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-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

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. 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.

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, as depicted in Figure 1.

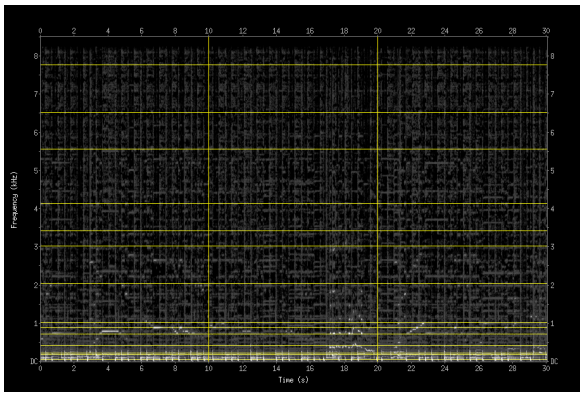

Figure 1: Mel scale zoning used to extract local information.

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;
- Mel, where the spectrogram is divided into 45 subwindows as described above and from each subwindow a different feature vector is extracted.

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;
- LPQ [Oja08a], multi-scale LPQ with radius 3 and 5;
- 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;
- 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;

[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).

- RICLBP [Nos12b], multi-scale rotation invariant co-occurrence of adjacent LBP with radius 1, 2 and 4;

- 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].

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).

METHOD	Global	Linear	Mel
LBP-HF	74.2	79.4	82.8
LPQ	77.8	79.9	83.3
HOG	70.2	72.3	77.2
LBP	78.8	81.2	84.9
HARA	68.6	69.3	49.9
LCP	66.2	55.8	41.0
NTLBP	67.4	74.9	77.4
DENSE	77.4	80.8	84.1
CoALBP	69.3	67.0	77.1
RICLBP	77.6	80.8	84.3
WLD	67.9	69.9	71.7
F1	80.1	80.5	84.7
F2	80.3	81.6	84.3
F3	81.8	82.9	**86.1**
WF	81.5	82.6	**86.1**

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.

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;
- 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.

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

[1] Visual features
[2] Acoustic features

Table 3: Comparison with the state-of-the-art in 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] [Sey10a] (MIREX 2009 winner)	82.7
Block-level[2][Poh10a] (MIREX 2010 winner)	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 in the ISMIR 2004 dataset.

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 [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 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 best 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.

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 ACKNOWLEDGMENTS

Our thanks to ...

7 REFERENCES

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

[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 (ISMIR), 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, volume 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 (SMC), 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 (IWSSIP), 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, volume 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 (WCCI-IJCNN), 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 (IWSSIP), 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 (CIARP), pp. xx-yy, Havana, Cuba, 2013.

[Dal05a] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection, Conference on Computer Vision and Pattern Recognition (CVPR), volume 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, volume 33, pp. 1093-1100, 2012.

[Fle07a] Flexer, A. A closer look on artist filter for musical genre classification, 8th International Conference on Music Information Retrieval (ISMIR), 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 (BMVC), pp. 1-10, Nottingham, UK, 2011.

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

[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 (ACCV), 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 (ICPR), 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 Transactions on Pattern Analysis

and Machine Intelligence, volume 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 (ICISP), pp. 236-243, Cherbourg-Octeville, France, 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 (ISMIR), pp. 249-254, Kobe, Japan, 2009.

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

[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 (ISMIR), pp. 451-456, Philadelphia, USA, 2008.

[Tza02a] Tzanetakis, G., and Cook, P. Musical genre classification of audio signals. IEEE Transactions on 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 (ICASSP), 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, volume 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 (ACCV), 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, volume 21, pp. 1465-1467, 2012.

Last page should be fully used by text, figures etc. Do not leave empty space, please.

Do not lock the PDF – additional text and info will be inserted, i.e. ISSN/ISBN etc.