33
Pagode	33.33	56.67	**60.00**	54.44
Salsa	84.44	83.33	**86.67**	85.56
Sertaneja	50.00	**61.11**	57.78	56.67
Tango	**95.56**	84.44	90.00	92.22
Overall	71.56 ± 1.26	72.56 ± 2.67	**77.78** ± 0.38	77.56 ± 1.17

Table 8
Recognition rates (%) obtained for each zone created with linear zoning.

Frequency band id.[a]	LMD			ISMIR 2004		
	Beginning segment	Middle segment	End segment	Beginning segment	Middle segment	End segment
10	48.56	56.56	49.89	64.10	69.99	59.75
9	56.44	57.67	52.22	62.27	71.11	63.39
8	55.89	55.33	51.89	63.11	71.39	63.25
7	56.33	59.00	53.78	66.62	70.97	64.10
6	55.44	61.00	59.33	63.53	73.35	63.53
5	53.56	55.89	53.44	66.76	73.49	63.11
4	54.22	53.89	56.33	68.72	70.83	66.62
3	56.00	51.67	52.22	70.55	70.97	67.04
2	52.89	56.67	54.44	68.02	71.67	67.04
1	54.00	61.67	64.78	70.55	72.79	68.16

[a] According to Tables 1 and 2.

the overall upper limit rate is equal to 98.67%, while in the ISMIR 2004 database this rate is equal to 98.46%, which indicates that even if the linear zoning did not provide better results than the global approach, there is a room for improvement if more sophisticated combination rules are employed. Using a dynamic classifier selection technique, one can select a specific ensemble of classifiers taking into account porperties of the specific feature vector to be classified, more details can be found in [32].

5.3. Results with Bark scale zoning

With Bark scale zoning, 66 classifiers were created for the LMD and 72 classifiers were created for the ISMIR 2004 database (as described in Section 4.2). Table 11 shows the recognition rates obtained individually for each created zone.

Table 10
Recognition rates (%) combining all zones of linear zoning with different rules in ISMIR 2004.

Genre	Max rule Rec. rate (%)	Min rule Rec. rate (%)	Product rule Rec. rate (%)	Sum rule Rec. rate (%)
Classical	**99.67**	97.71	98.69	99.02
Electronic	80.53	72.57	**89.38**	88.50
Jazz/blues	30.77	30.77	**38.46**	34.62
Metal/punk	**35.56**	17.78	20.00	26.67
Rock/pop	52.94	62.75	**70.59**	**70.59**
World	27.27	**66.94**	53.72	46.28
Overall	71.11	76.02	**78.40**	77.42

After training one classifier for each zone, their outputs were combined with the combination rules described in Section 4. The results obtained on the LMD and on the ISMIR 2004 database are presented, respectively, in Tables 12 and 13.

In general, Bark scale zoning did not present good results when faced with global feature extraction, both on the LMD and on the ISMIR 2004 database. Aiming to foresee the potential of using dynamic classifier selection in future works, we decided to check the upper limit between the classifiers created with Bark scale zoning. Using Bark scale zoning, the upper limit rates on the LMD and on the ISMIR 2004 database presented the best performance among all the zoning schemes experimented, probably favored by the larger number of classifiers.

Table 12
Average recognition rates (%) in LMD with Bark scale zoning using different rules.

Genre	Max rule Rec. rate (%)	Min rule Rec. rate (%)	Product rule Rec. rate (%)	Sum rule Rec. rate (%)
Axé	64.44	56.67	**95.56**	82.22
Bachata	92.22	80.00	81.11	**94.44**
Bolero	68.89	65.56	64.44	**91.11**
Forró	60.00	51.11	32.22	**75.56**
Gaúcha	24.44	**67.78**	21.11	50.00
Merengue	**96.67**	87.78	77.78	95.56
Pagode	35.56	56.67	10.00	**57.78**
Salsa	77.78	78.89	57.78	**85.56**
Sertaneja	50.00	51.11	34.44	**60.00**
Tango	**94.44**	81.11	82.22	87.78
Overall	66.44 ± 1.07	67.67 ± 1.00	55.67 ± 1.73	**78.00** ± 1.33

Table 11
Recognition rates (%) obtained for each zone with Bark scale zoning.

Frequency band id.[a]	LMD			ISMIR 2004		
	Beginning segment	Middle segment	End segment	Beginning segment	Middle segment	End segment
24	–	–	–	62.83	70.69	61.29
23	–	–	–	65.08	71.53	64.80
22	47.67	55.56	48.44	63.96	71.53	63.96
21	56.00	62.00	53.33	67.18	72.93	62.97
20	57.44	59.39	54.44	66.06	70.97	61.85
19	56.67	61.00	57.67	67.60	69.28	62.13
18	52.67	57.78	53.56	64.66	67.74	64.52
17	53.67	50.89	49.00	66.48	68.16	59.33
16	50.89	50.89	49.44	68.02	65.92	61.85
15	51.44	48.11	49.56	63.53	68.44	61.15
14	45.33	46.78	46.56	61.43	66.90	59.19
13	47.00	45.22	45.22	63.53	66.62	59.05
12	44.00	45.33	44.11	61.01	64.10	59.19
11	43.33	42.78	40.67	64.66	64.52	56.10
10	45.22	42.11	42.78	61.57	62.41	54.56
9	42.22	40.67	38.56	59.89	62.13	57.92
8	41.67	37.67	44.00	59.89	60.31	56.24
7	37.11	39.22	41.22	57.78	58.06	54.84
6	36.67	35.56	35.22	55.96	58.49	54.70
5	36.44	38.67	35.56	56.66	56.80	56.38
4	36.78	39.00	37.22	57.08	58.77	55.12
3	36.67	44.22	42.44	57.78	62.27	54.98
2	45.33	51.00	50.67	61.99	61.43	57.64
1	41.78	50.33	49.22	62.41	64.66	61.71

[a] According to Table 3.

Table 13
Recognition rates (%) combining all zones of Bark scale zoning with different rules in ISMIR 2004.

Genre	Max rule Rec. rate (%)	Min rule Rec. rate (%)	Product rule Rec. rate (%)	Sum rule Rec. rate (%)
Classical	**100.00**	96.08	99.67	99.35
Electronic	68.14	60.18	76.99	**88.50**
Jazz/blues	**15.38**	3.85	0.00	7.69
Metal/punk	**51.11**	0.00	6.67	11.11
Rock/pop	41.18	63.73	57.84	**64.71**
World	7.44	**67.77**	30.58	22.31
Overall	64.66	**71.53**	68.86	70.69

Table 14
Upper limit recognition rates (%) between the classifiers created with Bark scale zoning in LMD.

Genre	Upper limit
Axé	100.00
Bachata	100.00
Bolero	100.00
Forró	100.00
Gaúcha	98.89
Merengue	100.00
Pagode	100.00
Salsa	100.00
Sertaneja	100.00
Tango	100.00
Overall	**99.89**

Table 15
Upper limit recognition rates (%) between the classifiers created with Bark scale zoning in ISMIR 2004.

Genre	Upper limit
Classical	100.00
Electronic	100.00
Jazz/blues	84.62
Metal/punk	100.00
Rock/pop	99.02
World	100.00
Overall	**99.30**

Some details about these rates are shown in Tables 14 and 15, respectively.

5.4. Results with Mel scale zoning

In this case, 15 zones were created, as described in Section 4.3. Table 16 shows the recognition rates obtained individually for each zone.

The results obtained on the LMD and on the ISMIR 2004 database proceeding the fusion of the classifiers outputs with the four rules aforementioned are shown in Tables 17 and 18, respectively.

The zoning based in Mel scale provided the best overall performance in the LMD. The recognition rate improvement is about 3.33% when facing the best result with that obtained with global feature extraction, which is the second best obtained result. One can note that, considering the standard deviations taken from results obtained individually in each fold, the recognition rate obtained with Mel scale is quite promising, even implying in a greater number of classifiers than the global and linear zoning approaches.

In order to take a better idea about the classifier with best performance on the LMD, Table 19 shows the confusion matrix. We can observe that the music genres Gaúcha and Sertaneja present the worst recognition rates. Beyond being originally from the same country (Brazil), on can notice that in most of the cases such genres are confused with another Brazilian genre (Axé). As aforementioned, commonly genres from the same country present high similarity among themselves, what can explain such confusions.

The upper limit found between the classifiers created with Mel scale zoning for the LMD is equal to 99.78% in the LMD and 99.16% in the ISMIR 2004 database. These rates are very close to the best ones, obtained with Bark scale zoning.

5.5. Results with acoustic features

In order to compare the performance of the features presented here with some well-known acoustic features, we performed over the same datasets used in this work some experiments with the following features: Spectral Centroid, Roll-Off, Flux, Zero Crossing, and 13 Mel-Frequency Cepstral Coeficients (MFCCs). The features were extracted with the framework MARSYAS, more details about these features can be found in [10]. The obtained results are shown in Table 20.

One can observe that these features presented a significantly worse result on the LMD. Tables 21 and 22 show the confusion matrices generated in the classification with acoustic features on the LMD and on the ISMIR 2004 database, respectively. These matrices can help to take a good comprehension about the complementarity between the classifiers constructed with acoustic and visual features. The difference between the performance of the acoustic features in the LMD and the ISMIR 2004 datasets can be explained by the fact that the distribution of music pieces per genre in ISMIR 2004 is far from uniform. In this dataset, the genre which presented the best individual performance (classical) has much more music pieces than the others.

5.6. Discussion

Table 23 presents the best results obtained in this work considering all the zoning schemes proposed in this work. Furthermore, the experimental results with the acoustic features are also include here.

The Friedman test with the post hoc Shaffer's static procedure was employed to evaluate if there are statistically significant differences between the results originated from the different zoning schemes with LBP as well as the acoustic features. Since only the LMD was split into folds, we only performed the test on this database. For this purpose, we took the best result, in terms of fusion rule, obtained for each experimented zoning

Table 16
Recognition rates (%) obtained for each zone with Mel scale zoning.

Frequency band id.[a]	LMD			ISMIR 2004		
	Beginning segment	Middle segment	End segment	Beginning segment	Middle segment	End segment
15	48.89	54.00	49.00	68.86	73.63	66.06
14	55.11	59.67	55.00	68.16	72.37	62.41
13	56.89	57.11	52.33	67.74	73.21	64.10
12	58.44	63.22	61.00	71.11	70.13	62.97
11	55.11	51.67	51.22	68.30	68.58	63.81
10	51.44	48.11	50.67	64.52	67.46	57.78
9	57.44	55.22	55.11	67.88	70.83	65.08
8	54.00	58.56	57.44	68.30	68.44	65.36
7	42.56	38.22	41.00	59.89	61.15	56.94
6	39.67	42.11	43.78	59.47	61.15	56.80
5	45.11	51.00	49.56	59.75	63.81	59.61
4	44.44	48.44	47.11	60.17	65.22	60.03
3	36.78	41.11	41.33	58.63	62.97	57.08
2	48.00	57.89	56.78	60.31	65.92	62.55
1	35.78	43.44	39.56	59.47	64.38	59.89

[a] According to Table 4.

Table 17
Average recognition rates (%) in LMD using Mel scale zoning with different rules.

Genre	Max rule Rec. rate (%)	Min rule Rec. rate (%)	Product rule Rec. rate (%)	Sum rule Rec. rate (%)
Axé	65.56	64.44	83.33	**84.44**
Bachata	**95.56**	86.67	93.33	**95.56**
Bolero	75.56	65.56	91.11	**92.22**
Forró	73.33	58.89	82.22	**83.33**
Gaúcha	34.44	**71.11**	67.78	55.56
Merengue	**95.56**	87.78	**95.56**	**95.56**
Pagode	45.56	60.00	**71.11**	65.56
Salsa	**84.44**	80.00	**84.44**	**84.44**
Sertaneja	56.67	51.11	**67.78**	63.33
Tango	**96.67**	86.67	86.67	91.11
Overall	72.33 ± 3.33	71.22 ± 2.34	**82.33** ± 1.45	81.11 ± 1.35

Table 18
Recognition rates (%) using Mel scale zoning with different rules in ISMIR 2004.

Genre	Max rule Rec. rate (%)	Min rule Rec. rate (%)	Product rule Rec. rate (%)	Sum rule Rec. rate (%)
Classical	**99.67**	97.71	99.35	**99.67**
Electronic	73.45	67.26	88.50	**92.04**
Jazz/blues	**23.08**	**23.08**	19.23	19.23
Metal/punk	**33.33**	17.78	11.11	13.33
Rock/pop	49.02	65.69	**68.63**	64.71
World	**73.55**	17.33	46.28	33.88
Overall	67.32	76.44	**76.74**	73.91

scheme. The multiple comparison statistical test has show that the p value of the statistical test was lower than the corrected critical value only between the Mel scale zoning and the acoustic features, showing a statistically significant difference between these approaches at 95% confidence level. Furthermore, the statistical tests have shown that there is no statistically significant difference between the results obtained from the different zoning schemes.

Regarding to recognition rates, the use of zoning scheme in order to preserve local feature extraction did not perform as expected in most of cases. The only case in which the zoning scheme performed better than the global one occurred when Mel scale division was used in the experiments with the LMD. In this case, even considering the standard deviation between the three folds used in classification, the Mel scale zoning remains with the best result. Taking into account previous results in [18], it was expected that zoning would provide better results. However, obtained results suggest that the texture features captured by LBP from spectrogram images taken from music pieces of the same genre change less from zone to zone than when GLCM is used.

Also related to zoning scheme, it is worthy of mention that the upper limits provided by the pool of classifiers produced in all the zoning schemes experimented here are very promising. It would be very interesting to perform future works taking into account the use of some dynamic classifier/ensemble selection over these pool of classifiers.

Some results of the experiments described in this work are very promising when compared with the state of the art. Table 24 shows the best result obtained here with the LMD and the best results reported recently on the literature with this database. In order to proceed a fair comparison, it was chosen some works in which the LMD was used with artist filter. In some cases [23,18] the results refer to experiments developed with exactly the same folds, composed of the same music pieces. In 2010, Lopes et al. [23] presented an approach based on an instance selection method, where a music piece was represented by 646 instances. The instances consist of feature vectors representing short-term, low-level characteristics of music audio signal. The classifier used was an SVM and the final decision was done through majority voting. In 2011, Costa et al. [18] presented a classification scheme similar to the one presented in this work, based

Table 19
Confusion matrix (%) of the best case considering the Mel scale zoning in LMD.

Genre	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(0) Axé	**83.33**	0.00	1.11	0.00	0.00	0.00	6.67	3.33	5.56	0.00
(1) Bachata	2.22	**93.33**	2.22	1.11	0.00	0.00	0.00	1.11	0.00	0.00
(2) Bolero	1.11	0.00	**91.11**	1.11	0.00	0.00	0.00	1.11	4.44	1.11
(3) Forró	2.22	1.11	4.44	**82.22**	6.67	0.00	0.00	2.22	1.11	0.00
(4) Gaúcha	14.44	0.00	6.67	6.67	**67.78**	0.00	0.00	0.00	4.44	0.00
(5) Merengue	0.00	3.33	1.11	0.00	0.00	**95.56**	0.00	0.00	0.00	0.00
(6) Pagode	6.67	0.00	11.11	0.00	1.11	0.00	**71.11**	3.33	6.67	0.00
(7) Salsa	7.78	0.00	4.44	0.00	3.33	0.00	0.00	**84.44**	0.00	0.00
(8) Sertaneja	11.11	1.11	8.89	1.11	7.78	0.00	2.22	0.00	**67.78**	0.00
(9) Tango	1.11	0.00	11.11	1.11	0.00	0.00	0.00	0.00	0.00	**86.67**

Table 20
Recognition rates (%) obtained with acoustic features.

LMD	**Rec. rate**	**ISMIR 2004**	**Rec. rate**
Axé	57.78	Classical	95.42
Bachata	85.56	Electronic	76.11
Bolero	63.33	Jazz/blues	50.00
Forró	38.89	Metal/punk	51.11
Gaúcha	51.11	Rock/pop	59.80
Merengue	78.89	World	46.28
Pagode	46.67		
Salsa	57.78		
Sertaneja	42.22		
Tango	87.78		
Overall	**61.00** $\pm$ 1.53	**Overall**	**74.47**

Table 21
Confusion matrix (%) obtained in LMD using acoustic features.

Genre	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(0) Axé	**57.78**	2.22	2.22	4.44	6.67	1.11	4.44	7.78	12.22	1.11
(1) Bachata	0.00	**85.56**	2.22	2.22	1.11	2.22	1.11	5.56	0.00	0.00
(2) Bolero	0.00	3.33	**63.33**	4.44	5.56	0.00	8.89	3.33	7.78	3.33
(3) Forró	0.00	5.56	11.11	**38.89**	17.78	2.22	4.44	10.00	8.89	1.11
(4) Gaúcha	8.89	3.33	6.67	10.00	**51.11**	4.44	1.11	5.56	7.78	1.11
(5) Merengue	1.11	3.33	0.00	2.22	4.44	**78.89**	1.11	8.89	0.00	0.00
(6) Pagode	5.56	1.11	12.22	10.00	10.11	1.11	**46.67**	3.33	10.00	0.00
(7) Salsa	11.11	2.22	1.11	7.78	7.78	7.78	3.33	**57.78**	1.11	0.00
(8) Sertaneja	17.78	0.00	8.89	6.67	4.44	0.00	18.89	1.11	**42.22**	0.00
(9) Tango	0.00	0.00	5.56	1.11	2.22	0.00	1.11	0.00	2.22	**87.78**

Table 22
Confusion matrix (%) obtained in ISMIR 2004 using acoustic features.

Genre	(0)	(1)	(2)	(3)	(4)	(5)
(0) Classical	**95.42**	0.33	0.00	0.00	0.00	4.25
(1) Electronic	4.42	**76.11**	0.00	0.88	4.42	14.16
(2) Jazz/blues	7.69	3.85	**50.00**	0.00	7.69	30.77
(3) Metal/punk	0.00	6.67	0.00	**51.11**	42.22	0.00
(4) Rock/pop	5.88	8.82	0.98	11.76	**59.80**	12.75
(5) World	33.06	14.88	0.00	0.83	4.96	**46.28**

Table 23
Best results obtained in this work for each experimented zoning scheme.

Zoning/features	Recognition rate in LMD	Recognition rate in ISMIR 2004
LBP with global feature extraction	79.00 $\pm$ 1.00	**80.65**
LBP with linear zoning	77.78 $\pm$ 0.38	78.40
LBP with Bark scale zoning	78.00 $\pm$ 1.33	71.53
LBP with Mel scale zoning	**82.33** $\pm$ 1.45	76.74
Acoustic features	61.00 $\pm$ 1.53	74.47

on feature extracted from spectrograms, but using a different texture descriptor. The descriptor used was the well known GLCM, a statistical approach to describe texture content. Furthermore, only the linear zoning mechanism was used and only one classifier for all zones was created, with the final decision done through majority voting. In 2009, Cao and Li won MIREX Audio Genre Classification (Latin set) using basic acoustic features (e.g. MFCC) and the modeling framework of GSV-SVM [33].

Table 24
Best results obtained with LMD in this work and in the state of the art.

Genre	Mel scale zoning	Acoustic features	GLCM feature extraction [18]	Instance selection [23]	GLCM + inst. selection [18]	MIREX 2009 winner [33]	MIREX 2010 winner [34]
Axé	**83.33**	57.78	73.33	61.11	76.67	53.04	69.32
Bachata	93.33	85.56	82.22	91.11	87.78	**97.12**	95.84
Bolero	**91.11**	63.33	64.44	72.22	83.33	82.22	**91.11**
Forró	82.22	38.89	65.56	17.76	52.22	82.75	**83.38**
Gaúcha	67.78	51.11	35.56	44.00	48.78	**76.92**	72.11
Merengue	**95.56**	78.89	80.00	78.78	87.78	94.29	94.60
Pagode	71.11	46.67	46.67	61.11	61.11	52.94	**79.41**
Salsa	84.44	57.78	42.22	40.00	50.00	90.35	**93.56**
Sertaneja	**67.78**	42.22	17.78	41.11	34.44	19.94	36.13
Tango	86.67	87.78	93.33	88.89	**90.00**	84.31	83.08
Overall	**82.33**	61.00	60.11	59.67	67.20	74.66	79.86

Table 25
Best result obtained in this work and other existing approaches for the ISMIR 2004.

Work	Recognition rate
LBP with global feature extraction	80.65
Acoustic features	74.47
Marques et al. [35]	79.80
Lidy et al. [36]	81.40
Wu et al. [20]	86.10
Seyerlehner et al. [34]	**88.27**

In 2010, Seyerlehner et al. won the MIREX Audio Genre Classification (Latin) using a set of block-level features, more details can be found in [34].

Note that the overall recognition rate obtained in this work is higher than that obtained in other works. In addition, the standard deviation found between the classes is smaller with the classifier presented in this work.

The best result obtained on the ISMIR 2004 database (global feature extraction) and other results with this dataset presented in the literature are shown in Table 25. In order to make this table, it was selected some works in which the standard splitting of training and test sets was used. There are several works described in the literature aiming music genre classification using the ISMIR 2004 database. As not all of them show the recognition rates per genre, we could not show the detailed rates here.

The obtained results on the ISMIR 2004 database shows that the features presented here can perform well in different databases. Despite of performing better on the LMD, the classifier presented an acceptable performance on the ISMIR 2004 database, indicating its versatility.

6. Conclusion

In this paper we have presented an alternative approach for music genre classification which is based on texture images. Such visual representations are created by converting the audio signal representation into spectrograms images which can be divided into zones then that features can be extracted locally. We have demonstrated that, with LBP, there is a slight difference in terms of recognition rate when different zoning mechanisms are used and when a global feature extraction is performed, contrary to expectations generated by a previous work, that used GLCM.

Two different databases were used in the experiments: LMD, and ISMIR 2004. The best overall result in LMD was obtained with a zoning mechanism done according to the Mel scale, a scale that relates real frequency to perceived frequency. This result is not surprising since a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a non-linear Mel scale of frequency is commonly used as features in speech recognition systems, speaker recognition and in music information retrieval applications such as genre classification and audio similarity measures. Furthermore, one can note that with Mel scale zoning, the recognition rate obtained is 3.33 percentage point greater than those obtained without local feature extraction. One can say that without the use of zoning mechanism, the overall system complexity decreases, since a smaller number of classifiers is created. On the other hand, upper limit rates obtained with classifiers produced with zoning schemes are very promising which opens up interesting perspectives for future working on selection of classifiers.

The artist filter restriction was considered in all the experiments conducted on LMD. For ISMIR 2004 it was not possible, since the training and test sets are predefined in the original database splitting. There is no report in the literature about results with LMD, taking into account the artist filter restriction, better than those obtained in this work.

Regarding to ISMIR 2004, the obtained results are close to many others described in the literature. In general, it was possible to conclude that the features experimented here presents an interesting versatility in which concerns to different databases, specially when faced with some common acoustic features.

Although we did not carry out a rigorous analysis to find out if the LBP features used in this work capture different information from the features currently employed in the audio signal based approaches, there is some evidence that this can be true, since the distribution of the errors when comparing the confusion matrices are different. However, this particular aspect of the possible

complementarity of the image-based and audio signal-based features will be the subject of future work.

Our future work will also focus on the evaluation of other texture features, weighted zoning strategies, analysis of recognition rates versus multi classifier systems complexity, and experiments combining the different schemes presented here. There was not only one classifier providing best results for all the genres, it suggests that there is complementarity between them. With this in mind, we intend to explore the dynamic selection of classifiers.

Acknowledgments

This research has been partly supported by The National Council for Scientific and Technological Development (CNPq) grant 301653/2011-9, CAPES grant BEX 5779/11-1 and 223/09-FCT595-2009, Araucária Foundation grant 16767-424/2009, European Commission, FP7 (Seventh Framework Programme), ICT-2011.1.5 Networked Media and Search Systems, grant agreement No 287711; and the European Regional Development Fund through the Programme COMPETE and by National Funds through the Portuguese Foundation for Science and Technology, within projects ref. PTDC/EAT-MMU/112255/2009 and PTDC/EIA-CCO/111050/2009.

References

[1] J. Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, A. Toncheva, The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011, Technical Report, International Data Corporation (IDC), 2008.

[2] N. Orio, Automatic identification of audio recordings based on statistical modeling, Signal Processing 90 (4) (2010) 1064–1076.

[3] C.N. Silla, A.L. Koerich NS, C.A.A. Kaestner, Feature selection approach for automatic music genre classification, International Journal of Semantic Computing 3 (2) (2009) 183–208.

[4] J.J. Aucouturier, F. Pachet, Representing musical genre: a state of the art, Journal of New Music Research 32 (1) (2003) 83–93.

[5] C. McKay, 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.

[6] R.O. Gjerdingen, D. Perrott, Scanning the dial: the rapid recognition of music genres, Journal of New Music Research 37 (2) (2008) 93–100.

[7] F. Pachet, D. Cazaly, A taxonomy of musical genres, in: Proceedings of Content-Based Multimedia Information Access (RIAO), 2000, pp. 1238–1245.

[8] E. Pampalk, Computational Models of Music Similarity and Their Application in Music Information Retrieval, PhD Thesis, Viena University of Technology, 2006.

[9] T. Lidy, C.N. Silla, O. Cornelis, F. Gouyon, A. Rauber, C.A.A. Kaestner, andA.L. Koerich, On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-western and ethnic music collections, Signal Processing 90 (2010) 1032–1048.

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

[11] T. Li, M. Ogihara, Q. Li, A comparative study on content-based music genre classification, in: 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, pp. 282–289.

[12] F. Gouyon, S. Dixon, E. Pampalk, G. Widmer, Evaluating rhythmic descriptors for musical genre classification, in: 25th International AES Conference, 2004, pp. 196–204.

[13] A. Rauber, E. Pampalk, D. Merkl, Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by musical styles, in: 3rd International Conference on Music Information Retrieval, 2002, pp. 71–80.

[14] T. Lidy, A. Rauber, Evaluation of feature extractors and psycho-acoustic transformations for music genre classification, in: 6th International Conference on Music Information Retrieval, 2005, pp 34–41.

[15] S. Lippens, J.P. Martens, M. Leman, B. Baets, H. Meyer, G. Tzanetakis, A comparison of human and automatic musical genre classification, in: IEEE International Conference on Audio, Speech and Signal Processing, 2004, pp. IV-233–IV-236.

[16] M.R. French, R.G. Handy, Spectrograms: turning signals into pictures, Journal of Engineering Technology (2007) 32–35.

[17] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 971–987.

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

[19] H. Deshpande, R. Singh, U. Nam, Classification of music signals in the visual domain, in: Proceedings of the COST-G6 Conference on Digital Audio Effects, 2001, pp. 1–4.

[20] M. Wu, Z. Chen, J.R. Jang, J. Ren, Combining visual and acoustic features for music genre classification, in: 10th International Conference on Machine Learning and Applications, 2011, pp. 124–129.

[21] C.N. Silla, A.L. Koerich, C.A.A. Kaestner, The latin music database, in: 9th International Conference on Music Information Retrieval, 2008, pp. 451–456.

[22] E. Gomez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, P. Cano, N. Wack, Ismir 2004 Audio Description Contest, Technical Report, Music Technology Group - Universitat Pompeu Fabra, 2006.

[23] M. Lopes, F. Gouyon, A. Koerich, L.S. Oliveira, Selection of training instances for music genre classification, in: 20th International Conference on Pattern Recognition, 2010, pp. 4569–4572.

[24] MIREX, Music Information Retrieval Evaluation Exchange, 2010 〈http://www.music-ir.org〉.

[25] A. Flexer, A closer look on artists filters for musical genre classification, in: International Conference on Music Information Retrieval, 2007, pp. 341–344.

[26] C. Costa, J. Valle Jr., A. Koerich, Automatic classification of audio data, in: IEEE International Conference on Systems, Man, and Cybernetics, 2004, pp. 562–567.

[27] T. Mäenpää, M. Pietikäinen, Texture analysis with local binary patterns, in: Handbook of Pattern Recognition and Computer Vision, vol. 3, 2005, pp. 197–216.

[28] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, Inc., 1995.

[29] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 226–239.

[30] E. Zwicker, Subdivision of the audible frequency range into critical bands (frequenzgruppen), Acoustical Society of America Journal 33 (1961) 248.

[31] S. Umesh, L. Cohen, D. Nelson, Fitting the Mel scale, in: IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP'99. Proceedings, vol. 1. IEEE, 1999, pp. 217–220.

[32] A.H.R. Ko, R. Sabourin, A.S. Britto, et al., From dynamic classifier selection to dynamic ensemble selection, Pattern Recognition 41 (5) (2008) 1718–1731.

[33] C. Cao, M. Li, Thinkit's Submission for MIREX 2009 Audio Music Classification and Similarity Tasks, 2009.

[34] K. Seyerlehner, M. Schedl, T. Pohle, P. Knees, Using Block-level Features for Genre Classification, Tag Classification and Music Similarity Estimation, Submission to Audio Music Similarity and Retrieval Task of MIREX 2010, 2010.

[35] G. Marques, M. Lopes, M. Sordo, T. Langlois, F. Gouyon, Additional evidence that common low-level features of individual audio frames are not representative of music genres, in: Sound and Music Computing Conference, Barcelona, 2010.

[36] T. Lidy, A. Rauber, A. Pertusa, J.M. Inesta, Improving genre classification by combination of audio and symbolic descriptors using a transcription system, in: Proceedings of ISMIR, Vienna, Austria, 2007.