Set of Texture Descriptors for Music Genre Classification

Loris Nanni
Department of Information Engineering
University of Padua
viale Gradenigo 6
35131, Padua, Italy
loris.nanni@unipd.it

Yandre Costa
State University of Maringa (UEM)
Av. Colombo, 5790
87020-900, Maringa, Parana, Brazil
yandre@din.uem.br

Sheryl Brahnam
Computer Information Systems
Missouri State University
901 S. National
Springfield, MO 65804, USA
sbrahnam@missouristate.edu

ABSTRACT

This paper presents a comparison among different texture descriptors and ensembles of descriptors for music genre classification. The features are extracted from the spectrogram calculated starting from the audio signal. The best results are obtained by extracting features from subwindows taken from the entire spectrogram by Mel scale zoning. To assess the performance of our method, two different databases are used: the Latin Music Database (LMD) and the ISMIR 2004 database. The best descriptors proposed in this work greatly outperform previous results using texture descriptors on both databases: we obtain 86.1% accuracy with LMD and 82.9% accuracy with ISMIR 2004. Our descriptors and the MATLAB code for all experiments reported in this paper will be available at https://www.dei.unipd.it/node/2357 .

Keywords

Music genre, texture, image processing, pattern recognition.

1 INTRODUCTION

The field of music genre classification has grown significantly since 2002, when Tzanetakis and Cook [Tza02a] first introduced music genre classification as a pattern recognition task. This interest can be explained by the exponential growth of information available on the internet [Gan08a], especially the massive amounts of digital music being uploaded daily, which is making it more necessary than ever for search engines, music databases, and other web services to automatically organize music for easy retrieval. Musical genre is one of the most common ways people think about and organize music, and it is probably the most widely used scheme for managing digital music databases [Auc03a]. Automatic music genre classification is thus becoming an increasingly important machine learning problem.

In 2011, Costa et al. [Cos11a] started investigating the use of features extracted from spectrogram images for music genre recognition, the rational being that the textural content in spectrogram images contains information useful for musical genre discrimination. Several works have since been published describing the performance of some well-known texture operators on spectrogram images (e.g., for papers using the gray-level co-occurrence matrix, see (GLCM) [Cos11a, Cos12b], for local binary patterns (LBP), see [Cos12a, Cos12b, Cos13a], for Gabor Filters, see [Wu11a, Cos13b], and for local phase quantization (LPQ), see [Cos13b]). These operators both preserve and do not preserve local information about the extracted features. In all these studies, the texture descriptors were used to train a support vector machine (SVM) to discriminate genre.

In this work we expand previous studies by comparing and combining more than ten texture descriptors, and for more robust comparison, two different databases are used: the Latin Music Database (LMD) [Sil08a] and the ISMIR 2004 [Gom06a] database. Very impressive results are reported on both databases, with some of our descriptor sets outperforming previous state-of-the-art approaches based on texture descriptors. In our comparative studies, we also present the performance of each descriptor extracted from the following: a) the entire spectrogram, b) different subwindows of the spectrogram obtained by linear zoning, and c) different subwindows of the spectrogram obtained by Mel scale zoning. In general, better performances are obtained using Mel scale zoning, where, for each subwindow, a different feature vector is extracted and used to train a dif-

ferent SVM; the set of SVMs is then combined by sum rule.

2 FEATURE EXTRACTION

In order to reduce the amount of signal to be processed in further steps, we first perform the time decomposition approach presented in [Cos04a], using three 10-second segments extracted from the beginning, middle, and end of the original audio signals, as depicted in Figure 1. After performing signal decomposition, the next step converts the audio signal into a spectrogram. A spectrogram describes how the spectrum of frequencies varies with time and can be described by a graph with two geometric dimensions: one where the horizontal axis represents time and the other where the vertical axis represents frequency. A third dimension describing the signal amplitude in a specific frequency at a particular time is represented by the intensity of each point in the image. For spectrogram generation, the Discrete Fourier Transform is computed with a window size of 1024 samples using the Hanning window function, which has good all-round frequency-resolution and dynamic-range properties.

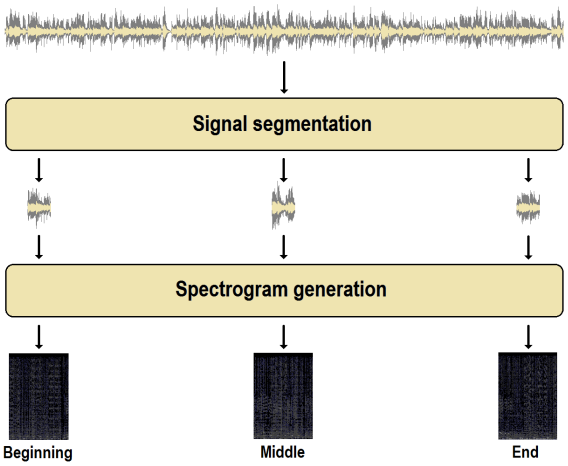

Figure 1: Mel scale zoning used to extract local information.

As described in previous works by Costa et al. [Cos11a, Cos12a, Cos12b], keeping some local information about the extracted features by zoning the spectrogram image is a good way to improve general performance in the classification task. Moreover, in [Cos12a] it was shown that a nonlinear image zoning, which takes into account frequency bands created according to the human perception of sound using the Mel scale [Ume99a], produces better results. Thus, in this work, we also examine results using Mel scale based zoning. In this case, 15 zones with different sizes are created in the region related to each one of the three segments originally extracted from the audio signal, which produces a total of 45 zones in the entire spectrogram image.

2.1 Global vs local

The texture descriptors are tested in three different ways:

- Global, where the features are extracted from the whole spectrogram;
- Linear, where the spectrogram is divided into 30 equal-sized subwindows and from each subwindow a different feature vector is extracted, as depicted in Figure 2;
- Mel, where the spectrogram is divided into 45 subwindows, as described previously, and from each subwindow a different feature vector is extracted. Figure 3 depicts this zoning scheme.

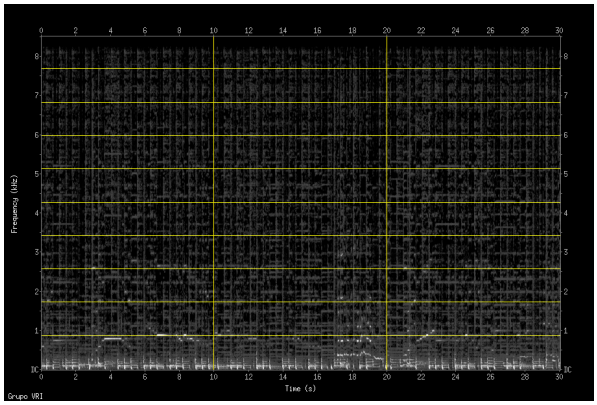

Figure 2: Linear zoning used to extract local information.

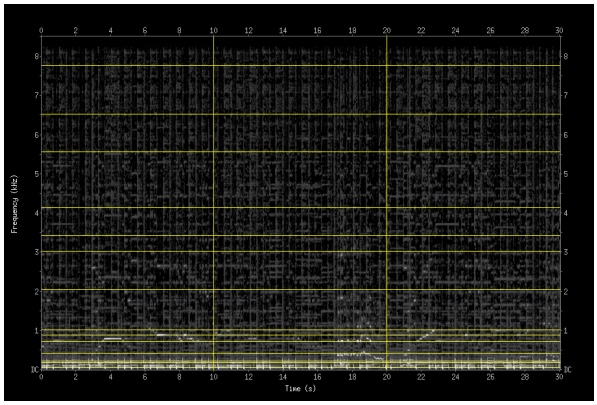

Figure 3: Mel scale zoning used to extract local information.

The features extracted with Linear/Mel are not concatenated and fed into one SVM as in Global. Rather an ensemble of 30/45 SVMs is trained (one for each subwindow), and the results of each SVM are then combined by sum rule.

2.2 Texture descriptors

The following approaches are compared in this paper[1]:

- LBP-HF [Zha12a], multi-scale LBP histogram Fourier feature vectors with radius 1 and 8 sampling points and with radius 2 and 16 sampling points. Details about this operator can be found in section 2.2.2;
- LPQ [Oja08a], multi-scale LPQ with radius 3 and 5. Details about this operator can be found in section 2.2.4;
- HOG [Dal05a], histogram of oriented gradients with number of cells = 5×6;
- LBP [Oja02a], multi-scale uniform LBP with radius 1 and 8 sampling points and with radius 2 and 16 sampling points. Details about this operator can be found in section 2.2.1;
- HARA [Har79a], Haralick texture features extracted from the spatial grey level dependence matrix;
- LCP [Guo11a], multi-scale linear configuration model with radius 1 and 8 sampling points and with radius 2 and 16 sampling points;
- NTLBP [Fat12a], multi-scale noise tolerant LBP with radius 1 and 8 sampling points and with radius 2 and 16 sampling points;
- DENSE [Yli12a], multi-scale densely sampled complete LBP histogram with radius 1 and 8 sampling points and with radius 2 and 16 sampling points;
- CoALBP [Nos12a], multi-scale co-occurrence of adjacent LBP with radius 1, 2 and 4;
- RICLBP [Nos12b], multi-scale rotation invariant co-occurrence of adjacent LBP with radius 1, 2 and 4. Details about this operator can be found in section 2.2.3;
- WLD [Che10a] , Weber law descriptor.

We use SVM with a radial basis function kernel for classification. For all approaches and for both datasets, we use the same SVM parameter set (to avoid the risk of overfitting since small training sets are used) where C=1000; gamma=0.1. Before the training step, the features are linearly normalized to [0,1].

Some of the texture operators aforementioned presented a noticeable performance in the results described in section 4. The next subsections present more details about these approaches.

[1] The MATLAB code we used is available so that misunderstandings in the parameter settings used for each method can be avoided (see abstract for MATLAB source code location).

2.2.1 LBP

The LBP texture operator was introduced by Ojala *et al.* in [Oja02a]. LBP takes into account for each pixel C, P neighbors equally spaced at a distance of R. LBP is an acronym that stands for Local Binary Pattern, these patterns are obtained taken into account the intensity differences of C and its P neighbors, and an histogram h of LBPs found in the image is used to describe the textural content of the image.

As stated by Mäenpää and Pietikäinen in [Mae05a], much of the information about the textural characteristics is preserved in the joint difference distribution:

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

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 neighbors 0 to $P-1$. The invariances to changes in the value of the central pixels when comparing with its neighbors is an important characteristic of this descriptor.

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. Equations 2 and 3 describe this.

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

where

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

By this way, 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 $0 \leq C' \leq 2^P$. Equation 4 describe how this code is obtained.

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

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

2.2.2 LBP-HF

In [Aho09a], Ahonen and Pietikäinen proposed a rotation invariant image descriptor based on uniform local binary patterns. The new approach was named Local Binary Pattern Histogram Fourier (LBP-HF). In this proposal, the Discrete Fourier transform (DFT) is used to extract a class of features that are invariant to rotation of the input image starting from the histogram rows of the uniform LBP patterns.

Let us denote a specific uniform LBP pattern by $U(n,r)$, it specifies an uniform pattern so that n is the number of 1-bits in the pattern and r is the rotation of the pattern. The uniform LBP histograms $h(U(n,r))$ is the number of occurrences of uniform pattern $U(n,r)$ in the image. The LBP-HF approach is based on the idea of applying the Discrete Fourier Transform (DFT) to the histogram of standard uniform LBPs, i.e:

$$H(n,u) = \sum_{r=0}^{P-1} h(U(n,r))e^{-\frac{i2\pi ur}{P}},\ 0 \leq u \leq P-1 \quad (5)$$

Finally, features are extracted using H, for details see [Aho09a].

2.2.3 RICLBP

To enhance the descriptive ability of LBP, it has been extended by introducing the concept of co-occurrence among LBPs, so that it is possible to extract information related to the global structures of the input image. The approach used in this paper, named Rotation Invariant Co-occurrence among adjacent LBPs (RICLBP),was proposed by Nosaka et al. [Nos12b]. RICLBP can simultaneously provide a high descriptive ability and invariance to image rotation. The basic idea is that LBP does not preserve structural information among binary patterns, and that such information could be useful for classifying the image. The Co-occurrence among adjacent LBP (LBP pair) at i ($i = (x,y)$ be a position vector in the image) is written as follows:

$$P(i,\Delta i) = (LBP(i), LBP(i+\Delta i)) \quad (6)$$

where $\Delta i = (i\ cos\theta, i\ sin\theta)$ is a displacement vector between an LBP pair. The value of i is an interval between an LBP pair, and $\theta = 0,\ \pi/4,\ \pi/2,\ 3\pi/4$.

The number of possible combination patterns of an LBP pair is significantly larger than that of the original LBP. The histogram feature generated from these LBP pairs contains information on the structure of the image, since it describes the frequency of LBP pairs that are located near to each other.

2.2.4 LPQ

Originally created to capture the textural content of blurred images, the Local Phase Quantization (LPQ) has shown good performance both on blurry and clear images. This operator is based on the blur invariance of the Fourier Transform Phase [Oja08b]. For each pixel, the blur insensitive information is found using the phase of 2D Short Term Fourier Transform (STFT) over a rectangular window.

Lets express $g(x)$, a blurred image resulted of the spatially invariant blurring of an original image $f(x)$, by

$$g(x) = f(x) * h(x) \quad (7)$$

where $x = [x,y]^T$ is the spatial cordinate vector and $h(x)$ is the point spread function. So, considering the Fourier space, one can express

$$G(u) = F(u).H(u) \quad (8)$$

where $G(u)$, $F(u)$, and $H(u)$ are the Discrete Fourier Transforms (DFT) of the blurred $g(x)$, $f(x)$ and $h(x)$, respectively, and $u = [u,v]^T$ is the frequency coordinate vector.

By this way, one can separate the magnitude from the phase with

$$|G(u)| = |F(u)| * |H(u)| \quad (9)$$

and

$$\angle G(u) = \angle F(u) * \angle H(u). \quad (10)$$

The Fourier transform is always real-valued when the blur $h(x)$ is centrally symmetric. Its phase is given by the following two-valued function

$$\angle H(u) = \begin{cases} 0 & \text{if } H(u) \geq 0 \\ \pi & \text{if } H(u) < 0 \end{cases} \quad (11)$$

so that $H(u)$ is positive at those frequencies where the original and the blurred image have the same phase. Taking into account the finite size of the observed image, it is known that the blurring invariance cannot be strictly achieved. If the image size is comparable to the blur size, the border effect causes a strong loss of information.

The aforementioned properties of blur invariance are the foundation of LPQ. From each image pixel position x of an image $f(x)$, a rectangular window N_x of size M by M is taken to calculate the local phase information using STFT:

$$F(u,x) = \sum_{y \in N_x} f(x-y)e^{-2\pi u^T y} = w_u^T f_x \quad (12)$$

where w_u is the 2-D DFT basis vector at frequency u, and f_x is a vector which contains all M^2 samples of image from N_x.

Four frequency vectors are considered on the LPQ operator: $u_1 = [a,0]^T$, $u_2 = [0,a]^T$, $u_3 = [a,a]^T$, and $u_4 = [a,-a]^T$, with a sufficiently small to last below the first zero crossing of $H(u)$ that safisfies

$$\angle G(u) = \angle F(u),\ for\ all \angle H(u) \geq 0. \quad (13)$$

If we put

$$F_x^c = [F(u_1,x), F(u_2,x), F(u_3,x), F(u_4,x)] \quad (14)$$

and

$$F_x = [Re\{F_x^c\},\ Im\{F_c^x\}]^T, \quad (15)$$

then the 8 by M^2 transform matrix is

$$W = [Re\{w_{u1},\ w_{u2},\ w_{u3},\ w_{u4}\},\ Im\{w_{u1},\ w_{u2},\ w_{u3},\ w_{u4}\}]^T \quad (16)$$

so,

$$F_x = W f_x. \quad (17)$$

In order to preserve the information as much as possible, the decorrelation of the coefficients need to be done before quantization.

Considering a Gaussian distribution, a whitening transform can achieve independence

$$G_x = V^T F_x \quad (18)$$

where V is an orthonormal matrix found by derivation from the singular value decomposition of the covariance matrix of the transform coefficient vector Fx.

G_x is computed for all the image positions. So, a quantization of the obtained vectors is done with the scalar quantizer:

$$q_j = \begin{cases} 1 & \text{if } g_j \geq 0 \\ 0 & \text{if } g_j < 0 \end{cases} \quad (19)$$

where the $j-th$ component of G_x is g_j. The following binary coding is used to turn the quantized coefficients in integers ranging from 0 to 255:

$$b = \sum_{j=1}^{8} q_j 2^{j-1} \quad (20)$$

Then, a feature vector is built with these integer values in order to be used in classification tasks.

3 MUSIC DATABASES

Our experiments are performed on the LMD and the ISMIR 2004 databases. These databases were chosen because they are among the most widely used in studies on music genre recognition; this makes comparing systems reported in the literature easier.

3.1 LMD

The Latin Music Database was specially created to support music information retrieval tasks. This database contains originally 3,227 music pieces assigned to 10 musical genres: axe, bachata, bolero, forro, gaucha, merengue, pagode, salsa, sertaneja, and tango. Training and classification experiments are carried out with LMD using a threefold cross-validation protocol. In this work, we decided to use the artist filter restriction [Fle07a], where all the music pieces of a specific artist are placed in one, and only one, fold of the dataset. As a result, a subset of 900 music pieces taken from the original dataset was used. This reduction is required since the distribution of music pieces per artist is far from uniform. The LMD results reported below refer to the average recognition rate obtained using the threefold cross-validation protocol.

3.2 ISMIR 2004

The ISMIR 2004 is one of the most widely used datasets in music information retrieval research. This database contains 1,458 music pieces assigned to six different genres: classical, electronic, jazz/blues, metal/punk, rock/pop, and world. The artist filter restriction cannot be used with this dataset as the number of music pieces per genre is not uniform. Due to the signal segmentation strategy used, it was also not possible to use all the music pieces: the training set used in our experiments is composed of 711 from the 728 music pieces originally provided and the testing set is composed with 713 from the 728 music pieces originally provided.

4 EXPERIMENTAL RESULTS

In tables 1 and 2, we compare our texture descriptors on both the LMD dataset (table 1) and on the ISMIR 2004 dataset (table 2). The following ensembles are also reported:

- F1, sum rule among LBP-HF, LPQ and LBP;
- F2, sum rule among LBP-HF, LPQ, LBP, RICLBP and DENSE;
- F3,sum rule among LBP-HF, LBP and RICLBP;
- WF, weighted sum rule among LBP-HF (weight 2), LBP (weight 3), and RICLBP (weight 1).

Examining tables 1 and 2, the following conclusions can be drawn:

- In both datasets the best stand-alone descriptor is the multi-scale uniform LBP;
- Mel typically outperforms Global and Linear;

METHOD	Global	Linear	Mel	Computation time(s)[1]
LBP-HF	74.2	79.4	82.8	0.141
LPQ	77.8	79.9	83.3	0.161
HOG	70.2	72.3	77.2	0.095
LBP	78.8	81.2	84.9	0.134
HARA	68.6	69.3	49.9	1.004
LCP	66.2	55.8	41.0	0.305
NTLBP	67.4	74.9	77.4	7.028
DENSE	77.4	80.8	84.1	0.596
CoALBP	69.3	67.0	77.1	0.289
RICLBP	77.6	80.8	84.3	0.464
WLD	67.9	69.9	71.7	0.767
F1	80.1	80.5	84.7	0.436
F2	80.3	81.6	84.3	1.496
F3	81.8	82.9	**86.1**	0.739
WF	81.5	82.6	**86.1**	0.739

[1] Computation time (seconds) coupled with Mel Using Matlab 2013a, CPU i5-3470 3.20 Ghz, 8GB RAM using the parallel toolbox.

Table 1: Performance on the LMD dataset.

METHOD	Global	Linear	Mel
LBP-HF	76.7	81.1	80.7
LPQ	78.3	80.6	80.5
HOG	74.3	70.7	72.1
LBP	80.5	81.1	81.4
HARA	72.1	76.3	77.3
LCP	73.2	4.6	42.9
NTLBP	72.4	74.9	76.2
DENSE	80.2	80.5	80.6
CoALBP	73.9	46.3	58.6
RICLBP	77.3	78.8	79.4
WLD	74.6	75.3	71.9
F1	**82.9**	80.9	82.0
F2	80.5	79.7	79.9
F3	81.9	80.8	80.9
WF	80.8	81.4	81.6

Table 2: Performance on the ISMIR 2004 dataset.

- The best result on both datasets is obtained by an ensemble of descriptors (F3 and WF in LMD and F1 in ISMIR 2004);
- The ensembles are mainly useful when a Global approach is used (note: this approach would be of value for reducing the computation time, e.g., when performing classification on a smartphone. Recall from subsection 2.1 that in Global, one SVM is trained for each descriptor, while Mel needs to train 45 SVMs for each descriptor).

In tables 3 and 4, our best approaches are compared with the state-of-the-art on both LMD and ISMIR 2004 datasets.

On the LMD dataset (table 3) our proposed ensemble outperforms all previous approaches when artist filter restriction is taken into account, while on the ISMIR 2004 dataset (table 4) our proposed ensemble outperforms previous works using texture descriptors (visual features), but it is outperformed by other approaches. Regarding these other approaches, it is important to underline the highly successful performance obtained using Block-level features, which are able to capture more temporal information than other features (see [Poh10a, Sey10a], for more details). The same can be said for LPNTF (Locality Preserving Non-negative Tensor Factorization), a multilinear subspace analysis technique (see [Pan09a], for more details). Both features are described here as acoustic features because

METHOD	**Accuracy (%)**
F1-Mel[1]	84.7
F3-Mel[1]	**86.1**
WF-Mel[1]	**86.1**
LBP-Mel[1][Cos12a]	82.3
LBP-Global[1][Cos12a]	79.0
GLCM[1][Cos12b]	70.7
LPQ[1][Cos13b]	80.8
Gabor filter[1][Cos13b]	74.7
MARSYAS features[2][Lop10a]	59.7
GSV-SVM+MFCC[2] [Cao09a] (MIREX 2009 winner)	74.7
Block-level[2] [Poh10a] (MIREX 2010 winner)	79.9
Principal Mel-spectrum components[2] [Ham11a] (MIREX 2011 winner)	82.3
Time Constrained Sequential Patterns[2] [Ren12a] (MIREX 2012 winner)	77.0
Multiple Rhythmic Signatures Patterns[2] [Pik13a] (MIREX 2013 winner)	77.6

[1] Visual features [2] Acoustic features

Table 3: Comparison with the state-of-the-art on the LMD dataset using artist filter restriction.

METHOD	**Accuracy (%)**
F1-Mel[1]	82.0
F1-Global[1]	82.9
F3-Mel[1]	80.9
Wf-Mel[1]	81.6
LBP-Mel[1][Cos12a]	76.7
LBP Global[1][Cos12a]	80.6
Gabor filter[1][Wu11a]	82.2
GSV+Gabor filter[3][Wu11a]	86.1
Block-level[2][Poh10a]	88.3
LPNTF[2] [Pan09a]	**94.4**

[1] Visual features [2] Acoustic features

[3] Visual plus acoustic features

Table 4: Comparison with the state-of-the-art on the ISMIR 2004 dataset.

they are extracted straight from the signal, without spectrogram generation.

The best results obtained in previous works that only used visual features (i.e. 82.3% [Cos12a] on LMD and 82.2% [Wu11a] on ISMIR 2004), however, were lower than those reported using our approach. Our proposed approach is very successful in its category, and produces the best reported result ever described on the LMD dataset using artist filter. Regarding the ISMIR 2004 dataset, our best result is not the best reported in the literature, but is the best one obtained using only visual features. Moreover, note that our proposed approach works well on both datasets without ad hoc tuning. The best previous work where visual features were tested on both datasets was [Cos12a]. In that work the best method for LMD (LBP-Mel) was different for the best method for ISMIR 2004 (LBP-global): here F1-Mel and F3-Mel outperform both these methods on both datasets.

5 CONCLUSION

In this work an examination of 10 different texture descriptors (and their combinations) for music genre classification is performed. Three different methods are tested for feature extraction: Global, Linear, and Mel, where the descriptors are extracted from 45 subwindows taken from the spectrogram, calculated starting from the audio signal and obtained with Mel scale zoning. For each subwindow, a different feature vector is extracted and a set of 45 SVMs are trained for each texture descriptor. This set of SVMs is then combined by sum rule.

The presented results are obtained on two well-known datasets (ISMIR 2004 and LMD) by combining different texture descriptors. Our ensembles outperform previous studies on both datasets using texture descriptors extracted from spectrogram. The best result obtained on the LMD dataset is the best ever obtained on this dataset considering the use of artist filter.

In the future, we plan on investigating bag-of-feature-based approaches. Moreover, we plan on coupling acoustic features with the ensemble propose in this paper (i.e., acoustic features + texture features) to see whether this combination enhances performance further.

6 REFERENCES

[Auc03a] Aucouturier, J.J., and Pachet, F. Representing musical genre: A state of the art. Journal of New Music Research, pp. 83-93, vol. 32, number 1, 2003.

[Aho09a] Ahonen, T., and Pietikäinen, M. Image description using joint distribution of filter bank responses. Pattern Recognition Letters, vol. 30, number 4, pp. 368-376, 2009.

[Cao09a] Cao, C., and Li, M. Thinkit's Submission for MIREX 2009 Audio Music Classification and Similarity Tasks (MIREX-09), International Conference on Music Information Retrieval, Kobe, Japan, 2009.

[Che10a] Chen, J., and Shan, S., and He, C., and Zhao, G., and Pietikäinen, M., and Chen, X. et al., WLD: A robust local image descriptor, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, pp. 1705-1720, 2010.

[Cos04a] Costa, C.H.L., and Valle Jr, J.D., and Koerich, A.L. Automatic Classification of Audio Data. International Conference on Systems, Man, and Cybernetics, pp. 562-567, The Hague, The Neterlands, 2004.

[Cos11a] Costa, Y. M. G., and Oliveira, L. E. S., and Koerich, A. L., and Gouyon, F. Music Genre Recognition Using Spectrograms. 18th International Conference on Systems, Signals and Image Processing, pp. 151-154, Sarajevo, Bosnia and Herzegovina, IEEE Press, 2011.

[Cos12a] Costa, Y. M. G., and Oliveira, L. E. S., and Koerich, A. L., and Gouyon, F., and Martins, J. G. Music Genre Classification Using LBP Textural Features. Signal Processing, vol. 92, number 11, pp. 2723-2737, 2012.

[Cos12b] Costa, Y. M. G., and Oliveira, L. E. S., and Koerich, A. L., and Gouyon, F. Comparing Textural Features for Music Genre Classification. IEEE World Congress on Computational Intelligence, pp. 1867-1872, Brisbane, Australia, IEEE Press, 2012.

[Cos13a] Costa, Y. M. G., and Oliveira, L. E. S., and Koerich, A. L., and Gouyon, F. Music Genre Recognition Based on Visual Features with Dynamic Ensemble of Classifiers Selection. 20th International Conference on Systems, Signals and Image Processing, pp. X-Y, Bucharest, Romania, IEEE Press, 2013.

[Cos13b] Costa, Y. M. G., and Oliveira, L. E. S., and Koerich, A. L., and Gouyon, F. Music Genre Recognition Using Gabor Filters and LPQ Texture Descriptors. 18 th Iberoamerican Congress on Pattern Recognition, pp. 67-74, Havana, Cuba, 2013.

[Dal05a] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection, Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886-893, San Diego, USA, 2005.

[Fat12a] Fathi, A., and Naghsh-Nilchi, A. R. Noise tolerant local binary pattern operator for efficient texture analysis, Pattern Recognition Letters, vol. 33, pp. 1093-1100, 2012.

[Fle07a] Flexer, A. A closer look on artist filter for mu-

sical genre classification, 8th International Conference on Music Information Retrieval, pp. 341-344, Vienna, Austria, 2007.

[Gan08a] Gantz, J.F., and Chute, C., and Manfrediz, A., and Minton, S., and Reinsel, D., and Schlichting, W., and Toncheva, A.The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011. Technical report: International Data Corporation (IDC), 2008.

[Gom06a] Gomez, E., and Gouyon, F., and Herrera, P., and Koppenberger, M. and Ong, B., and Serra, X., and Streich, S., and Cano, P., and Wack, N. ISMIR 2004 Audio Description Contest, Technical Report, Music Technology Group - Universitat Pompeu Fabra, 2006.

[Guo11a] Guo, Y., and Zhao, G., and Pietikainen, M. Texture classification using a linear configuration model based descriptor, British Machine Vision Conference, pp. 1-10, Nottingham, UK, 2011.

[Ham11a] Hamel, P. Pooled Features Classification MIREX 2011. International Conference on Music Information Retrieval, Miami, USA, 2011.

[Har79a] Haralick, R. M. Statistical and structural approaches to texture, Proceedings of the IEEE, vol. 67, pp. 786-804, 1979.

[Mae05a] Mäenpää, T., and Pietikäinen, M. Texture analysis with local binary patterns. Handbook of pattern recognition and computer vision, vol. 3, pp. 197-216, 2005.

[Nos12a] Nosaka, R., and Ohkawa, Y., and Fukui, K. Feature extraction based on co-occurrence of adjacent local binary patterns, Lecture Notes in Computer Science - Advances in image and video technology, pp. 82-91, 2012.

[Nos12b] Nosaka, R., and Suryanto, C. H., and Fukui, K. Rotation invariant co-occurrence among adjacent LBPs, Asian Conference on Computer Vision, pp 15-25, Daejon, Korea, 2012.

[Lop10a] Lopes, M., and Gouyon, F., and Koerich, A. L., and Oliveira, L. E. S. Selection of training instances for music genre classification, 20th International Conference on Pattern Recognition, pp. 4569-4572, Istanbul, Turkey, 2010.

[Oja02a] Ojala, T., and Pietikainen, M., and Maeenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, pp. 971-987, 2002.

[Oja08a] Ojansivu, V., and Heikkila, J. Blur insensitive texture classification using local phase quantization. International Conference on Image and Signal Processing, pp. 236-243, Cherbourg-Octeville, France, 2008.

[Oja08b] Ojansivu, V., and Heikklä, J. Blur Insensitive Texture Classification Using Local Phase Quantization. Image and Signal Processing, Springer Berlin Heidelberg, pp. 236-43, 2008.

[Pan09a] Panagakis, Y., and Kotropoulos, C., and Arce, G. R. Music genre classification using locality preserving non-negative tensor factorization and sparse representations, 10th International Conference on Music Information Retrieval, pp. 249-254, Kobe, Japan, 2009.

[Pik13a] Pikrakis, A. A MIREX 2013 Submission Based on a Deep Learning Approach to Rhythm Modeling. 14th International Conference on Music Information Retrieval, Curitiba, Brazil, 2013.

[Poh10a] Pohle, T., and Seyerlehner, K., and Schnitzer, D. Audio Music Similarity and Retrieval Task of MIREX 2010, Utrecht, The Netherlands, 2010.

[Ren12a] Ren, J-M., and Wu, M-J., and Jang, J. S-R. Time Constrained Sequential Patterns for Music Genre Classification, Porto, Portugal, 2012.

[Sey10a] Seyerlehner, K., and Schedl, M., and Pohle, T., and Knees, P. Using block-level features for genre classification, tag classification and music similarity estimation, 6th Annual Music Information Retrieval Evaluation eXchange (MIREX-2010), Utrecht, The Netherlands, 2010.

[Sil08a] Silla, C.N., and Koerich, A. L., and Kaestner, C.A.A. The Latin Music Database. 9th International Conference on Music Information Retrieval, pp. 451-456, Philadelphia, USA, 2008.

[Tza02a] Tzanetakis, G., and Cook, P. Musical genre classification of audio signals. IEEE Transactions speech and audio processing, pp. 293-302, 2002.

[Ume99a] Umesh, S., and Cohen, L., and Nelson, D. Fitting the Mel Scale. International Conference on Acoustics, Speech, and Signal Processing, pp. 217-220, Phoenix, USA, 1999.

[Wu11a] Wu, M.J., and Chen, Z.S., and Jang, J.S.R., and Ren, J.M. and Li, Y.H., and Lu, C.H. Combining visual and acoustic features for music genre classification. International Conference on Machine Learning and Applications, vol. 2, pp. 124-129, Honolulu, Hawaii, 2011.

[Yli12a] Ylioinas, J., and Hadid, A., and Guo, Y., and Pietikäinen, M. Efficient image appearance description using dense sampling based local binary patterns, Asian Conference on Computer Vision, pp.375-388, Daejon, Korea, 2012.

[Zha12a] Zhao, G., and Ahonen, T., and Matas, J., and Pietikäinen, M. Rotation-invariant image and video description with local binary pattern features. IEEE Transactions on Image Processing, vol. 21, pp. 1465-1467, 2012.