Applied Soft Computing 52 (2017) 28–38

Contents lists available at ScienceDirect

Applied Soft Computing

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

An evaluation of Convolutional Neural Networks for music classification using spectrograms

Yandre M.G. Costa [a,*], Luiz S. Oliveira [b], Carlos N. Silla Jr. [c]

[a] PCC/DIN, State University of Maringá (UEM), Maringá, PR, Brazil
[b] PPGInf, Federal University of Paraná, Curitiba, PR, Brazil
[c] PPGIa, Pontifical Catholic University of Paraná, Curitiba, PR, Brazil

ARTICLE INFO

Article history:
Received 29 April 2016
Received in revised form 6 December 2016
Accepted 10 December 2016
Available online 19 December 2016

Keywords:
Music genre recognition
Pattern recognition
Neural network applications

ABSTRACT

Music genre recognition based on visual representation has been successfully explored over the last years. Classifiers trained with textural descriptors (e.g., Local Binary Patterns, Local Phase Quantization, and Gabor filters) extracted from the spectrograms have achieved state-of-the-art results on several music datasets. In this work, though, we argue that we can go further with the time-frequency analysis through the use of representation learning. To show that, we compare the results obtained with a Convolutional Neural Network (CNN) with the results obtained by using handcrafted features and SVM classifiers. In addition, we have performed experiments fusing the results obtained with learned features and handcrafted features to assess the complementarity between these representations for the music classification task. Experiments were conducted on three music databases with distinct characteristics, specifically a western music collection largely used in research benchmarks (ISMIR 2004 Database), a collection of Latin American music (LMD database), and a collection of field recordings of ethnic African music. Our experiments show that the CNN compares favorably to other classifiers in several scenarios, hence, it is a very interesting alternative for music genre recognition. Considering the African database, the CNN surpassed the handcrafted representations and also the state-of-the-art by a margin. In the case of the LMD database, the combination of CNN and Robust Local Binary Pattern achieved a recognition rate of 92%, which to the best of our knowledge, is the best result (using an artist filter) on this dataset so far. On the ISMIR 2004 dataset, although the CNN did not improve the state of the art, it performed better than the classifiers based individually on other kind of features.

1. Introduction

The recent literature shows that spectrograms obtained from audio signal have been successfully applied to musical genre classification [4,7,6,55], one of the most important tasks in music information retrieval (MIR). Since texture is the main visual content found in the spectrogram, different types of texture representations have been used to describe the content of these images, such as Gray-Level Co-Occurrence Matrix (GLCM) [4], Local Binary Patterns (LBP) [7], Local Phase Quantization [6], Gabor Filters [55,6], and Weber Local Descriptor (WLD) [33].

In all those works, the authors report good results most of the time using Support Vector Machine (SVM) classifiers or Multiple Classifier Systems (MCS) with different fusion strategies. We can also observe that in several occasions, classifiers trained with visual representations outperform the classifiers trained with acoustic features.

Since extracting features from the whole music piece can be prohibitive due to the required processing time, the number and size of music segments used in MIR have also been subject of investigation. Costa et al. [3] used three segments taken from the beginning, middle, and end parts of the music pieces. This strategy was also adopted in [47,7,6,4,32]. George and Shamir [10] achieved good results considering a 60 s-music segment taken from 0:30 to 1:30 of the original track. However, some databases make available only 30-s clips of the music [40], which limit the choice of size and location of the segment.

Considering that the selected music clip contains relevant information for a given MIR problem, it is undeniable that the aforementioned texture descriptors can offer a good representation to train machine learning classifiers. The recent literature corroborates to that. However, some researchers advocate that the main weakness of the current machine learning methods lies exactly on

* Corresponding author.
E-mail addresses: yandre@din.uem.br (Y.M.G. Costa), lesoliveira@inf.ufpr.br (L.S. Oliveira), carlos.sillajr@gmail.com (C.N. Silla Jr.).

http://dx.doi.org/10.1016/j.asoc.2016.12.024
1568-4946/

this feature engineering [1,21]. To them, machine learning algorithms should be less dependent on feature engineering by being able to extract and organize the discriminative information from the data. In other words, to learn the representation.

The idea of representation learning is not new and it can be found in the literature on several different flavors, such as single Layer Learning Models, Probabilistic Models, Auto-Encoders and Convolutional Neural Networks. A good review about representation learning can be found in [1]. Among the different methods of learning representations, the most common are the deep learning methods, which are formed by the composition of multiple non-linear transformations, with the goal of producing more useful representations.

It has only recently emerged as a viable alternative due to the appearance and popularization of the Graphic Processing Units (GPUs) which are capable of delivering high computational throughput at relatively low cost, achieved through their massively parallel architecture. Among the different architectures, the Convolutional Neural Network (CNN) introduced by LeCun in [22] has been widely used to achieve state-of-the-art in different pattern recognition problems [20,34]. In the case of texture classification it has not been different. Hafemann et al. [12] have shown that CNN is able to surpass traditional textural descriptors for images of microscopic and macroscopic texture. In the field of content-based music informatics, as pointed out by Humphrey et al. [16], research is dominated by hand-crafted feature design and despite the progress in several areas community efforts are yielding diminishing returns. Therefore, they advocate that MIR can somehow benefit of representation learning techniques.

The main contribution of this work is to investigate whether or not there is complementarity between features learned from a visual representation of the sound and other handcrafted features taken directly from the audio signal or from a time-frequency image of the sound. For this purpose, we compare the representation learning approach with a recent variation of LBP, the RLBP [56] and three types of acoustic features, Rhythm Patterns (RP), Statistical Spectrum Descriptors (SSD) and Rhythm Histograms (RH). The SVM classifier was used with these handcrafted features. To better assess the methods, western (ISMIR 2004 database) [2], Latin (Latin Music Database) [46], and ethnic (African music database)[1] music collections were used in this work. Classifiers were built with these different kinds of features, and their outputs were combined using a late fusion strategy, as depicted in Fig. 1.

A set of comprehensive experiments shows that the representation learning technique compares favorably to other classifiers in several scenarios. In the case of the LMD dataset, the combination of CNN and RLBP achieved 92% of recognition rate, which to the best of our knowledge, is the best result on this dataset so far (using the artist filter). Considering the ISMIR 2004 database, the literature shows higher accuracies by combining different types of representation (e.g., visual and acoustic), however, the CNN classifier performs better than all individual classifiers trained with visual representation. Finally, on the African database, the CNN surpassed the other representations and the state-of-the-art by a fair margin.

The remainder of this paper is organized as follows: Section 2 presents the related work on content-based music informatics that use the concept of representation learning. Section 3 describes the music databases used in the experiments. Section 4 describes the hand-crafted features and the CNN architecture. Section 5 presents the classifiers used in this work, while Section 6 reports all the experiments that have been carried out on music classification. Finally, the last section concludes this work.

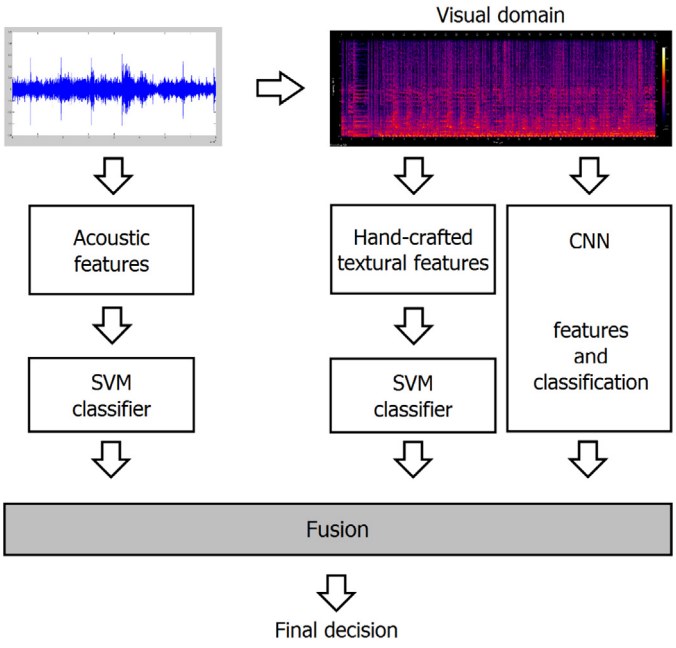

Fig. 1. An overview of the classification scheme.

2. Related work

In 2002, Tzanetakis and Cook presented music genre classification as a pattern recognition task [51]. In that work, the authors assessed the classification using acoustic features extracted from the sound on a dataset with 1000 music pieces labeled according to 10 musical genres. The authors also introduced a comprehensive set of features to describe music content (i.e. timbral texture features, rhythmic content features, and pitch content features), and the MARSYAS framework was made available to the music information retrieval research community.

Since then, the MARSYAS framework has been extensively used to extract acoustic features. Shen et al. [44] combined multiple acoustic features (timbre, rhythm, and pitch) extracted with the MARSYAS framework. In a pre-processing step, the authors used Principal Component Analysis to make a dimensionality reduction in the extracted feature vectors. After that, the features were concatenated to form a 25-dimensional feature vector. This feature vector was used as input to a three-layer neural network in order to perform a nonlinear dimensionality reduction. The neural network output layer had one unit for each class of the dataset. Further, the authors performed a human musical perception integration. The authors claim about the effectiveness of their method on music retrieval and about its novelty, since additional information (human perception) was included for the first time in the music retrieval task.

In the sense of using complementary information in music genre classification, Song and Zhang [48] proposed an information fusion framework for distance based music genre classification algorithms. The results obtained with such fusion performed better than those obtained with the single feature set (i.e. timbre and rhythm).

Valero and Alías [52] investigated the Gammatone Cepstral Coefficients (GTCC), an audio content descriptor already used in speech research field alternative to traditional acoustic features (i.e. Mel Frequency Cepstral Coefficient). The experiments were carried out on general sounds and audio scenes, and they concluded

[1] Kindly provided by Royal Museum of Central-Africa (RMCA).

that GTCC is more effective than MFCC in representing the spectral characteristics of non-speech audio signals.

Wang et al. [54] explored the music recommendation task based on user's general and contextual preferences from his/her playing records. For this purpose, a music embedding model was built from user's historical playing sequences. The user's historical playing sequence was treated as a "sentence" and every record of music was regarded as a "word", which were borrowed from the Natural Language Processing (NLP) domain. Experiments were conducted on a dataset collected from an online music service website, and the authors claim that the proposed approach is effective.

Similarly to the aforementioned works, the literature shows that the research on content-based music informatics is dominated by hand-crafted feature design [16,49]. However, the literature shows different attempts of using representation learning in this field.

Gwardys and Grzywczak [11] used a CNN trained on the Large Scale Visual Challenge (ILSVRC) as a feature extractor. In their case, they did not have enough data to train a classifier, but enough data was available in another domain of interest where the data could be in a different feature space or follow a different data distribution. Pan and Yang [36] showed that this knowledge transfer, also known as transfer learning, if done successfully, would greatly improve the performance of the learning algorithm hence avoid expensive data-labeling efforts. In [11], the authors generated three images of frequency spectrogram for each music track: one for the original music, one for the harmonic content and one for the percussive content. A 4096-dimensional vector was computed for each image using the CNN and three SVM were trained, one for each image. The final decision was provided by combining the SVM scores. They evaluated their work on five classes of the GTZAN dataset [51] and achieved an accuracy of 78%. In [43], the authors presented EMIF, an intelligent indexing framework designed to facilitate scalable and accurate content based music retrieval. In the indexing module, the authors generated a music signature using deep learning to combine various low-level acoustic features.

Sigtia and Dixon [45] trained a neural network with rectified linear units (ReLUs) using Stochastic Gradient Descent (SGD). They used the activations of the hidden layers of the neural networks as features and trained a Random Forest classifier on top of these features to predict the classes. The authors validated their ideas on two datasets, GTZAN and ISMIR 2004. In the former, they reported an accuracy of 83% while in the latter the best performance was 74.4% of recognition rate.

Nakashika et al. [31] presented a different approach. Instead of using the spectrogram images to train the CNN, they used Gray-Level Co-occurrence Matrices (GLCM) calculated from the Mel map. According to the authors, the GLCM has more efficient features for genre classification than the normal spectrogram. This architecture was assessed on 10 classes of the GTZAN dataset and reached 72% of accuracy, which does not compare to the state-of-the-art.

Feng [8] explored a 5-layer Restricted Boltzman Machine (RBM) on the GTZAN database. In [8] the 30-second audio file was represented by 2080 length of MFCC features. The best result reported in this work, 61% of recognition rate is quite far from the state-of-the-art, though.

Following a similar approach, Li et al. [24] also used feature learning to train the classifier. Instead of an RBM, the authors have employed a CNN with three convolutional layers to learn the representation. Then, the CNN was used as feature extractor to train a variety of decision tree classifiers available in the WEKA framework. Combining those classifiers using the majority voting rule the authors reported a classification accuracy of 84% on the GZTAN dataset. In [14], the authors applied a Deep Belief Network (DBN) on Discrete Fourier Transforms (DFTs) of the audio for representation learning. Experiments were conducted on the GTZAN dataset and the reported results were very encouraging, around 84% accuracy in the best scenario.

Still in the context of content-based music informatics, we found some research on representation learning in other applications such as chord recognition [17], and music onset detection [41]. In [17], the authors argue that representation learning using CNN is a viable alternative to the traditional approach of classifying short-time features and relying on post-filtering to smooth the results into a musically plausible chord path. Their experimental results corroborate to their initial hypothesis. Schluter and Bock [41] compared Recurrent Neural Networks (RNN) and CNN in the context of musical onset detection, which consists in finding the starting points of musically relevant events in audio data. They show that CNN perform comparable to the RNN with less manual preprocessing, but at higher computational cost. Finally, Hamel et al. [15] performed automatic annotation and ranking of music audio by analyzing the impact of the selection of pooling functions for summarization of the features over time.

3. Music databases

Since 2002, when Tzanetakis and Cook [51] presented music genre classification as a pattern recognition task, the focus of the research on this subject has been put preponderantly on music datasets from the western culture. However, some intriguing questions about the feasibility of using the proposed approaches on different cultural scenarios remain open. Although some authors, such as Lidy et al. [25] and Lee et al. [23], have done some effort in this direction, there is still a lack of investigations aiming to bring answers to these questions. For this reason, in this work we have decided to make the experiments on three quite different databases: a traditional western database (i.e. ISMIR 2004); a Latin Music Database (i.e. LMD), with some genres taken from countries with very similar cultural aspects; and the third one, an African music collection, whose recordings were collected in the field and did not use any studio resources. To make this work self-contained, we describe these three datasets in the following sub-sections.

3.1. Latin Music Database

The Latin Music Database (LMD) [46] contains 3227 full-length music pieces of 501 different artists. The database is uniformly distributed along 10 musical genres: Axé, Bachata, Bolero, Forró, Gaúcha, Merengue, Pagode, Salsa, Sertaneja, and Tango. The dataset brings together many genres with a significant similarity among themselves with regard to instrumentation, rhythmic structure, and harmonic content. 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 the 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 the protocol proposed in [30], which considers 900 music pieces from the LMD split into 3 folds of equal size (30 music pieces per genre). The splitting is done using an artist filter [9], 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 database since the distribution of music pieces per artist is far from uniform. It is worth mentioning that the artist filter makes the classification task harder.

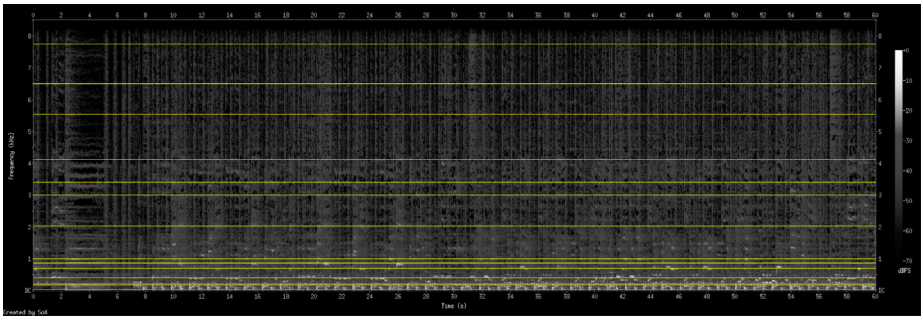

Fig. 2. Mel scale zoning superimposed over a spectrogram.

Table 1
Classes and number of samples in the ISMIR 2004 dataset.

Classes	Samples
Classical	640
Electronic	229
Jazz/Blues	52
Metal/Punk	90
Rock/Pop	203
World	244
Total	1458

3.2. ISMIR 2004

The ISMIR 2004 [2] dataset contains 1458 music pieces divided into training (50%) and testing (50%). It was proposed for the music information retrieval contest organized by the Music Technology Group of Pompeu Fabra University and it is composed of music pieces of six different genres: classical, electronic, jazz/blues, metal/punk, rock/pop, and world. The distribution is not uniform, and the training and test sets are predefined. It was not possible to use artist filter with this database because there is no information about the performing artist of each music piece. The number of samples per class in the ISMIR 2004 dataset is reported in Table 1.

3.3. African music database

As mentioned before, we have decided to present in this work a comprehensive study regarding the characterization of musical content using spectrograms. For this reason, we also used the collection of African music of the Royal Museum of Central-Africa (RMCA)[2] in Belgium, kindly provided by the museum. The music pieces of this database are not previously classified in terms of genre. Genre is a typical musical attribute in western music collections, but it is not necessarily true when we consider non-western music collections. In the African music database, the music pieces can be classified according to different categories such as country, function, ethnic group or instrumentation.

The choice of this database was based on the fact that the characteristics of the musical content in this collection are quite different when compared with typical western music databases. By using this database we can verify the feasibility of the visual-spectrogram technique in music classification tasks in different scenarios.

The experiments using the African music database were carried out using 10-fold cross-validation. In order to ensure that all the folds had at least one music piece of each class, we have discarded the classes in which there were not at least 10 music pieces considering each different classification type (i.e. country, function, ethnic group, and instrumentation). Table 2 describes the number of music pieces per class according to the different types of categories of each class in all the classification types of the African collection.

[2] http://www.africamuseum.be.

4. Feature extraction

In order to perform the visual feature extraction, we need to represent the original audio signal in the visual domain. For this reason, we have built spectrograms from the audio signal content. In all cases, we have used only one audio channel. The audio sample size is 16 for both LMD and ISMIR 2004. Regarding the bit rate, and audio sample size, the LMD and the ISMIR 2004 databases do not have the same values considering the original form in which they are available. The bit rate in the LMD is equal to 352 kbps, while in the ISMIR 2004 it is equal to 706 kbps. The audio sample rate in the LMD is 22.05 kHz, while in the ISMIR 2004 it is 44.1 kHz. In the African music database the audio files technical features are not standardized, since these recordings were taken in the field along many years. However, we have noticed that in the music pieces of this database the bit rate ranges from 705 kbps to 1536 kbps, and the audio sample rate is always of 44.1 kHz. In all cases, the Discrete Fourier Transform was computed with a window size of 1024 samples using the Hann window function which has good all-round frequency-resolution and dynamic-range properties.

Having in mind the good results achieved by George and Shamir [10], where they used a single 60 s-music segment taken from 0:30 to 1:30 of the original track, we have decided to adopt the same strategy. It should be noted that not all songs in the different databases have a duration of at least 90 s. Therefore, in order to use as many songs as possible from the different databases the following strategy was adopted. For music pieces with more than 60 s and less than 90 s, we consider the 60-s around the middle point of the music (i.e. from the middle −30 to middle +30). For music pieces with less than 60 s, we took the whole music piece content. In the next subsections we describe both hand-crafted features and also the CNN architecture used to learn the representation.

4.1. Hand-crafted features

Regarding the hand-crafted features, two different types of descriptors were used. The first one is the Robust LBP (RLBP), a variation of LBP that has been successfully applied to different texture problems [56]. The second is a family of acoustic features that has been successfully applied to the African music database.

4.1.1. RLBP

The hand-crafted features taken in the visual domain were obtained considering a Mel-scale zoning of the images, as made in [7]. The Mel scale represents the frequency bands according to the human perception. Fig. 2 shows this division superimposed over a spectrogram image used in this work.

Table 2
Classes and number of music pieces in the African music collection.

Country		Ethnic		Function		Instrumentation	
Burundi	17	Agni	18	Birth	19	Aerophone	56
Congo DRC	479	Dagomba	28	Cattle	44	Aerophone+	12
Ethiopia	25	Fanti	14	Court	45	Idiophone	
Gabon	11	Hutu	29	Dance	112	Aerophone+	19
Ghana	42	Luba	348	Entertainment	178	Idiophone+	
Ivory Coast	18	Mbuti	18	Festive	48	Membranophone	
Republic of	18	Ntandu	12	Funeral	21	Cordophone	152
the Congo		Sala	59	Historical	25	Cordophone+	27
Rwanda	402	Mpasu		Hunting	68	Idiophone	
Senegal	10	Tutsi	36	Lullaby	21	Idiophone	253
		Twa	49	Mourning	26	Idiophone+	147
		Wolof	10	Narrative	72	Membranophone	
				Praise	54	Membranophone	36
				Religious	14		
				Ritual	79		
				Transmitting Message	21		
				War	22		
				Wedding	22		
				Work	13		
Total	1022	Total	621	Total	904	Total	702

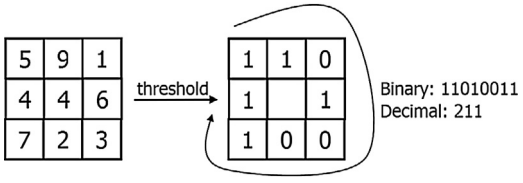

Fig. 3. Original LBP operator, extracted from [29].

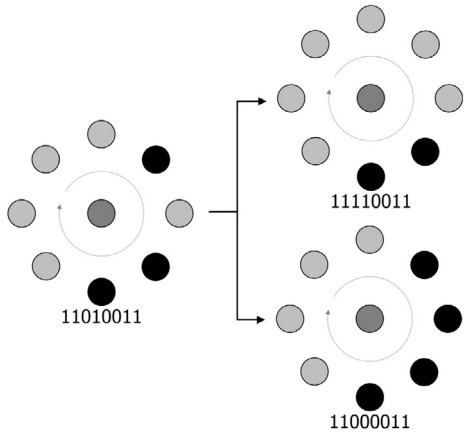

Fig. 4. Non-uniform LBP taken as uniform patterns in RLBP.

LBP has been one of the most powerful and successful texture descriptors used in works described in the literature in the last ten years. With LBP, a binary pattern is taken from each pixel of the image. The binary code is built considering the differences between the pixel and its equally spaced neighbors according to previously defined distance. In this case, if the difference between the pixel and a neighbor is bigger or equal to 0, the position in the binary code is set to 1, otherwise it is set to 0. Fig. 3 illustrates this process, in which the LBP is calculated by thresholding a 3×3 neighborhood of the central pixel.

In [35], the authors presented the concept of uniform LBP where a LBP is considered uniform if the number of transactions from 0 to 1, or from 1 to 0, in the binary code is less than or equal to 2, considering that the code is seen as a circular list. For example, the LBP code seen on the left side of Fig. 4 is a non-uniform pattern. According to Ojala et al. [35], uniform patterns provide better results because of their statistical properties.

Despite the good performance of LBP, researchers have continued looking for novel ways to improve it. In the work of Chen et al. [56] the Robust Local Binary Pattern (RLBP) is presented. In the case of the RLBP, Chen et al. considered a more flexible concept of uniformity. They claim that if there is one, and only one, value in the binary code which makes the LBP non-uniform, it is possibly caused by some noise and, for this reason, it must be considered as a uniform pattern. Fig. 4 shows a non-uniform LBP code that must be taken as uniform codes in RLBP.

Recent works have shown that RLBP is able to achieve better results than LBP in some circumstances [56]. In this work, as far as we know, RLBP is used for the first time to describe the textural content of the spectrogram for the task of music content categorization.

4.1.2. Acoustic features

In this work we have employed three different feature sets that extract features from the audio signals, namely Rhythm Patterns (RP), Statistical Spectrum Descriptors (SSD) and Rhythm Histograms (RH).

The Rhythm Patterns (RP) were originally proposed in [38] and enhanced in [39] to incorporate psycho-acoustic models [57]. In order to extract the RPs, the audio signal is segmented into a series of 6 s sequences and a spectrogram is generated using the short time fast Fourier transforms (STFT) with a Hanning window function of 23 ms and 50% overlap. After the spectrogram is generated, the Bark scale [57] is used to aggregate the critical band frequencies of the spectrogram into 24 frequency bands. The spectrum energy values of the Bark scale critical bands are then transformed into the Decibel scale (dB) to compute the loudness level curves (using the Phon scale) and the specific loudness sensation per critical band (using the Sone scale). These processing steps produce a Bark scale Sonogram that reflects the specific loudness sensation of an audio segment by the human ear [25]. A fast Fourier transform is then applied to the Bark scale Sonogram in order to obtain a time-invariant representation of the 24 critical bands. A gradient filter and Gaussian smoothing are then applied to improve the similarity between the rhythm patterns. The final feature matrix is obtained by computing the median of segment wise rhythm patterns.

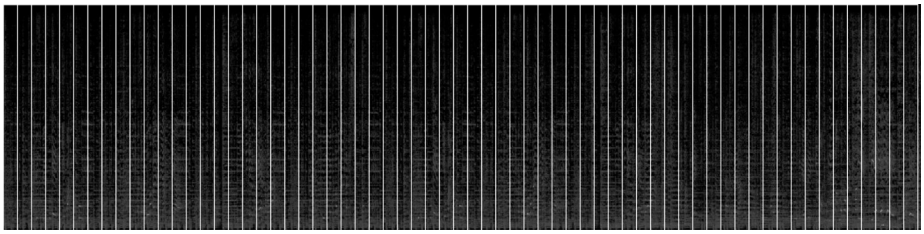

Fig. 5. Spectrogram divided into patches.

The Statistical Spectrum Descriptors (SSD) also employ the Bark scale to aggregate the critical band frequencies into 24 frequency bands. For each frequency band, it computes the statistical measures of mean, median, variance, skewness, kurtosis, min-value and max-value.

The Rhythm Histogram (RH) contains 60 bins, one for each modulation frequency between 0 Hz and 10 Hz, and the value of each bin is computed by summing the corresponding modulation amplitude value of each of the 24 critical frequency bands of the Bark scale spectrogram.

4.2. Representation learning

In spite of the growing popularity of the CNN, it has some constraints. First, it does not deal quite well with high-resolution images and second, it requires a certain amount of samples for training due to the huge number of parameters that must be adjusted during the training phase. To deal with these problems, we first downscale the high-resolution spectrogram images in 50% and then split the image into several patches (sub-images). The resized image has 256×800 pixels.

Then, the resized image can be divided into several patches of 256×16 depending on the overlap between patches. Fig. 5 shows the patches extracted from spectrogram with no overlap.

The deep neural network architecture used in this study was based on the model that achieved high levels of accuracy on different pattern recognition tasks. In summary, it contains repeated use of convolutional layers with 64 filters followed by max-pooling layers as used in Hafemann et al. [12]. A fully connected layer at the end is responsible for the classification. The architecture is illustrated in Fig. 6.

The input is a patch of 256×16. The convolutional layers have trainable filters that are applied across the entire image. The definition of the layers includes the filter size and stride (distance between the application of the filters). If the stride is smaller than the filter size, the filter will be applied in overlapping windows. In this study, the best results were achieved using 5×5 kernels with stride 1.

The pooling layers implement a linear downsampling function to reduce dimensionality and capture small translation invariances. In our experiments, different kernels and strides were used but the best results always were achieved with window size 2×2 and stride 2. Similarly to [12], the fully-connected layers are the standard for neural networks and connect, using unshared weights, all the neurons from one layer to the next one. In this case, it uses softmax activation.

Given the nature of the spectrogram image, it makes sense using kernels wide in time and narrow in frequency as in [17,41]. We have tried this kind of strategy, but the architecture of Fig. 6 with squared filters produced better results in our experiments. One of the advantages of using a representation learning approach is not requiring the design of feature extractors by a domain expert, but instead let the model learn them.

The CNN was trained using the Stochastic Gradient Descent (SGD) using back-propagation with for 80 epochs mini-batches of 128 instances, momentum factor of 0.9 and weight_decay of 5×10^{-4}. The learning rate is set to 10^{-3} in the beginning to make the weights quickly fit the long ravines in the weight space, then it is reduced over the time (until 5×10^{-4}) to make the weights fit the sharp curvatures. The network makes use of the well known cross-entropy loss function.

In order to monitor the generalization performance and select the best training model, we have divided the original training set into training set (70%) and validation set (30%). During training, the performance of the network on the training set will continue to improve, but its performance on the validation set will only improve to a point, where the network starts to overfit the training data, that the learning algorithm is terminated. To implement the CNN models we have used the Caffe framework [18] on a Tesla C2050 GPU.

5. Classification

The classification with hand-crafted features was performed using Support Vector Machine (SVM), presented by Vapnik in [53]. SVM was chosen based on our previous works on music genre recognition [4,7,6,5]. SVM is the classifier that provides the best results, hence, it has been selected for our experiments in this work. The normalization was done so that each attribute value ranges from −1 to +1. The results presented here were obtained by using a RBF kernel, where the parameters C and γ were determined through a grid search.

Regarding the three acoustic features investigated in this work, one feature vector was extracted from each audio sample. The RH feature dimensionality is 60, the RP feature vector dimensionality is 1380, and the SSD feature vector dimensionality is 161.

The classification with visual features was performed as described in the following: firstly, the 60-s segment of the audio signal was converted into a visual representation (spectrogram). Following, feature vectors were extracted considering the subwindows depicted in Fig. 2. In this way, one 59-dimensional RLBP feature vector is taken from each subwindow and a unique SVM is built for each subwindow. After that, each classifier provides one prediction for each class. Thus, we used some well-known fusion rules, presented by Kittler et al. [19], in order to get the combination of the classifiers outputs. The fusion rules used were Sum, Product, Max, and Min. As described in Section 6, one can see that, in general, Sum or Product rules provided better results.

The subwindowing strategy is based on the fact that by zoning the images we can preserve some useful local information that can potentially help to better discriminate different music genres. As claimed in [7], by zoning the images we can extract local information and try to highlight the specificities of each music genre. The reason for this is that classifiers specialized in specific regions of the image are more able to capture specificities, at times related to instruments, at times related to rhythmic patterns. Fig. 7 presents

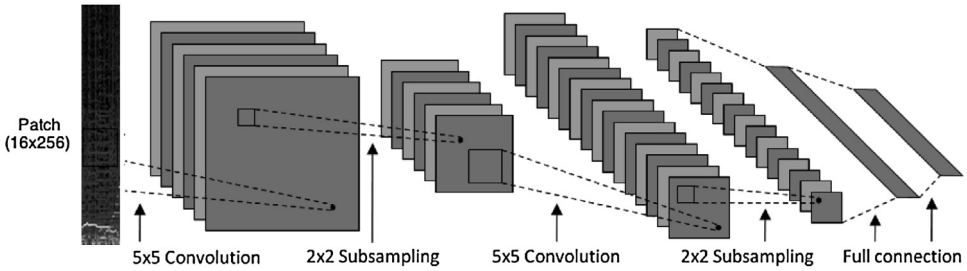

Fig. 6. The deep Convolutional Neural Network architecture.

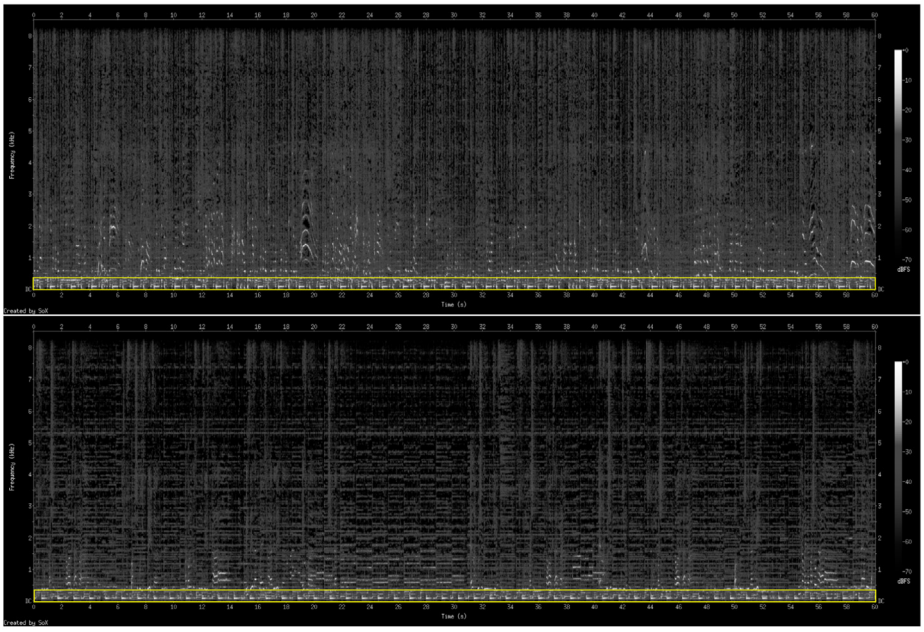

Fig. 7. Spectrograms of different music genres of LMD (Axé in the top and Forró in the bottom) with some areas of similarity at low frequencies.

two spectrograms taken from different music genres but with high similarity between the texture content at low frequencies.

Since the models are trained on image patches, a strategy to divide the testing images into patches is necessary. In this work, we have applied the same strategy used for creating images for training (with no overlap). During classification, the CNN assigns a probability for each possible class given a patch image. Those probabilities can be combined in several different ways. In our study, the best results were always produced by the Sum rule. Therefore, the prediction for a given test image is the class that maximizes the sum of the probabilities on all patches of the image.

6. Experimental results

The following subsections present the experiments carried out on the three databases introduced in Section 3. For each database we present the performance for each representation and also the combination of these representations using a late fusion strategy. In all cases, the performances were assessed with accuracy and *F*-measure.

Remember that the performance of the RLBP and CNN is a combination of several individual decisions. In the case of the RLBP, 15 SVMs (one for each frequency band in the Mel scale) are combined to produce a final decision using some well-known fusion rules, with low computational cost, presented by Kittler et al. [19], as pointed out in Section 5. In the case of the CNN the final decision is provided by combining the classification result for each image patch.

Table 3
Recognition accuracy (%) with standard deviation (σ) and *F*-measure (%) individually obtained using each kind of feature on the LMD dataset.

Feature		1st-level fusion	Accuracy (%)	*F*-measure (%)
Acoustic	RH	–	52.0 ± 2.3	51.2
	RP	–	67.0 ± 1.0	66.9
	SSD	–	63.3 ± 0.7	63.2
Visual	CNN	Sum	83.0 ± 2.2	83.6
	RLBP	Product	87.4 ± 1.1	**87.5**

To avoid any confusion, in the following subsection we refer to this fusion as 1st-level fusion. The late fusion, where different classifiers and representations are combined, is referred to as 2nd-level fusion. In the case of 2nd-level fusion, we have assessed different fusion rules (presented in Section 5) but the best results were always produced by the Sum or Product rules.

6.1. Results on LMD

Table 3 shows the results obtained on the LMD dataset considering each kind of hand-crafted feature (both acoustic and visual) and the results obtained with CNN. As stated previously, the cross validation was carried out with factor 3 in the experiments on the LMD because we decided to apply the artist filter. Taking into account that the number of music pieces per artist is not balanced in this

Table 4
State-of-the-art on the LMD database using the artist filter.

Authors	Features	Accuracy (%)
Lopes et al. [28]	Acoustic with instance selection	59.6
Hamel [13]	Principal Mel-spectrum Components	82.3
Costa et al. [7]	Visual features (LBP)	82.3
Nanni et al. [33]	Visual (LPB-HF, LBP, RICLBP) and Acoustic (MFCC, DFB, OSC)	85.1
Nanni et al. [33]	Visual features (LPB-HF, LBP, RICLBP)	**86.1**

Table 5
Recognition accuracy (%) with standard deviation (σ) and *F*-measure (%) obtained on the LMD dataset by combining different representations.

Representations	2nd-level fusion	Accuracy (%)	*F*-measure (%)
RP, RH, SSD	Sum	68.1 ± 2.4	67.9
RP, CNN, RLBP	Sum	91.1 ± 0.6	90.9
SSD, CNN, RLBP	Sum	91.1 ± 1.3	91.2
RH, CNN, RLBP	Sum	91.2 ± 1.5	91.3
CNN, RLBP	Product	92.0 ± 1.6	**92.0**

dataset, it was not possible to create more than three-fold. The results described here refer to the three folds average.

The best result (87.4% accuracy) was achieved by SVM trained with RLPB, followed by the CNN with 83% (accuracy). Comparing these results with the state-of-the-art (Table 4) we may notice that the RLBP outperforms all the methods reported in the literature that use artist filter. The classifiers trained with acoustic features, on the other hand, produced very poor performance compared to the classifiers trained with visual features.

By analyzing the confusion matrices, we noticed some complementarity between the classifiers based on different features, given that the classification errors were not necessarily the same. With that in mind, we decided to combine the results obtained individually using a late fusion strategy. Table 5 shows the best results obtained by fusing the outputs of the classifiers reported in Table 3.

The results presented in Table 5 show that better accuracy rates can be obtained by combining the classifiers based on both visual representations. When combined, these two representations are able to surpass the best result reported in the literature (Table 4) in about six percentage points. To the best of our knowledge, this is the best result obtained on the LMD using the artist filter. The addition of the classifiers trained with acoustic features in the combination did not bring any improvement given their poor performance on the LMD dataset.

To better analyze these results, we performed the Wilcoxon Signed Rank Test. It shows that both RLBP and CNN are better than the classifiers trained with acoustic features, but there is no statistical difference between RLBP and CNN (p-value = 0.2377 for $\alpha = 0.05$). The combination CNN-RLBP is significantly better than all single classifiers reported in Table 3 and also better than the combination of all acoustic features reported in Table 5. Table 6 shows the p-values found during the Wilcoxon Signed Rank Test when comparing the CNN-RLBP with significance level $\alpha = 0.05$.

Table 7 presents the confusion matrix of the best result reported in Table 5. By analyzing the confusion matrix presented in Table 7, it can be seen that the bigger occurrence of confusion is from Forró (3) with Gaúcha (4), two Brazilian music genres. These genres are very popular in Brazil and their rhythmic and timbral content have noticeable similarities, which might explain such confusions.

Table 6
Wilcoxon Signed Rank Test (p-values for $\alpha = 0.05$) comparing CNN-RLPB against all single classifiers reported in Table 3 and combinations reported in Table 5. (+) Indicates statistical difference and (−) no difference.

Strategy	Representation	p-Value ($\alpha = 0.05$)
Single classifiers	RH	0.000001 (+)
	RP	0.000001 (+)
	SSP	0.000001 (+)
	RLBP	0.033901 (+)
	CNN	0.004215 (+)
Combinations	RP, RH, SSH	0.000001 (+)
	RP, CNN, RLBP	0.350300 (−)
	SSD, CNN, RLBP	0.466900 (−)
	RH, CNN, RLBP	0.422200 (−)

Table 7
Confusion matrix (%) of the best result on LMD.

	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(0) Axé	**92.2**	0.0	1.1	0.0	4.4	0.0	1.1	0.0	1.1	0.0
(1) Bachata	0.0	**98.9**	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(2) Bolero	0.0	0.0	**96.7**	0.0	1.1	0.0	0.0	0.0	1.1	1.1
(3) Forró	1.1	0.0	4.4	**71.1**	13.3	0.0	3.3	2.2	4.4	0.0
(4) Gaúcha	4.4	0.0	2.2	0.0	**91.1**	0.0	0.0	0.0	2.2	0.0
(5) Merengue	1.1	0.0	0.00	0.0	0.0	**98.9**	0.0	0.0	0.0	0.0
(6) Pagode	1.1	0.0	3.3	0.0	2.2	0.0	**90.0**	2.2	1.1	0.0
(7) Salsa	1.1	0.0	5.6	0.0	0.0	0.0	0.0	**93.3**	0.0	0.0
(8) Sertaneja	2.2	0.0	4.4	0.0	3.3	0.0	0.0	0.0	**90.0**	0.0
(9) Tango	0.0	0.0	1.1	0.0	1.1	0.0	0.0	0.0	0.0	**97.8**

Table 8
Recognition accuracy (%) with standard deviation (σ) and *F*-measure (%) individually obtained using each kind of feature on the ISMIR 2004 database.

Feature		1st-level fusion	Accuracy (%)	*F*-measure (%)
Acoustic	RH	–	63.4	58.5
	RP	–	73.5	71.9
	SSD	–	76.8	76.4
Visual	CNN	Sum	**85.9**	**86.3**
	RLBP	Sum	83.4	82.5

Table 9
Recognition accuracy (%) with standard deviation (σ) and *F*-measure (%) obtained on the ISMIR 2004 dataset by combining different representations.

Representations	2nd-level fusion	Accuracy (%)	*F*−measure (%)
RP, CNN, RLBP	Sum	83.8	85.9
RH, CNN, RLBP	Sum	85.3	84.7
CNN, RLBP	Product	86.6	**87.1**
SSD, CNN, RLBP	Product	**86.7**	**86.6**

6.2. Results on ISMIR 2004

As described in Section 4, we have used an adaptable signal segmentation strategy. This strategy was particularly important to allow the usage of the whole ISMIR 2004 dataset in the experiments, since there are music pieces with less than 90 seconds duration in it. Table 8 shows the results individually obtained using each kind of feature. In this case the best result, 85.9% (accuracy), was achieved by the CNN, followed by the SVM trained with the RLBP, 83.4% (accuracy). Similarly to the experiments on the LMD, the acoustic features were outperformed by the visual features.

Differently from the experiments on the LMD, the combination of different representations brought a slight improvement compared to the best classifier. This behavior may be explained by the fact that all classifiers make quite similar mistakes. Table 9 presents the best results obtained by combining the outputs of the classifiers based on CNN and hand-crafted features.

Table 10
Confusion matrix (%) of the best result on the ISMIR 2004 dataset.

	(0)	(1)	(2)	(3)	(4)	(5)
(0) Classical	**99.4**	0.0	0.0	0.0	0.0	0.6
(1) Electronic	3.5	**88.6**	0.0	0.0	4.4	3.5
(2) Jazz/blues	15.4	3.8	**69.2**	0.0	3.8	7.7
(3) Metal/punk	0.0	0.0	0.0	**73.3**	24.4	2.2
(4) Rock/pop	7.8	3.9	0.0	2.0	**84.3**	2.0
(5) World	28.7	7.4	0.0	0.0	1.6	**62.3**

Table 11
State-of-the-art on the ISMIR 2004 dataset.

Authors	Features	Accuracy (%)
Sigtia and Dixon [45]	Representation Learning	74.4
Costa et al. [7]	Visual features (LBP)	80.6
Nanni et al. [33]	Visual features (LPB-HF, LBP, RICLBP, LPQ)	82.9
Seyerlehner et al. [42]	Acoustic features	88.3
Lim et al. [27]	Acoustic features	89.9
Nanni et al. [33]	Visual (LPB-HF, LBP, RICLBP) and Acoustic (DBF, OSC)	**90.2**

By analyzing the obtained results, it can be seen that the best results were obtained by the classifiers based on visual features (CNN and RLBP). The best results with or without the use of any classifier based on acoustic feature are practically the same. These results are confirmed by the Wilcoxon Signed Rank Test, which shows that both CNN and RLPB are statistically better than the classifiers trained with acoustic features, but there is no statistical difference between the CNN and RLBP (p-value = 0.2575 for $\alpha = 0.05$), neither between the best combination (SSD, CNN, RLBP) and CNN (p-value = 0.5530 for $\alpha = 0.05$) or RLPB (p-value = 0.2575 for $\alpha = 0.05$).

Table 10 presents the confusion matrix of the best result reported in Table 9. It shows that Jazz/Blues and World music are very often confused with Classical. Besides, there is a lot of room for improvement between Metal/punk and Rock/pop. These confusions may be explained in part by the distribution of the database, which is far from uniform (Table 1).

The best results on the literature for the ISMIR 2004 dataset are presented in Table 11. The highest performance achieved on this dataset is reported by Nanni et al. in [33], which uses a combination of weighted classifiers trained with textural features and an ensemble of SVMs combined with Random Space Adaboost trained with acoustic features.

Such a combination is able to reach an accuracy of 90.2%. The average performance of the individual classifiers, though, is 80%, i.e., significantly lower than the performance achieved by the CNN. This shows that the CNN is a viable alternative for music genre recognition and can be further explored to build robust ensembles of heterogeneous classifiers.

Here it is worth mentioning that Panagakis et al. [37] report an accuracy of 94.3% for this database, however, as pointed out by Sturm [50], these results arise from a flaw in the experiment inflating accuracies from around 60%.

6.3. Results on African music database

As mentioned before, the musical content in this collection is quite different from the typical western music databases, which makes it an interesting dataset to assess the power of representation learning. Table 12 describes the results obtained on the African ethnic music collection for the four different classification tasks: country, ethnic, function, and instrumentation.

Table 12
Recognition accuracy (%) with standard deviation (σ) followed by F-measure (%) individually obtained using each kind of feature on the African music database.

Features		Country	Ethnic	Function	Instrumentation
Acoustic	RH	64.1 ± 2.4/ 61.4	62.4 ± 3.7/ 54.2	26.0 ± 3.5/ 18.6	44.7 ± 4.9/ 39.1
	RP	78.1 ± 2.4/ 76.9	80.2 ± 6.1/ 79.1	39.1 ± 4.2/ 35.2	63.7 ± 4.6/ 60.1
	SSD	84.1 ± 3.4/ 83.5	87.1 ± 3.4/ 87.0	49.1 ± 2.3/ 46.3	70.6 ± 4.7/ 68.4
Visual	CNN	**92.9 ± 0.5**[a]/ **93.2**[a]	**93.9 ± 0.7**[a]/ **93.4**[a]	**61.6 ± 2.2**[a]/ **63.0**[a]	**75.9 ± 1.3**[a]/ **78.0**[a]
	RLBP	88.0 ± 1.7[b]/ 86.0[b]	88.2 ± 4.8[c]/ 87.9[c]	49.1 ± 3.8[b]/ 44.9[b]	72.5 ± 5.5[a]/ 68.8[a]

[a] Sum fusion rule in 1st level.
[b] Max fusion rule in 1st level.
[c] Min fusion rule in 1st level.

Table 13
Wilcoxon Signed Rank Test (p-values for $\alpha = 0.05$) comparing the CNN against all other classifiers for the African music database. (+) indicates statistical difference.

Task	Classifiers			
	RH	RP	SSD	RLBP
Country	0.000090 (+)	0.000090 (+)	0.000122 (+)	0.000364 (+)
Ethnic	0.000090 (+)	0.000090 (+)	0.000289 (+)	0.010519 (+)
Function	0.000091 (+)	0.000091 (+)	0.001388 (+)	0.001793 (+)
Instrum.	0.000090 (+)	0.000090 (+)	0.000289 (+)	0.022537 (+)

Table 14
Recognition accuracy (%) with standard deviation (σ) followed by F-measure (%) obtained on the African database by combining different representations. The product rule was used as 2nd level fusion rule in all cases.

Features	Country	Ethnic	Function	Instrumentation
RH, CNN, RLBP	90.4 ± 0.7/ 91.9	90.1 ± 2.0/ 91.2	58.3 ± 1.0/ 59.5	72.5 ± 1.5/ 76.2
RP, CNN, RLBP	91.6 ± 0.6/ **92.7**	92.7 ± 0.8/ **93.3**	57.3 ± 2.0/ 60.2	72.9 ± 1.7/ 76.6
SSD, CNN, RLBP	91.4 ± 0.8 92.3	93.1 ± 0.8 **93.4**	58.2 ± 1.7 60.0	73.1 ± 2.3 77.4
CNN, RLBP	93.0 ± 0.8/ **93.5**	93.2 ± 1.2/ **93.3**	60.9 ± 1.2/ **62.2**	76.5 ± 1.5/ **78.5**

The adaptable signal segmentation strategy used here (described in Section 4) was important to avoid discarding several music pieces (that have less than 60 s) from this database.

Differently from the two previous experiments where the classifiers trained with acoustic features reached quite poor performances, in this dataset they achieve an accuracy comparable to the classifier trained with RLBP. This corroborates with the hypothesis pointed out by Lidy et al. [25] that ethnic music has some peculiarities. Table 12 also shows that the CNN outperforms all other classifiers in all classification tasks of the African music database, what lead us to conclude that representation learning is a promising strategy to deal with such peculiarities. Such good performance is corroborated by the Wilcoxon Signed Rank Test reported in Table 13, which indicates statistical difference among the CNN and all other classifiers for all the different classification tasks of the African music database.

On the other hand, the big gap among the performance of the classifiers did not allow the 2nd-level fusion to further improve the results. On the contrary, in some tasks, such as Ethnic and Function, the 2nd-level fusion decreased the accuracy. The results of the 2nd-level fusion are presented in Table 14, where all the results were obtained with the product rule.

Table 15 presents the results on the African music collection already published in the literature. Note that the results reported

Table 15
State-of-the-art on the African music collection.

Features			Accuracy (%)
Country			
Lidy et al. [25]	Acoustic	Hybrid-SSD	82.2
Lidy et al. [26]	Acoustic	RP, RH, SSD, MVD, TRH, TSSD	**89.0**
Ethnic group			
Lidy et al. [25]	Acoustic	Hybrid-TSSD	**88.6**
Lidy et al. [26]	Acoustic	RP, RH, SSD, MVD, TRH, TSSD	83.0
Function			
Lidy et al. [25]	Acoustic	Hybrid-SSD	48.3
Lidy et al. [26]	Acoustic	RP, RH, SSD, MVD, TRH, TSSD	**54.8**
Instrumentation			
Lidy et al. [25]	Acoustic	TSSD	69.1
Lidy et al. [26]	Acoustic	RP, RH, SSD, MVD, TRH, TSSD	**73.0**

in this work are, in all the cases, better than those reported in the literature.

7. Conclusion

In this work we have evaluated the Convolutional Neural Network for music content characterization using three music databases with distinct characteristics. Our experiments have shown that the CNN compares favorably to other classifiers in several scenarios. On the African music database the CNN surpassed all the results reported in the literature while in the LMD database, the combination of a CNN with an SVM trained with RLBP achieved 92% of recognition rate, which is to the best of our knowledge, the best result (using the artist filter) on this dataset so far. In the ISMIR 2004 database, the CNN did not improve the state-of-the-art, but when compared to all individual classifiers, it performed better than all of them.

The good performance of the CNN, compared to the classifiers trained with hand-crafted features, can be explained by the robust representation that it is capable to learn from the data. Even in the cases where the CNN is outperformed by the hand-crafted features, such as RLBP, it brings some complementarity, which can be attested by the results produced through the combination of representation learning and hand-crafted features.

All those results support the hypothesis that representation learning using deep learning is a viable alternative for automatic feature engineering for music content characterization, which has been dominated over the last years by hand-crafted feature design. We believe that further investigation of different architectures of CNN or other approaches can bring some improvements to the field of music genre recognition and music information retrieval in general. As future work, we will explore techniques of data augmentation for spectrogram images, since it is well known that deep learning techniques such as CNN are highly dependent on large training sets. We also aim to make investigations on other datasets, related to other music classification tasks, like artist recognition and mood classification.

Acknowledgements

The authors would like to thank the Royal Museum of Central-Africa (RMCA), Belgium, for providing the data set of African music. This research has been partially supported by The National Council for Scientific and Technological Development (CNPq) grant 303513/2014-4.

References

[1] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[2] P. Cano, E. Gomez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, N. Wack, ISMIR 2004 audio description contest. Technical report, Music Technology Group – Universitat Pompeu Fabra, 2006.
[3] C.H.L. Costa, J.D. Valle-Jr, A.L. Koerich, Automatic classification of audio data, in: International Conference on Systems, Man and Cybernetics, 2004, pp. 562–567.
[4] Y.M.G. Costa, L.S. Oliveira, A.L. Koerich, F. Gouyon, Music genre recognition using spectrograms, in: International Conference on Systems, Signals and Image Processing, 2011, pp. 154–161.
[5] Y.M.G. Costa, L.S. Oliveira, A.L. Koerich, F. Gouyon, Music genre recognition based on visual features with dynamic ensemble of classifiers selection, in: International Conference on Systems, Signals and Image Processing, 2013, pp. 55–58.
[6] Y.M.G. Costa, L.S. Oliveira, A.L. Koerich, F. Gouyon, Music genre recognition using Gabor filters and LPQ texture descriptors, in: Iberoamerican Congress on Pattern Recognition, 2013, pp. 67–74.
[7] Y.M.G. Costa, L.S. Oliveira, A.L. Koerich, F. Gouyon, J. Martins, Music genre classification using LBP texture features, Signal Process. 92 (11) (2012) 2723–2737.
[8] T. Feng, Deep learning for music genre classification. Technical report, University of Illinois, 2014.
[9] A. Flexer, A closer look on artists filters for musical genre classification, in: International Conference on Music Information Retrieval, 2007, pp. 341–344.
[10] J. George, L. Shamir, Computer analysis of similarities between albums in popular music, Pattern Recognit. Lett. 45 (2014) 78–84.
[11] G. Gwardys, D. Grzywczak, Deep image features in music information retrieval, Int. J. Electron. Telecommun. 60 (4) (2014) 321–326.
[12] L.G. Hafemann, L.S. Oliveira, P. Cavalin, Forest species recognition using deep convolutional neural networks, in: International Conference on Pattern Recognition, 2014, pp. 1103–1107.
[13] P. Hamel, Pooled features classification., Submission to Audio Train/Test Task of MIREX, 2011.
[14] P. Hamel, D. Eck, Learning features from music audio with deep belief networks, in: ISMIR, Utrecht, The Netherlands, 2010, pp. 339–344.
[15] P. Hamel, S. Lemieux, Y. Bengio, D. Eck, Temporal pooling and multiscale learning for automatic annotation and ranking of music audio, in: ISMIR, 2011, pp. 729–734.
[16] E. Humphrey, J.P. Bello, Y. LeCun, Moving beyond feature design: deep architectures and automatic feature learning in music informatics, in: International Conference on Music Information Retrieval, 2012, pp. 403–408.
[17] E.J. Humphrey, J.P. Bello, Rethinking automatic chord recognition with convolutional neural networks, in: International Conference on Machine Learning and Applications, 2012, pp. 357–362.
[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, 2014 arXiv:1408.5093.
[19] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 226–239.
[20] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[21] Y. LeCun, Y. Bengio, G.E. Hinton, Deep learning, Nature 521 (2015) 436–444.
[22] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551.
[23] J.H. Lee, K. Choi, X. Hu, J.S. Downie, K-pop genres: a cross-cultural exploration, in: ISMIR, 2013, pp. 529–534.
[24] T. Li, A.B. Chan, A. Chun, Automatic musical pattern feature extraction using convolutional neural network, in: International Conference Data Mining and Applications, 2010.
[25] T. Lidy, C.N. Silla Jr., O. Cornelis, F. Gouyon, A. Rauber, C.A.A. Kaestner, 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 Process. 90 (4) (2010) 1032–1048.
[26] T. Lidy, R. Mayer, A. Rauber, P.J. Ponce de León Amador, A. Pertusa Ibá nez, J.M. I nesta Quereda, A Cartesian ensemble of feature subspace classifiers for music categorization, in: International Society for Music Information Retrieval Conference, 2010, pp. 279–284.
[27] S.C. Lim, J.S. Lee, S.J. Jang, S.P. Lee, M.Y. Kim, Music-genre classification system based on spectro-temporal features and feature selection, IEEE Trans. Consum. Electron. 58 (4) (2012) 1262–1268.
[28] M. Lopes, F. Gouyon, A. Koerich, L.S. Oliveira, Selection of training instances for music genre classification, in: International Conference on Pattern Recognition, 2010, pp. 4569–4572.
[29] J.G. Martins, L.S. Oliveira, S. Nisgoski, R. Sabourin, A database for automatic classification of forest species, Mach. Vis. Appl. 24 (3) (2013) 567–578.
[30] MIREX, Music Information Retrieval Evaluation eXchange, 2010, http://www.music-ir.orghttp://www.music-ir.org.
[31] T. Nakashika, C. Garcia, T. Takiguchi, Local-feature-map integration using convolutional neural networks for music genre classification, in: Interspeech, 2012, pp. 1752–1755.

[32] L. Nanni, Y.M.G. Costa, S. Brahnam, Set of texture descriptors for music genre classification, in: International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, 2014.
[33] L. Nanni, Y.M.G. Costa, A. Lumini, M.Y. Kim, S.R. Baek, Combining visual and acoustic features for music genre classification, Expert Syst. Appl. 45 (2016) 108–117.
[34] X.X. Niu, C.Y. Suen, A novel hybrid CNN-SVM classifier for recognizing handwritten digits, Pattern Recognit. 45 (2012) 1318–1325.
[35] T. Ojala, M. Pietikä inen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[36] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[37] Y. Panagakis, C. Kotropoulos, T.G. Arce, Music genre classification using locality preserving non-negative tensor factorization and sparse representations, in: ISMIR, 2009, pp. 249–254.
[38] A. Rauber, M. Fruhwirth, Automatically analyzing and organizing music archives, in: Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries, 2001, pp. 402–414.
[39] A. Rauber, E. Pampalk, D. Merkl, The SOM-enhanced jukebox: organization and visualization of music collections based on perceptual models, J. New Music Res. 32 (2) (2003) 193–210.
[40] A. Schindler, R. Mayer, A. Rauber, Facilitating comprehensive benchmarking experiments on the million song dataset, in: International Conference on Music Information Retrieval, 2012, pp. 469–474.
[41] J. Schluter, S. Bock, Musical onset detection with convolutional neural networks, in: International Workshop on Machine Learning and Music, 2013, pp. 1–4.
[42] K. Seyerlehner, M. Schedl, T. Pohle, P. Knees, Using block-level features for genre classification, tag classification and music similarity estimation, Submission to Audio Music Similarity and Retrieval Task of MIREX, 2010.
[43] J. Shen, T. Mei, D. Tao, X. Li, Y. Rui, EMIF: towards a scalable and effective indexing framework for large scale music retrieval, in: Proceedings of the ACM on International Conference on Multimedia Retrieval, ACM, 2015, pp. 543–546.
[44] J. Shen, J. Shepherd, A.H.H. Ngu, Towards effective content-based music retrieval with multiple acoustic feature combination, IEEE Trans. Multimed. 8 (6) (2006) 1179–1189.
[45] S. Sigtia, S. Dixon, Improved music feature learning with deep neural networks, in: IEEE International Conference on Acoustic, Speech and Signal Processing, 2014, pp. 6959–6963.
[46] C.N. Silla, A.L. Koerich, C.A.A. Kaestner, The Latin music database, in: International Conference on Music Information Retrieval, 2008, pp. 451–456.
[47] C.N. Silla-Jr, C.A.A. Kaestner, A.L. Koerich, Automatic music genre classification using ensemble of classifiers, in: IEEE International Conference on Systems, Man and Cybernetics, 2007, pp. 1687–1692.
[48] Y. Song, C. Zhang, Content-based information fusion for semi-supervised music genre classification, IEEE Trans. Multimed. 10 (1) (2008) 145–152.
[49] S. Stober, A. Nurnberger, Adaptive music retrieval – a state of the art, Multimed. Tools Appl. 65 (3) (2013) 467–494.
[50] B. Sturm, Classification accuracy is not enough: on the evaluation of music genre recognition systems, J. Intell. Inf. Syst. 41 (3) (2013) 371–406.
[51] G. Tzanetakis, P. Cook, Musical genre classification of audio signals, IEEE Trans. Speech Audio Process. 10 (5) (2002) 293–302.
[52] X. Valero, F. Alias, Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification, IEEE Trans. Multimed. 14 (6) (2012) 1684–1689.
[53] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag New York, Inc, 1995.
[54] D. Wang, S. Deng, X. Zhang, G. Xu, Learning music embedding with metadata for context aware recommendation, in: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ACM, 2016, pp. 249–253.
[55] N.J. Wu, J.S.R. Jang, Combining acoustic and multilevel visual features for music genre recognition, ACM Trans. Multimed. Comput. Commun. Appl. 12 (2015) 1–17.
[56] Y. Zhao, W. Jia, R.X. Hu, H. Min, Completed robust local binary pattern for texture classification, Neurocomputing 106 (2013) 68–76.
[57] E. Zwicker, H. Fastl, Psychoacoustics – Facts and Models, Springer, 1999.