Expert Systems With Applications 45 (2016) 108–117

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Combining visual and acoustic features for music genre classification

Loris Nanni[a,*], Yandre M.G. Costa[b], Alessandra Lumini[c], Moo Young Kim[d], Seung Ryul Baek[d]

[a] DEI, University of Padua, viale Gradenigo 6, Padua, Italy
[b] DIN, State University of Maringa (UEM), Maringa, PR, Brazil
[c] DISI, University of Bologna, Cesena, Italy
[d] DTIC, Sejong University, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Music genre
Texture
Image processing
Acoustic features
Ensemble of classifiers
Pattern recognition

ABSTRACT

Since musical genre is one of the most common ways used by people for managing digital music databases, music genre recognition is a crucial task, deep studied by the Music Information Retrieval (MIR) research community since 2002. In this work we present a novel and effective approach for automated musical genre recognition based on the fusion of different set of features. Both acoustic and visual features are considered, evaluated, compared and fused in a final ensemble which show classification accuracy comparable or even better than other state-of-the-art approaches. The visual features are locally extracted from sub-windows of the spectrogram taken by Mel scale zoning: the input signal is represented by its spectrogram which is divided in sub-windows in order to extract local features; feature extraction is performed by calculating texture descriptors and bag of features projections from each sub-window; the final decision is taken using an ensemble of SVM classifiers. In this work we show for the first time that a bag of feature approach can be effective in this problem. As the acoustic features are concerned, we propose an ensemble of heterogeneous classifiers for maximizing the performance that could be obtained starting from the acoustic features. First timbre features are obtained from the audio signal, second some statistical measures are calculated from the texture window and the modulation spectrum, third a feature selection is executed to increase the recognition performance and decrease the computational complexity. Finally, the resulting descriptors are classified by fusing the scores of heterogeneous classifiers (SVM and Random subspace of AdaBoost). The experimental evaluation is performed on three well-known databases: the Latin Music Database (LMD), the ISMIR 2004 database and the GTZAN genre collection. The reported performance of the proposed approach is very encouraging, since they outperform other state-of-the-art approaches, without any ad hoc parameter optimization (i.e. using the same ensemble of classifiers and parameters setting in all the three datasets). The advantage of using both visual and audio features is also proved by means of *Q*-statistics, which confirms that the two sets of features are partially independent and they are suitable to be fused together in a heterogeneous system. The MATLAB code of the ensemble of classifiers and for the visual features extraction will be publicly available (see footnote 1) to other researchers for future comparisons. The code for acoustic features is not available since it is used in a commercial system.

1. Introduction

Music genre recognition was originally introduced as a pattern recognition task by Tzanetakis and Cook (2002). Since then, many works related to this field have been presented by the Music Information Retrieval (MIR) research community. The large growth of the amount of data available on the internet (Gantz et al., 2008), which includes digital music, has motivated the development of these works. The need for tools which help in automatically organize music for easy retrieval, like search engines and music databases can explain this interest. There are a number of studies concerning audio content analysis using different features and methods. Automatic music genre recognition is a crucial task for a content based music information retrieval system. According to Aucouturier and Pachet (2003), musical genre is one of the most common ways used by people for managing digital music databases.[1]

From 2011, Costa, Oliveira, Koerich, and Gouyon (2011) have investigated the analysis of spectrogram image content searching for features useful for music genre recognition. Noticeably, texture is

* Corresponding author. Tel.: +39 3493511673.
E-mail addresses: loris.nanni@unipd.it (L. Nanni), yandre@din.uem.br (Y.M.G. Costa), alessandra.lumini@unibo.it (A. Lumini), mooyoung@sejong.ac.kr (M.Y. Kim), daey7979@nate.com (S.R. Baek).

[1] https://www.dei.unipd.it/node/2357

http://dx.doi.org/10.1016/j.eswa.2015.09.018
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

the main visual content found in a spectrogram image. Since then, some texture descriptors widely known in the image processing literature have been used to capture the content of these images: Gray-Level Co-occurrence Matrix (GLCM) has been applied in Costa et al. (2011, 2012a); results obtained using Local Binary Patterns (LBP) can be seen in Costa, Oliveira, Koerich, and Gouyon (2012a, 2013a); Costa, Oliveira, Koerich, Gouyon, and Martins (2012b). Gabor filters were tried in Costa, Oliveira, Koerich, and Gouyon (2013b); Wu et al., (2011); and Local Phase Quantization (LPQ) was experimented in Costa et al. (2013b). In all of the cases, the operators were experimented both using and not using image zoning in order to preserve local information about the extracted features. The genre classification was obtained using a Support Vector Machine (SVM) trained with the aforementioned features.

Some works published in last two years show that investigations related to musical genre in music information retrieval scenario are not exhausted, and they still remain as an active research topic, as one can note in the following:

- In Panagakis, Kotropoulos, and Arce (2014), the authors present a novel framework (i.e. joint sparse low-rank classification (JSLRR)) for music genre classification with the purpose of correct the noise and identify the subspace structures in data contaminated by outliers. In that work, a novel classifier is proposed, which is referred to as JSLRR-based classifier. Two special cases of this classifier are presented: the joint sparse representation-based classifier, and the low-rank representation-based. At the end, the experimental results show that JSLRR performs well in the music genre classification.
- In Srinivas, Roy, and Mohan (2014), the authors present music genre classification using On-line Dictionary Learning (ODL). They achieved an impressive accuracy rate of 99.41% on the LMD dataset using MARSYAS features and a sparsity based classifier. However, the artist filter restriction was not applied, probably because a large number of samples is required to construct the ODL. By this way, it is difficult to make a fair comparison between the obtained result and the state-of-the-art.
- In Schindler and Rauber (2015), the authors propose an audio–visual approach for music genre classification. For this, they explore affective visual information taken from music videos. Results show that a combination of the modalities can improve non-timbral and rhythmic features but show insignificant effects on high performing audio features.
- In Lee, Shin, Jang, Jang, and Yoon (2015), the authors aim to predict user favorite songs in a music recommendation system. For this purpose, they make use of features typically used in genre recognition tasks. The authors conclude that the proposed system can be applied to various audio devices, apps and services.
- Finally, in Sarkar and Saha (2015), the music signal is categorized according to its genre. For this, the authors decompose the audio signal to obtain the component reflecting the desired degree of local characteristics using empirical mode decomposition (EMD). The authors make experiments on the GTZAN dataset and claim that the proposed methodology is effective in comparison to the state-of-the-art (see results in Table 13). It is noteworthy, however, that the authors have tested their system in a single dataset with randomly created folds, consequently without artist filter restriction. Thus, more tests should be performed for confirming that their method works well in different datasets.

In this work, we expand previous studies based on texture descriptors extracted from the spectrogram calculated starting from the audio signal (Nanni, Costa, & Brahnam, 2014). Each spectrogram is divided in different sub-windows by Mel scale zoning. For each sub-window a set of descriptors is extracted and a different classifier is trained, then the classifiers outputs are combined by sum rule (Kittler, Hatef, Duin, & Matas, 2002).

Moreover, we combine our best set of texture descriptors with the acoustic features proposed in Lim, Lee, Jang, Lee, and Kim (2012), used for training a heterogeneous ensemble built by Support Vector Machine and a random subspace of Adaboost of neural networks.

A wide set of experiments is carried out over three benchmark databases in order to compare the performance obtained by varying several descriptors, different bag of feature approaches and different classifiers. The main contribution of this work is the design and evaluation of an ensemble of descriptors and classifiers, combined by weighted sum rule, that works very well in the datasets here tested (i.e. the Latin Music Database (LMD) (Silla, Koerich, & Kaestner, 2008), the ISMIR 2004 (Cano et al., 2006) database and the GTZAN genre collection (Tzanetakis & Cook, 2002) without an ad hoc optimization. Therefore we believe that the proposed system could be scalable to any music genre classification problem without requiring precise tuning. Very impressive results are reported on the three databases, with some of our visual descriptor sets outperforming previous state-of-the-art approaches based on texture descriptors. When the visual features are combined with acoustic features, performance comparable or better to the state-of-the-art approaches is obtained.

The main strengths of the proposed approach are the following:

- We propose to extract texture features from the spectrogram image of an audio signal and we show, in our experiments, improved performance with respect to the previous approaches based on visual features. Moreover, our ensemble of texture features, named EnsVis in Table 13, reaches results comparable also with standard audio approaches. The advantage of using (visual) texture features is related to the fact that they are partially independent, using the *Q*-statistic, from audio features, as proved by the results in Table 13, where the fusion of these different types of features gains better classification results than the single-type approaches.
- Please note that the system based on audio features (referenced as Lim et al., 2012 in Table 13) which is used as baseline approach is a commercial system with very high performance: therefore it is a valuable result that our final ensemble EnsVisAc outperforms this approach.
- The approach based on the extraction of visual features is implemented in MATLAB and made freely available to other researchers for future comparisons.

The main drawbacks of our approach with respect to other methods proposed in the literature, and based on audio features, are related to the increased computational cost needed for feature extraction.

2. Proposed approach: visual features

In this section we focus on music classification from its spectrogram representation and we propose an ensemble of texture descriptors and classifiers for maximizing the performance that could be obtained starting from the visual features. As shown Fig. 1, first the input signal is represented by its spectrogram (step 1), then the resulting image is divided in sub-windows (step 2) in order to extract local features, feature extraction is performed by calculating texture descriptors (step 3) and bag of features projections (step 4) from each sub-windows. The resulting descriptors are classified by SVM. Then the final decision is obtained by fusing the scores using the weighed sum rule (step 6).

In Fig. 1 the complete scheme of the system based on visual features is reported, while the single steps are detailed in the following sections.

2.1. Steps 1 and 2: spectrogram representation and subwindowing

In this work, the authors decided to use a signal segmentation strategy suggested by Costa, Valle Jr, and Koerich (2004). This strategy