# Music Genre Recognition Using Spectrograms

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 present an alternative approach for music genre classification which converts the audio signal into spectrograms and then extracts features from this visual representation. The idea is that treating the time-frequency representation as a texture image we can extract features to build reliable music genre classification systems. The proposed approach also takes into account a zoning mechanism to perform local feature extraction, which has been proved to be quite efficient. On a very challenging dataset of 900 music pieces divided among 10 music genres, we have demonstrated that the classifier trained with texture compares similarly to the literature. Besides, when it was combined with other classifiers trained with short-term, low-level characteristics of the music audio signal we got an improvement of about 7 percentage points in the recognition rate.**

## I. Introduction

Music genres are categorical labels created by humans to determine the style of music. Because of the human perception subjectiveness, assigning a genre to a music piece is not a trivial task. In spite of that, music 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 [1].

The literature shows us that in the last decade several researchers have devoted a considerable amount of efforts towards automatic music genre classification.One of the earlier works was introduced by Tzanetakis and Cook [2] 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], [4]. Other characteristics such as Inter-Onset Interval Histogram Coefficients, Rhythm Patterns and its derivatives Statistical Spectrum Descriptors, and Rhythm Histograms also have been proposed in the literature recently [4].

In spite of all efforts done during the last years, the 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 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 realize 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 [6] (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 different features, which we expect to be complementary to the traditional ones, to build a more robust music genre classification system. The characteristics used in this work are some of the Gray Level Co-occurrence Matrix (GLCM) descriptors introduced by Haralick [7].

By analyzing the spectrogram images, we have noticed that the textures are not uniform, so it is important to consider a local feature extraction rather than a global one. With this in mind, we propose an efficient zoning technique to obtain local information of the given pattern. We also demonstrate through experimentation that certain zones of the spectrogram have no discriminant information for most of the music genres.

Through a set of comprehensive experiments on the Latin Music Database [3], a very challenging dataset of 900 music pieces divided among 10 music genres, we demonstrate that the proposed approach compares favorably to the traditional approaches reported in the literature. In addition, the experimental results show that the classifier trained with the texture features produces complementary information which can be used to build more reliable music genre classification systems. When combining the results of the proposed system with another one based on short-term, low-level characteristics of the music audio signal [8] we were able to get an improvement of about 6% on the recognition rate.

## II. Texture Features

Since our approach is based on visual representation of the audio signal, the first step of the feature extraction process

Thanks to Fundação Araucária, CNPq, and CAPES for funding.

consists in converting the audio signal to a spectrogram. The spectrograms were created using a bit rate = 352kbps, audio sample size = 16 bits, one channel, and audio sample rate = 22.5kHz. In this work we have used the idea of time decomposition [9] 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 [3] which considers three 30-second segments from the beginning ($\overline{\Upsilon}_{beg}$), middle ($\overline{\Upsilon}_{mid}$), and end ($\overline{\Upsilon}_{end}$) parts of the original music. Figure 1 depicts this process.

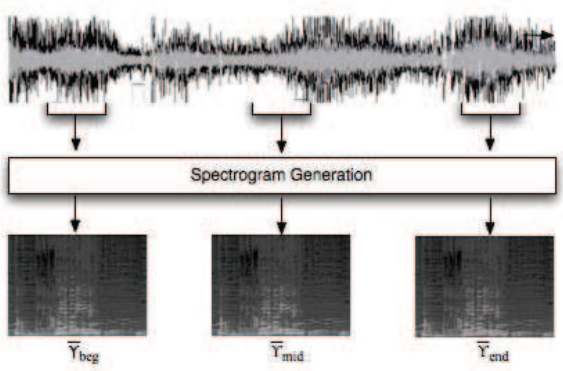


Figure 1. Creating spectrograms using time decomposition.

After generating the spectrograms, the next step consists in extracting the features from the images. As stated before, the approach proposed in this work considers the spectrogram as a texture and it uses the well-known GLCM texture descriptors as features. 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 $\theta$. Given a GLCM, some statistical information can be extracted from it. Assuming that $Ng$ is the gray level depth, and $p(i, j)$ is the probability of the co-occurrence of gray level $i$ and gray level $j$ observing consecutive pixels at distance $d$ and angle $\theta$, to describe the texture.

Haralick [7], the precursor of this technique suggested a set of 14 characteristics, but most of 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.

$$\text{Entropy} = -\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} p(i,j)\log(p(i,j)) \quad (1)$$

$$\text{Correlation} = \frac{p(i,j) - \mu_x\mu_y}{\sigma_x^2\sigma_y^2} \quad (2)$$

where $\mu_x = \sum_{i=1}^{Ng} i \times p_x(i)]$, $p_x(i) = \sum_{j=1}^{Ng} p(i,j)$, $\sigma_x^2 = \sum_{i=1}^{Ng}(i-\mu_x)^2 p_x(i)$, $\mu_y = \sum_{j=1}^{Ng} j \times p_y(j)$, $p_y(j) = \sum_{i=1}^{Ng} p(i,j)$ $e$ $\sigma_y^2 = \sum_{j=1}^{Ng}(j-\mu_y)^2 p_y(j)$.

$$\text{Homogeneity} = \sum_{i=1}^{Ng}\sum_{j=1}^{Ng} \frac{p(i,j)}{1+(i-j)^2} \quad (3)$$

$$\text{3rd Order Momentum} = \sum_{i=1}^{Ng}\sum_{j=1}^{Ng}(i-j)^3 \times p(i,j) \quad (4)$$

$$\text{Maximum Likelyhood} = \max p(i,j) \quad (5)$$

$$\text{Contrast} = \sum_{i=1}^{Ng}\sum_{j=1}^{Ng}(i-j)^2\ p(i,j) \quad (6)$$

$$\text{Energy} = \sum_{i=1}^{Ng}\sum_{j=1}^{Ng}(p(i,j))^2 \quad (7)$$

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.

At this point we could have a piece of music represented by three 28-dimensional feature vectors. However, by analyzing the texture images, we have noticed that the texture produced by the spectrograms are not uniform, so it is important to consider a local feature extraction rather than a global one. With this in mind, we have used a zoning technique which is a simple but efficient way to obtain local information of a given pattern. Figure 2 depicts the zoning mechanism used in this work. We discuss in Section IV that several other configurations of zoning have been tried out but this one produced the better results. Therefore each spectrogram image is represented by ten 28-dimensional feature vectors, summing up 30 vectors for a music piece.

## III. Methodology used for Classification

The classifier used in this work was the Support Vector Machine (SVM) introduced by Vapnik in [10]. 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 yielded using a Gaussian kernel. Parameters $C$ and $\gamma$ were tuned using a grid search.

Training and classification were carried out using the standard 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. 2-3, 3-1, 1-2). Considering that 30 feature

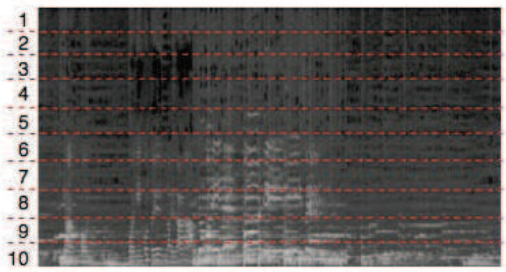


Figure 2. Zoning mechanism used to extract local information

vectors are extracted for each piece of music, for each fold we have 18,000 and 9,000 feature vectors for training and testing, respectively.

After training, the classification process is done as follows: The three 30-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 10 zones and one feature vector is extracted from each zone. The 28-dimensional feature vector is then sent to the classifier which assigns it to one of the 10 possible classes. The final decision is performed through a majority voting scheme as depicted in Figure 3, where each square represents the output of the classifier for each zone.

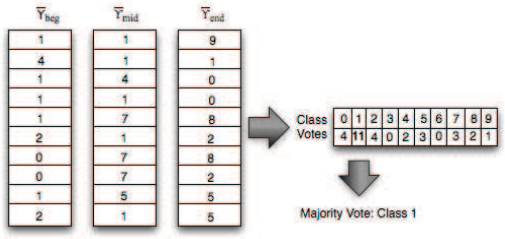


Figure 3. Voting mechanism used for classification

The rationale behind the zoning and voting 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 Figure 4 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. In this work we have investigated fixed zones. Other zoning strategies using different scales will be the subject of further investigation.

## IV. Experimental Results and Discussion

The experiments are carried out on a subset of the Latin Music Database (LMD) [3]. The LMD is made up of 3,227 full-length music pieces uniformly distributed along 10 classes of music genres: “Axe”, “Bachata”, “Bolero”, “Forro”, “Gaúcha”, “Merengue”, “Pagode”, “Salsa”, “Sertaneja”, and “Tango”. In our experiments we use 900 music pieces from the LMD,

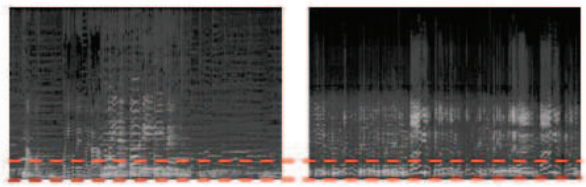

Figure 4. Spectrograms of different music genres with areas of similarity

which are split into 3 folds of equal size (30 music pieces per class). The splitting is done using an artist filter [11], 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.

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.

Before discussing the experiments it is important to mention that we have tried out different configurations of zoning. Figure 5 reports the average performance for the number of zones ranging from 1 to 40. As we can observe, after ten zones there is no improvement in terms of recognition rate.

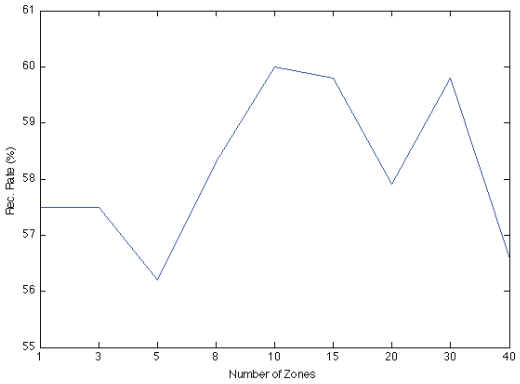


Figure 5. Evolution of the recognition rates for different number of zones

Table I reports the average recognition rate for 10 zones considering the three folds aforementioned. To have a better insight of these results we also report the results achieved by Lopes et al in [8]. In this case both works can be directly compared since they rely on the same experimental protocol. Lopes et al, though, use a different approach based on instances, which are feature vectors representing short-term, low-level characteristics of music audio signals.

Table I
AVERAGE RECOGNITION RATE (RR) FOR 10 ZONES WITH STANDARD DEVIATIONS ($\sigma$)

| Genre | RR(%) | $\sigma$ | RR (%) [8] | $\sigma$ |
|---|---|---|---|---|
| Axé | 73.3 | 8.8 | 61.1 | 13.4 |
| Bachata | 82.2 | 15.0 | 91.1 | 6.9 |
| Bolero | 64.4 | 8.4 | 72.2 | 9.6 |
| Forró | 65.5 | 8.4 | 17.7 | 17.1 |
| Gaúcha | 35.5 | 5.1 | 44.0 | 8.4 |
| Merengue | 80.0 | 6.6 | 78.8 | 11.7 |
| Pagode | 46.6 | 17.6 | 61.1 | 8.4 |
| Salsa | 42.2 | 6.9 | 40.0 | 12.0 |
| Sertaneja | 17.7 | 6.9 | 41.1 | 35.5 |
| Tango | 93.3 | 6.6 | 88.9 | 9.6 |
| Average | 60.1 | 9.0 | 59.6 | 13.4 |

As we can observe from Table I both classifiers perform very poorly for some music genres. In spite of the fact that both systems have similar recognition rates, they produce different confusions due to the different representations used to train the classifiers. From Table I we can see that the difference of performances can reach more than 10% for certain music genres. Sertaneja and Forró feature the biggest differences.

Such differences indicate some complementarity between both feature sets. As stated in the beginning of this work, our motivation was to explore a different representation for music genre expecting that it could be complementary to the traditional features reported in the literature. Table I indicates that the classifier trained with texture features offers some degree of complementarity to that developed by Lopes et al [8].

In order to show how complementary are both classifiers we combined our results with the results published by Lopes et al in [8]. Their approach is based on an instance selection method where a music piece is represented by 646 instances. The classifier used is a SVM and the final decision is also done through majority voting.

We converted the outputs of both systems into probabilities simply dividing the number of votes received for a given hypothesis from the classifiers by the total number of votes. With this in hand we have applied several different combination methods described by Kittler et al in [12]. The two methods that provided best results were Max and Sum, which are reported in Table II.

Table II
RESULTS USING MAX AND SUM COMBINATION RULES

| Genre | Max (%) | Sum (%) |
|---|---|---|
| Axé | 72.2 | 76.6 |
| Bachata | 87.7 | 87.7 |
| Bolero | 80.0 | 83.3 |
| Forró | 53.3 | 52.2 |
| Gaúcha | 47.7 | 48.8 |
| Merengue | 86.6 | 87.7 |
| Pagode | 56.6 | 61.1 |
| Salsa | 48.8 | 50.0 |
| Sertaneja | 34.4 | 34.4 |
| Tango | 90.0 | 90.0 |
| Average | 65.7 | 67.2 |

From Table II we can observe that the Sum rule explores better the diversity produced by the two different classifiers achieving an recognition rate of 67.2%. In other words, an improvement of 7 percentage points compared with the baseline system presented in Table I. Those results compare to the literature, specially to those published in the last MIREX contest on audio genre classification [13].

## V. 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 which are divide into zones so that features can be extracted locally. We have demonstrated that after 10 zones there is no further improvement in term of recognition rate.

Experiments combining two different systems using different combination strategies have proved that the proposed approach can provide complementary information to that provided by short-term, low-level characteristics of the music audio signal. Using the Max rule we were able to reach an improvement of about 7 percentage points in the recognition rate. Our future works will be focused towards the development and tests of other texture features and other strategies for zoning.

## REFERENCES


[1] J. J. Aucouturier, and F. Pachet, "Representing musical genre: A state of the art", *Journal of New Music Research*, vol. 32, no. 1, pp. 83-93, 2003.

[2] 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.

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

[4] T. Lidy, C. N. Silla Jr, 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.

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

[6] M. French, and R. Handy, "Spectrograms: turning signals into pictures", *Journal of Engineering Technology*, vol. 24, pp. 32-35, 2007.

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

[8] M. Lopes, F. Guyon, A. L. Koerich and L. S. Oliveira, "Selection of Training Instances for Music Genre Classification", in *20th International Conference on Pattern Recognition*, 2010.

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

[10] V. Vapnik, "The nature of statistical learning theory", *Springer-Verlag New York*, Inc, 1995.

[11] A. Flexer, "A closer look on artist filters for musical genre classification", in *International Conference on Music Information Retrieval*, pp. 341-347, 2007.

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

[13] Mirex, "Music Information Retrieval Evaluation Exchange", http://www.music-ir.org/mirex/wiki/2010:MIREX2010_Results, 2010.