COMPARING TEXTURAL FEATURES FOR MUSIC GENRE CLASSIFICATION

Yandre M. G. Costa*, Luiz S. Oliveira†, Alessandro L. Koerich‡, and Fabien Gouyon§
*State University of Maringá
Maringá, Brazil
Email: yandre@din.uem.br
†Federal University of Paraná
Curitiba, Brazil
‡Pontifical Catholic University of Paraná
Curitiba, Brazil
§INESC Porto
Porto, Portugal

***Abstract*—In this paper we compare two different textural feature sets for automatic music genre classification. The idea is to convert the audio signal into spectrograms and then extract features from this visual representation. Two textural descriptors are explored in this work: the Gray Level Co-Occurrence Matrix (GLCM) and Local Binary Patterns (LBP). Besides, two different strategies of extracting features are considered: a global approach where the features are extracted from the entire spectrogram image and then classified by a single classifier; a local approach where the spectrogram image is split into several zones which are classified independently and final decision is then obtained by combining all the partial results.**

The database used in our experiments was the Latin Music Database, which contains music pieces categorized into 10 musical genres, and has been used for MIREX (Music Information Retrieval Evaluation eXchange) competitions. After a comprehensive series of experiments we show that the SVM classifier trained with LBP is able to achieve a recognition rate of 80%. This rate not only outperforms the GLCM by a fair margin but also is slightly better than the results reported in the literature.

I. Introduction

From 2002, when Tzanetakis and Cook [1] introduced music genre classification as a pattern recognition task, many other works has been developed for this purpose [2], [3], [4], [5], [6], [7]. According to Lidy et al.[8], most of the works rely on the content-based approach, which extracts representative features from the digital audio signal. Among the most common features used we can mention for example, timbral texture, beat-related, pitch-related, and rhythm histograms.

In spite of all efforts done so far, the automatic music genre classification still remains an open problem. McKay and Fujinaga [9] pointed out some problematic aspects of genre and refer to some experiments where human beings were not able to correctly classify more than 76% of the musics. 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, Costa et al [10] proposed an alternative approach for automatic genre classification. It converted the audio signal into spectrograms [11] (short-time Fourier representation) and then extracted textural features from the visual representation. The experiments reported in [10], using the Latin Music Database, took into account the Gray Level Co-Ocurrence Matrix (GLCM) textural descriptors and achieved similar results to those methods based on traditional features. However, the authors have shown that the classifiers based on textures carry some complementary information when compared to the traditional ones. When both strategies were combined, a significant improvement of about 10% was achieved.

The GLCM and its descriptors were proposed by Haralick [12] almost 40 years ago. Since then other textural descriptors have been developed and successfully applied into different areas, but one of them, the Local Binary Pattern (LBP) has gained a lot of attention because of its performance and simplicity of implementation. The concept of LBP was first proposed by Ojala et al. in [13] as a simple and robust approach in terms of grayscale variations. It was proved to discriminate a large range of rotated textures efficiently. Later, they extend their work [14] to be a gray-scale and rotation invariant texture operator.

With this in mind, in this work we pursue the investigation initiated in [10] by comparing both GLCM and LBP as textural descriptors to perform music genre classification. By analyzing the spectrogram images one can notice that different patterns of texture may occur in the same image. To deal with this, two strategies for feature extraction were considered. The first one is a local approach where the spectrogram image is divided into several zones that are independently classified and the final result is achieved by combining all the partial decisions. The second strategy, on the other hand, is a holistic one. In this case the features are extracted from the entire spectrogram image.

Our experiments were carried out on the Latin Music Database [15], a very challenging dataset of 900 music pieces

divided among 10 music genres. The results reported in this work show that the SVM classifier trained with LBP is able to achieve a recognition rate of 80%. This rate not only outperforms the GLCM by a fair margin but also is slightly better than the results reported in the literature. Taking into account the best results obtained in MIREX 2009 and MIREX 2010 [16] competitions, the improvement was about six and one percentage points, respectively.

This paper is organized as follows: Section II describes some basic aspects about the LMD. Section III describes details about feature extraction performed in this work. Section IV introduces the methodology used for classification while Section V reports the experimental results. Finally, Section VI concludes this work.

II. LATIN MUSIC DATABASE

Presented by Silla et al. [15], the LMD is a digital music database created for support research in music information retrieval. This database is composed of 3,227 full-length music samples in MP3 format originated from music pieces of 501 artists. 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 dataset 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. The professionals classified around 300 music pieces per month, and the development of the complete database took around one year.

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 [17], 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. [7] and by Costa et al. [10]. 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 Music Information Retrieval Evaluation eXchange (MIREX) [16].

III. FEATURE EXTRACTION

Before proceed the generation of the visual representation, we performed a time decomposition based on the idea presented by Costa et al. [18] 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 $S_{pq} =< s_p, \ldots, s_q >$. In the generic case, one may extract K (overlapping or non-overlapping) sub-signals and obtain a sequence of spectrograms $\overline{\Upsilon}_1, \overline{\Upsilon}_2, \ldots, \overline{\Upsilon}_K$. We have used the strategy proposed by Silla et al. [15] which considers three 10-second segments from the beginning ($\overline{\Upsilon}_{beg}$), middle ($\overline{\Upsilon}_{mid}$), and end ($\overline{\Upsilon}_{end}$) parts of the original music. In order to avoid segments that do not provide good discrimination among genres, we decided to ignore the first ten seconds and the last ten seconds of the music pieces. The rationale behind this strategy is that some common effects present in these parts of the music signal, like fade in and fade out, as well as kinds of noise, like those produced by the audience, could turn these signal samples less discriminant than the others.

After the signal decomposition, the next step consists in converting the audio signal into a spectrogram. The spectrograms were created using a bit rate = 352kbps, audio sample size = 16 bits, one channel, and audio sample rate = 22.05 kHz. Figure 1 depicts the signal segmentation and spectrogram generation.

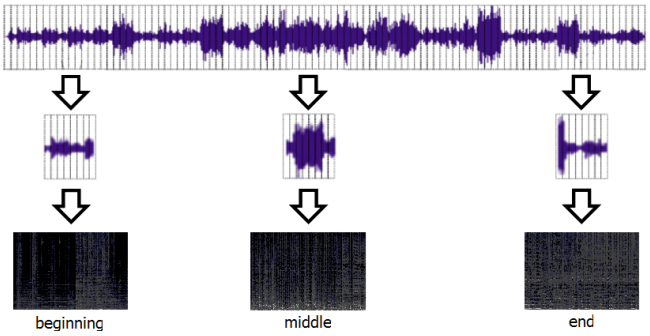

Figure 1. Creating spectrograms using time decomposition.

Once the spectrograms were generated we proceeded the texture feature extraction from these images. As stated before, the approach proposed in this work considers that the main visual content present in the spectrogram images is the texture. With this in mind, we used the GLCM and LBP texture operators, described respectivelly in Sections III-A and III-B, to get features.

A. Gray Level Co-occurencce Matrix

Among the statistical techniques of texture recognition, the GLCM has been one of the most used and successful ones. This technique consists of statistical experiments conducted on how a certain level of gray occurs on other levels of gray. It intuitively provides measures of properties such as smoothness, coarseness, and regularity. By definition, a GLCM is the joint probability occurrence of gray level i and j within

a defined spatial relation in an image. That spatial relation is defined in terms of a distance d and an angle θ. Given a GLCM, some statistical information can be extracted from it.

Haralick [19], the precursor of this technique, suggested a set of 14 characteristics, but most of the works in the literature consider a subset of these descriptors. In our case, we have used the following seven descriptors, which have produced interesting results for other texture problems: Entropy, Correlation, Homogeneity, 3rd Order Momentum, Maximum Likelihood, Contrast, and Energy. Those readers interested in the mathematical formulation can refer to [19].

In our experiments we have tried different values for d as well as different angles. The best setup we have found is $d = 1$ and $\theta = [0, 45, 90, 135]$. Considering the seven descriptors aforementioned, in the end we have a feature vector of 28 components for each image zone.

B. Local Binary Pattern

Presented by Ojala et al. [14], LBP is a model that describes the texture taking into account for each pixel *C*, a set of neighbors *P*, equally spaced at a distance of *R*, as shown in Figure 2.

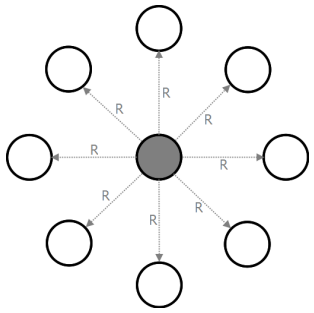

Figure 2. The LBP operator. A pixel *C*, dark circle in the middle, and its neighbors P_n, lighter circles.

An histogram *h* is defined by the texture intensity differences of *C* and its neighbors *P*. 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 *P*, it is defined that: if the sign is positive the result is 1, otherwise 0. Thus, it is possible to obtain this invariance of the intensity value of pixels in gray-scale format. With this, the LBP value can be obtained by multiplying the binary elements for a binomial coefficient. So, it is generated a value $0 \leq C' \leq 2^P$ (corresponding to the vector).

Observing the non-uniformity of the vector obtained, Ojala et al. [14] 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. Figure 3 illustrates this idea.

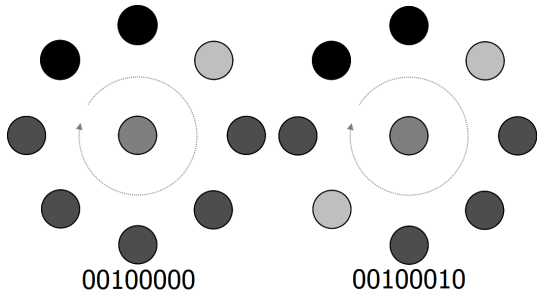

Figure 3. LBP uniform pattern. (a) the two transitions showed identifies the pattern as uniform. (b) with four transitions, it is not considered a uniform pattern.

So, instead of using the whole histogram, which size is 2^P, it is possible to use only the uniform values, constituting a smaller feature vector, with only 59 features. 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: $LBP^{label}_{P,R}$.

During the experiments, we observed that the feature extraction with $LBP^{u2}_{8,2}$ is fast and accurate enough for the proposed application. Then, we choose to use *P*=8 and *R*=2 on the tests described in this paper.

C. Global and Local Feature Extraction

The global approach is the simplest way to perform feature extraction of a given spectrogram image. This is a holistic by nature where the features are extracted from the entire image and the final decision is produced by a single classifier.

However, by analyzing the spectrogram images one can notice that different patterns of texture may occur in the same image. This can be observed in the spectrogram depicted in Figure 4. To deal with that, in our previous work [10] we proposed a zoning mechanism to obtain local information rather than a global one. The idea was to take advantage of these different texture patterns by processing them in an independent way. Differently from [10], where only one classifier was created with feature vectors from all zones, here we train one classifier for each zone and the final decision is obtained using traditional combination rules as described in Section IV.

In order to proceed the local feature extraction, we have evaluated six different number of linear zones (1, 3, 5, 10, 15, and 20), which were applied to the spectrogram image before extracting textural features.

Thus, considering that three spectrogram images were generated from each music piece, since we extracted three segments, the number of total zones, and consequently the number of classifiers is $3n$. The rationale behind the zoning and combining scheme is that music signals may include similar instruments and similar rhythmic patterns which leads to similar areas in the spectrogram images. By zoning the images we can extract local information and try to highlight

the specificities of each music genre. In some cases we can notice that at low frequencies the textures are quite similar but they get different as the frequency increases. The opposite can happen as well and for this reason the zoning mechanism becomes an interesting alternative.

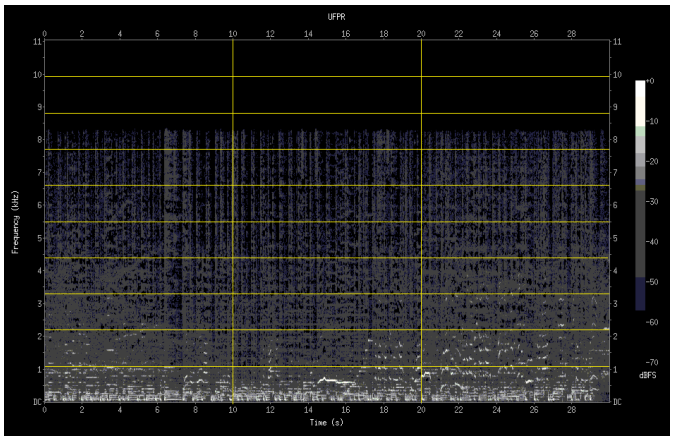

Figure 4. Zoning with n = 10.

In the next Section we show some details about the methodology used for classification.

IV. METHODOLOGY USED FOR CLASSIFICATION

The classifier used in this work was the Support Vector Machine (SVM) introduced by Vapnik in [20]. Normalization was performed by linearly scaling each attribute to the range [-1,+1]. The Gaussian kernel was used, with parameters C and γ tuned using a grid search.

The classification process is done as follows: as aforementioned, the three 10-second segments of the music are converted to the spectrograms ($\overline{\Upsilon}_{beg}$, $\overline{\Upsilon}_{mid}$, and $\overline{\Upsilon}_{end}$). Each of them is divided into n zones, according to the values of n described in subsection III-C. Then, a 28-dimensional GLCM feature vector and a 59-dimensional LBP feature vector are extracted from each zone. Next, each one of these feature vectors is sent to a specific classifier, which assigns a prediction to each one of the ten possible classes. Training and classification were carried out using the 3-fold cross-validation: 1 fold used for training a 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 created $3n$ classifiers with 600 and 300 feature vectors for training and testing, respectively.

With this amount of classifiers, we used estimation of probabilities to proceed the combination of outputs in order to get a final decision. In this situation, is very useful to have a classifier producing a posterior probability $P(class|input)$. Here, we are interested in estimation of probabilities because we want to try different fusion strategies like Max, Min, Product, and Sum. The following equations, presented by Kittler et al. [21], describe how the outputs are combined with these four decision rules in order to get a final decision:

$$\text{Max Rule}(x) = \max_{k=1}^{c} \max_{i=1}^{m} P(\omega_k|y_i(x)) \quad (1)$$

$$\text{Min Rule}(x) = \max_{k=1}^{c} \min_{i=1}^{m} P(\omega_k|y_i(x)) \quad (2)$$

$$\text{Product Rule}(x) = \max_{k=1}^{c} \prod_{i=1}^{m} P(\omega_k|y_i(x)) \quad (3)$$

$$\text{Sum Rule}(x) = \max_{k=1}^{c} \sum_{i=1}^{m} P(\omega_k|y_i(x)) \quad (4)$$

where x represents the pattern to be classified, m is the number of classifiers (in this case 3 times n, the number of zones), y_i represents the output label of the $i-th$ classifier in a problem in which the possible class labels are $\Omega = \omega_1, \omega_2, ..., \omega_c$, and $P(\omega_k|y_i(x))$ is the estimation of probability of pattern x belong to class ω_k according to $i-th$ classifier.

V. EXPERIMENTAL RESULTS AND DISCUSSION

The results presented here refer to the average recognition rate considering the three folds aforementioned. Subsection V-A presents the results obtained with GLCM features, while subsection V-B presents the results obtained with LBP features. Finally, subsection V-C presents a brief discussion about all the obtained results.

A. Results with GLCM features

Table I reports the results obtained when GLCM features were used with four different combination rules and with six different zoning configurations for each spectrogram generated from a music piece. The results achieved using the holistic approach (no zoning) compare to the results reported by Lopes et al [7]. On the other hand, by increasing the number of zones up to a certain point we observe an important improvement. In this experiment, using a local approach with five zones brought us an improvement of more than five percentage points.

Table I
RECOGNITION RATES (%) OBTAINED WHEN DIFFERENT NUMBER OF ZONES AND DIFFERENT COMBINATION RULES ARE USED WITH GLCM FEATURES.

Number of zones	Max. rule	Min. rule	Product rule	Sum rule
No zoning	59.56	60.78	**64.67**	63.44
3	57.56	61.56	69.89	69.22
5	57.11	60.44	**70.78**	69.33
10	55.56	58.00	69.78	68.22
15	53.22	58.78	69.11	68.22
20	47.22	54.78	41.56	67.22

Table II shows the confusion matrix produced by the by the combination of classifiers trained with GLCM features using five zones. The results are very similar to those reported in [7] where the highest confusions are related to classes Gaúcha (4) and Sertaneja (8).

Table II
CONFUSION MATRIX (%) GLCM

	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(0)	**70.0**	0.0	4.4	0.0	6.7	1.1	11.1	1.1	5.6	0.0
(1)	1.1	**85.6**	4.4	0.0	4.4	1.1	0.0	3.3	0.0	0.0
(2)	0.0	1.1	**82.2**	3.3	2.2	0.0	3.3	1.1	3.3	3.3
(3)	2.2	1.1	1.1	**65.6**	8.9	1.1	1.1	3.3	5.6	0.0
(4)	18.9	1.1	1.1	8.9	**50.0**	2.2	1.1	0.0	5.6	1.1
(5)	0.0	1.1	0.0	2.2	2.2	**85.6**	2.2	6.7	0.0	0.0
(6)	12.2	0.0	12.2	3.3	6.7	0.0	**55.6**	5.6	4.4	0.0
(7)	2.2	0.0	2.2	7.8	3.3	8.9	6.7	**66.67**	2.2	0.0
(8)	20.0	0.0	12.2	7.8	0.0	0.0	4.4	0.00	**55.6**	0.0
(9)	0.0	0.0	6.7	0.0	1.1	0.0	1.1	0.00	0.00	**91.1**

(0) Axé,(1) Bachata, (2)Bolero, (3) Forró, (4) Gaúcha, (5) Merengue, (6) Pagode, (7) Salsa, (8) Sertaneja (9) Tango

Table IV
CONFUSION MATRIX (%) LBP

	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(0)	**74.4**	0.0	2.2	0.0	0.0	0.0	11.1	3.3	8.9	0.0
(1)	1.1	**92.2**	2.2	1.1	0.0	1.1	0.0	1.1	1.1	0.0
(2)	0.0	1.1	**91.1**	0.0	1.1	0.0	1.1	0.0	4.4	1.1
(3)	0.0	1.1	4.4	**77.8**	5.5	3.3	3.3	2.2	2.2	0.0
(4)	11.1	1.1	5.5	6.7	**67.8**	0.0	0.0	0.0	7.8	0.0
(5)	1.1	3.3	0.0	1.1	0.0	**92.2**	0.0	2.2	0.0	0.0
(6)	6.7	0.0	13.3	1.1	1.1	0.0	**64.4**	7.8	5.6	0.0
(7)	2.2	0.0	2.2	0.0	3.3	1.1	3.3	**86.7**	1.1	0.0
(8)	11.1	1.1	7.8	2.2	10.0	0.0	2.2	0.0	**65.6**	0.0
(9)	1.1	0.0	5.6	1.1	1.1	0.0	0.0	0.0	0.0	**91.1**

(0) Axé,(1) Bachata, (2)Bolero, (3) Forró, (4) Gaúcha, (5) Merengue, (6) Pagode, (7) Salsa, (8) Sertaneja (9) Tango

B. Results with LBP features

Table III shows the results obtained with LBP features. Differently from the previous experiments where the local approach produces a remarkable improvement relative to the global one, here, of both approaches achieve similar results. Of course, in this context the global approach is more appealing since it uses only one classifier. In the top of that, the best result achieved by the classifier trained with the LBP feature set is about 10 percentage points better than the best results achieved with the GLCM features.

Table III
RECOGNITION RATES (%) OBTAINED WHEN DIFFERENT NUMBER OF ZONES AND DIFFERENT COMBINATION RULES ARE USED WITH LBP FEATURES.

Number of zones	Max. rule	Min. rule	Product rule	Sum rule
No Zoning	76.56	75.67	78.78	**79.22**
3	73.44	74.56	78.67	79.00
5	72.89	74.78	**80.33**	80.11
10	72.33	72.11	78.44	78.67
15	70.89	73.89	79.33	77.78
20	69.00	72.67	63.89	76.78

In Table IV we can visualize the confusion matrix produced by the combination of classifiers trained with LBP features using a local approach with five zones. It shows that the classifier trained with LBP is able to reduce several confusions perceived in Table II, except for class Tango (9) where the GLCM performs well.

C. Discussion

When comparing the performance of GLCM features with the LBP features in this application, one can notice that the classifiers trained with LBP achieved recognition rates significantly better than that achieved with the GLCM features. In the best case, the recognition rate with the LBP feature was about 10 percentage points greater than the recognition rate achieved with the GLCM features.

Regarding the local and global approaches, it is easy to observe that the local strategy pays off when dealing with the GLCM features. This feature set is not able to deal with the great variability of the spectrogram image, therefore training different classifiers for specific parts of the image is better than using just one classifier for the whole spectrogram image. However, the experiments have shown, however, that after five zones there is no further improvement. In the case of the LBP features, both local and global approaches perform almost evenly. The local approach using five zones yields slightly better results. As explained earlier, the global approach could be more interesting in this case since it requires a single classifier.

The relevance of the results achieved by the LBP feature set is clear when compared with the literature. Table V reports recent results on the LMD using artist filter. Such results can be directly compared since all of them use exactly the same experimental protocol. In addition, we have access to the performance for each class. Lopes et al. [7] 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. Costa et al. [10], presented a classification scheme similar to some experiments presented in this work, based on GLCM features extracted from spectrograms, but using only a single classifier for the feature vectors extracted from all the zones. The final decision was taken simply through majority voting.

Table V
COMPARISON OF DIFFERENT STRATEGIES ON LMD WITH ARTIST FILTER.

Genre	LBP 5 zones	GLCM 5 zones	GLCM [10]	Instance Selection [7]	GLCM +Inst. Selection [10]
Axé	74.44	70.00	73.33	61.11	**76.67**
Bachata	**92.22**	85.56	82.22	91.11	87.78
Bolero	**91.11**	82.22	64.44	72.22	83.33
Forró	**77.78**	65.56	65.56	17.76	52.22
Gaúcha	**67.78**	50.00	35.56	44.00	48.78
Merengue	**92.22**	85.56	80.00	78.78	87.78
Pagode	**64.44**	55.56	46.67	61.11	61.11
Salsa	**86.67**	66.67	42.22	40.00	50.00
Sertaneja	**65.56**	55.56	17.78	41.11	34.44
Tango	91.11	91.11	**93.33**	88.89	90.00
Average	**80.33**	70.78	60.11	59.67	67.20

From Table V it is easy to see that the LBP-based system surpasses the others by a large margin. For some classes, such as Gaúcha and Sertaneja, which are quite difficult to discriminate, this strategy is able to get recognition rates of around 65%, while others stay below 50%. Besides, it provides the best performance for 8 out of 10 classes. In spite of that, there is room for some improvement. Other strategies have good results for some classes which could be combined somehow to improve the classification rates. This will be subject of future works.

As mentioned before, the Latin Music Database has been used in the competitions organized by the MIREX (Music Information Retrieval Evaluation eXchange). From Table VI we can notice the outstanding improvement over the last few years. As we can see, our result using the LBP-based system compares favorably to best results reported in the literature.

Table VI
RECOGNITION RATES (%) OBTAINED IN THIS WORK, BY HUMANS AND IN THE LAST AUTOMATIC MUSIC GENRE RECOGNITION CONTESTS.

Work reference	Recognition rate (%)
MIREX 2008 - LMD [22]	65.1
MIREX 2009 - LMD [23]	74.6
MIREX 2010 - LMD [16]	79.8
This work (GLCM)	70.7
This work (LBP)	80.3

VI. CONCLUSION

In this paper we have compared two different textural descriptors to perform music genre classification. The idea was to convert the music signal into a spectrogram image and them extract textural features from it. Two textural features, GLCM and LBP, were evaluated in this paper as well as two different feature extraction approaches to deal with the intra-class variability of the spectrogram image. The local approach divided the image into several different zones which are independently classified by different classifiers. The second one is holistic by nature since it process the entire spectrogram image as a whole.

The experimental results have shown that the local approach performs very well when the classifier is trained with the GLCM feature set. We have also seen that after a certain number of zones (five in our experiments) there is no further improvements in terms of correct classification. For the classifier trained with LBP, both local and global approach achieve similar results. In such a case, the holistic approach is more attractive because it requires only one classifier.

Our experiments also have shown that the LPB-based system achieves a overall recognition rate of 80%, which compares to the best results reported in MIREX 2010 for the Latin Music Database. In the light of this promising results, in future work we plan to combine them with other traditional strategies to enhance the final recognition rates.

ACKNOWLEDGMENT

The authors would like to thank CNPq (Grants #301653/2011-9, 402357/2009-4), CAPES and Araucaria Foundation (Grant #16767).

REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[2] C. N. Silla and A. L. K. N. C. A. A. Kaestner, "Feature selection approach for automatic music genre classification," *International Journal of Semantic Computing*, vol. 3, no. 2, pp. 183–208, 2009.

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

[4] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *25th International AES Conference*, 2004.

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

[6] T. Lidy and 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.

[7] M. Lopes, F. Gouyon, A. Koerich, and L. S. Oliveira, "Selection of training instances for music genre classification," in *20th Int. Conf. on Pattern Recognition*, 2010.

[8] T. Lidy, C. N. Silla, O. Cornelis, F. Gouyon, A. Rauber, C. A. A. Kaestner, and A. 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*, vol. 90, pp. 1032–1048, 2010.

[9] C. McKay and I. Fujinaga, "Musical genre classification: Is it worth pursuing and how can it be improved?" in *7th Int. Conf. on Music Information Retrieval*, 2006.

[10] 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.

[11] M. R. French and R. G. Handy, "Spectrograms: turning signals into pictures," *Journal of engineering technology*, pp. 32–25, 2007.

[12] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man and cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

[13] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on featured distribution," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[14] T. Ojala, M. Pietikainen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

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

[16] Mirex, "Music information retrieval evaluation exchange," 2010, http://www.music-ir.org/mirex/wiki/2010:Main_Page.

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

[18] C. Costa, J. Valle-Jr, and A. Koerich, "Automatic classification of audio data," in *IEEE Int. Conf. on Systems, Man, and Cybernetics*, 2004, pp. 562–567.

[19] R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of IEEE*, vol. 67, no. 5, 1979.

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

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

[22] Mirex, "Music information retrieval evaluation exchange," 2008, http://www.music-ir.org/mirex/wiki/2008:Main_Page.

[23] ——, "Music information retrieval evaluation exchange," 2009, http://www.music-ir.org/mirex/wiki/2009:Main_Page.