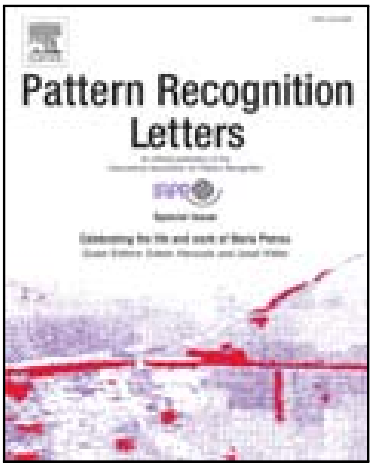


# Combining visual and acoustic features for audio classification tasks

L. Nanni, Y.M.G. Costa, D.R. Lucio, C.N. SillaJr., S. Brahnam

PII: S0167-8655(17)30013-2
DOI: 10.1016/j.patrec.2017.01.013
Reference: PATREC 6728

To appear in: *Pattern Recognition Letters*

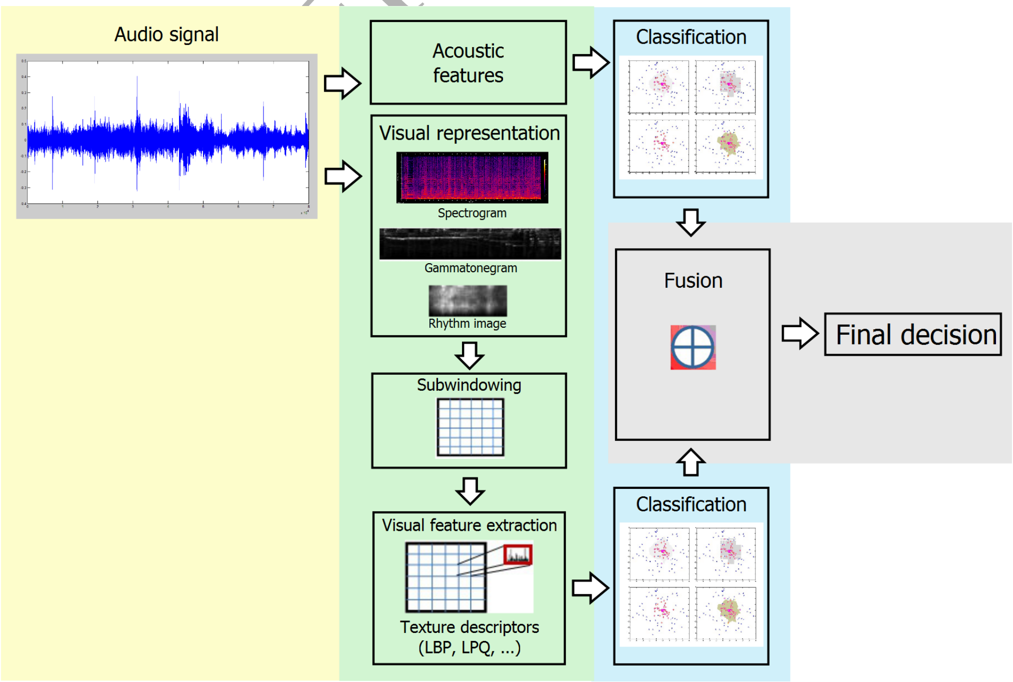
Received date: 7 May 2016
Revised date: 30 December 2016
Accepted date: 16 January 2017

Please cite this article as: L. Nanni, Y.M.G. Costa, D.R. Lucio, C.N. SillaJr., S. Brahnam, Combining visual and acoustic features for audio classification tasks, *Pattern Recognition Letters* (2017), doi: 10.1016/j.patrec.2017.01.013

**Highlights**

- Coupling texture descriptors and acoustic features
- Different methods for representing an audio as an image are compared
- Heterogeneous ensemble of different classifiers improves performance

## Graphical Abstract



Pattern Recognition Letters
journal homepage: www.elsevier.com

# Combining visual and acoustic features for audio classification tasks

L. Nanni[1], Y. M. G. Costa[2], D. R. Lucio[2], C. N. Silla Jr.[3], S. Brahnam[4]
[1] DEI, University of Padua
[2] PCC/DIN, State University of Maringá
[3] PPGIa, Pontifical Catholic University of Paraná
[4] CIS, Missouri State University

## ABSTRACT

In this paper a novel and effective approach for automated audio classification is presented that is based on the fusion of different sets of features, both visual and acoustic. A number of different acoustic and visual features of sounds are evaluated and compared. These features are then fused in an ensemble that produces better classification accuracy than other state-of-the-art approaches. The visual features of sounds are built starting from the audio file and are taken from images constructed from different spectrograms, a gammatonegram, and a rhythm image. These images are divided into subwindows from which a set of texture descriptors are extracted. For each feature descriptor a different Support Vector Machine (SVM) is trained. The SVMs outputs are summed for a final decision. The proposed ensemble is evaluated on three well-known databases of music genre classification (the Latin Music Database, the ISMIR 2004 database, and the GTZAN genre collection), a dataset of Bird vocalization aiming specie recognition, and a dataset of right whale calls aiming whale detection. The MATLAB code for the ensemble of classifiers and for the extraction of the features will be publicly available (https://www.dei.unipd.it/node/2357 +Pattern Recognition and Ensemble Classifiers).

*Keywords*: audio classification, texture, image processing, acoustic features, ensemble of classifiers, pattern recognition

## 1. Introduction

In recent years, an increasing number of works has been noticed exploring audio classification tasks using features taken from the visual domain. In 2008, Yu and Slotine (42) described musical instruments classification from time-frequency texture. Inspired in that work, Costa et al. (5) started to investigate some well-known texture descriptors, taken from the image processing literature, aiming to capture the spectrogram content to perform music genre classification. For this purpose, the authors experimented with the Gray Level Co-occurrence Matrix (13), and found some promising results. In a follow up work, Costa et al. used Local Binary Patterns (LBP) to describe the spectrogram content (8). Costa et al. also introduced a novel way to split the spectrogram image into zones which corresponds to frequency bands. In addition, Ming-Ju et al. started to investigate the use of both acoustic and visual features in the music genre classification task, as made in (41). The use of several features taken from both acoustic and visual domains is still under investigation, as one can see in (26).

More recently, the visual-spectrogram technique has been investigated to perform audio classification in other application domains, such as language identification (21) and bird species classification (20; 25).

In this paper we report the results of a new ensemble system that expands our previous work on texture descriptors extracted from spectrograms of audio signals (24; 26; 25).

The main contribution of this work is the design and evaluation of an ensemble of visual and acoustic descriptors that work well across the following benchmark datasets: the Latin Music Database (LMD) (14), the ISMIR 2004 (2) database, the GTZAN genre collection (39) database, a subset of bird vocalizations for bird species classification extracted from the Xeno-Canto database, and a dataset composed of clips which contain any mixture of right whale calls, non-biological noise, or other whale calls, aiming right whale calls detection[1].

In comparison with our previous work, this work has the following novelty aspects: we employ Gammatonegrams and Rhythm Images as visual representations of sound. From these representations several texture descriptors are extracted; we also employ three different types of spectrograms, however in this work we employ both gray level image as well as color images. Our previous work was limited to gray-level images; we evaluate these novel representations, several texture descriptors extracted from them and their combination in the same datasets that we have used in previous research (LMD, GTZAN, ISMIR, BIRD-46) and a novel dataset of right whale calls.

## 2. Proposed approach

The input audio signal is first represented by the following types of image (step V1):

- Different spectrograms: three spectrogram images are created with the lower limits of the amplitudes set to -70 dBFS, -90 dBFS, and -120 dBFS, respectively. We consider the spectrogram both as gray level image and as a color image;
- Gammatonegram: this is a spectrogram that is based on a gammatone filter band that approximates how sounds are processed by the mammalian chochlea;
- Rhythm image: this image describes the fluctuations or modulation amplitudes on a number of frequency bands in the human auditory range.

Next, each image is divided into a set of subwindows (step V2) from which a set of local texture descriptors are extracted (step V3). Each descriptor is classified by SVM. The final decision is obtained by fusing all the SVM scores using the weighed sum rule (step A3/V5). The rest of this section describes steps V1-V5 and A1-A3 in greater detail. The classification scheme is depicted in Figure 1.

### *2.1. Step V1: Spectrogram representation*

### *2.1.1. Spectrograms*

The signal segmentation strategy utilized in this work is the one suggested by Costa et al. in (4). By following this strategy, the original signal content can be reduced to three 10s segments. To minimize the possibility of an unrepresentative sample, each of these segments are taken from the beginning, middle, and end of the original audio signals for the three music genre datasets.

A different scenario is used for bird vocalization classification. Segmentation is important for extracting the most relevant parts of the signal (called shots). The samples are built from shots manually extracted from the original bird samples. In order to standardize the size of the useful content, we concatenate the sample with itself until the size is equal to 30s. In right whale call detection, segmentation was not used. In that dataset the clips are all standardized with two seconds length.

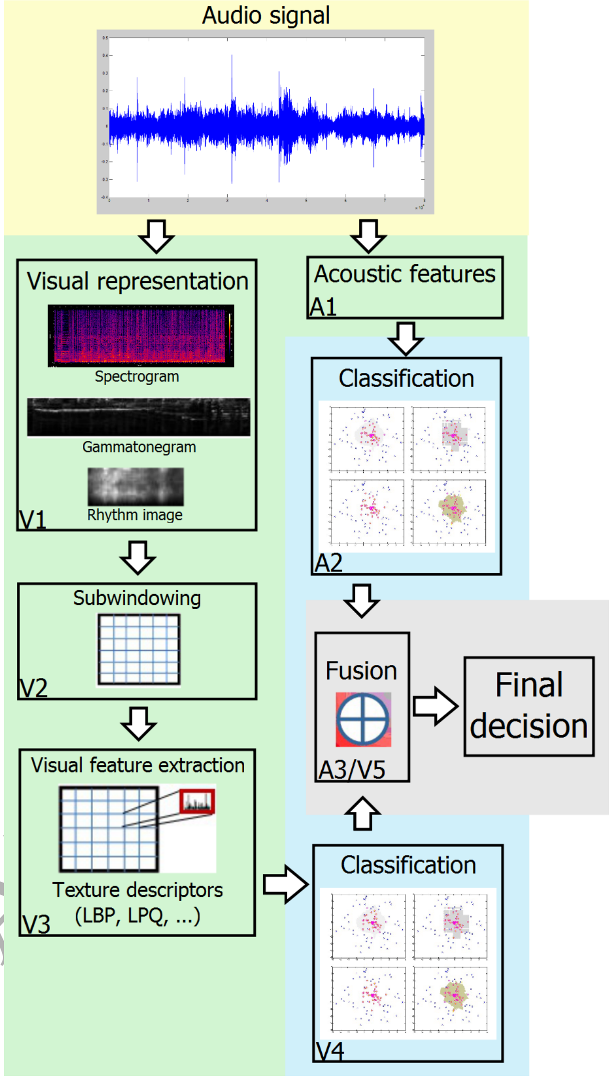


**Figure 1. Classification overall scheme.**

Once segmented, the audio signal is converted into a spectrogram image that shows the spectrum of frequencies (vertical axis) varying according to time (horizontal axis). The intensity of each point in the image represents the signal's amplitude. The spectrograms are generated using the Hanning window function with the DFT computed with a window size of 1024 samples.

After a battery of tests, we decided to use spectrograms created with lower limits set to the following three values: -70 dBFS, -90 dBFS, and -120 dBFS. The rationale behind this decision was the expectation of finding some complementarity between these different representations.

### *2.1.2. Gammatone-like spectrograms*

Spectrograms, however, do not match the way sound is processed by the ear.

In this work we use a spectrogram that is based on a Gammatone filter band to approximate how sounds are processed by the mammalian cochlea since this filter band has been shown to provide a simple linear fit to experimental observations of the cochlea (33).

Formally, a Gammatone filter describes an impulse response that is the product of a gamma envelope, as described in equation 1:

$$g(t) = at^{n-1}e^{-2\pi bt}cos(2\pi ft + \phi), \qquad (1)$$

where $t$ measured in seconds is time, $f$ in Hz is the center

[1] Organized by The Marinexplore and Cornell University and available at https://www.kaggle.com/c/whale-detection-challenge/ .

frequency, $\phi$ in radians is the phase of the carrier, $a$ is the amplitude, $n$ is the filter's order, and $b$ in Hz is the filter's bandwidth.

To convert a Gammatone filter bank into a time-frequency visualization requires summing up the energy within regular time bins. A spectrogram that is similar to a Gammatone spectrogram is proposed in (9), a MATLAB toolbox that constructs a Gammatone-like spectrogram by calculating a conventional fixed-bandwidth spectrogram that is then combined with the fine frequency resolution of the FFT-based spectra into the Gammatone responses using a weighting function. This method has been shown in (11) to approximate the full approach 30 to 40 times faster with little loss of information.

### *2.1.3. Rhythm image*

In this section we discuss a method proposed in (17) for extracting features based on rhythm patterns, which describe the fluctuations or modulation amplitudes of a number of frequency bands, specifically the critical bands, or overlapping of frequencies, in the human auditory range.

There are two main steps in extracting Rhythm Patterns Features (17), but before applying the two steps, the audio is preprocessed by converting it into raw digital audio and averaging all channels into one channel. After preprocessing the signal, 6s excerpts are extracted that are transformed into a power spectrum using STFT.

The following two steps are then applied:

- Step V1: the aim of this step is to compute the human loudness sensation as a spectrogram, or sonogram, of different frequency bands. First, the resulting frequency bands produced by the preprocessing application of STFT are grouped into 24 psychoacoustically critical-bands by applying Bark scale (44). A spreading function is then applied to account for spectral masking affects (36). Next the spectrum energy values on the critical bands are transformed into the decibel scale (dB) followed by the successive calculation of the loudness levels through incorporating equal-loudness contours (Phon) and the specific loudness sensation of each critical band (Sone).
- Step V2: FFT is applied to the sonogram, thereby transforming it into a time-invariant representation based on the modulation frequency of the 24 critical bands. In this way the reoccurring patterns in the audio signal are captured. Modulation amplitudes are then weighted according to their effects on human hearing. A gradient filter is applied to emphasize distinctive beats, and Gaussian smoothing is applied to diminish variations unnoticed by human hearing.

The Rhythm Pattern (RP) is reshaped to a matrix of size 24 (i.e. the number of critical bands) × 60, and from this matrix the visual descriptors are extracted. The RP without reshaping is also used to train an SVM.

## *2.2. Step V2: Subwindowing*

Costa et al. have recommended in several works (5; 6; 8) employing a zoning mechanism to preserve local information about the extracted features. With this in mind, 15 zones for each spectrogram that vary in sizes defined by the Mel scale (40) were selected for this study. In (26) we showed a subwindowing based on the Mel scale outperforms a method where the features are extracted from the whole image. Considering that one spectrogram is created for each segment taken from the original signal, we have a total of 45 zones, and 45 SVMs are trained for each descriptor, i.e. one for each of the 45 zones. The SVM decisions are combined by sum rule.

The Gammatonegram is divided into 30 × 3 non overlapping subwindows of size 21 × 99, while the Rhythm image is divided into three non overlapping subwindows of size 24 × 20. A different SVM is trained for each subwindow and combined by sum rule.

## *2.3. Step V3: Texture descriptors*

The following state-of-the-art texture descriptors[2]are evaluated in this work:

- LBP (28): this is a multiscale uniform local binary pattern (LBP). The final descriptor is obtained from the concatenation of patterns at different radii $R$ and different sampling points $P$, viz.($R$=1, $P$=8) and ($R$=2, $P$=16);
- LBP-HF (43): this is a multiscale LBP histogram Fourier descriptor obtained from the concatenation of LBP-HF with values ($R$=1, $P$=8) and ($R$=2, $P$=16);
- RICLBP (27): this is a multiscale rotation invariant co-occurrence of adjacent LBP with values($R$=1, $P$=8), ($R$=2, $P$=8) e ($R$=4, $P$=8);
- LPQ (29): this is a multiscale Local Phase Quantization with radius values 3 and 5;
- MLPQ (23): this is an ensemble of LPQ that is based on a ternary encoding. It is possible to design an effective ensemble by combining sets of LPQ extracted by varying the parameters $r$ (the neighborhood sizes, $r \in$[1, 3, 5]), a (the scalar frequency, $a \in$[0.8, 1, 1.2, 1.4, 1.6]), and $\rho$ (the correlation coefficient between adjacent pixel values and with $\rho \in$[0.75, 0.95, 1.15, 1.35, 1.55, 1.75, 1.95]). Each LPQ trains a different classifier. The classifiers are combined by sum rule. This ensemble is built up with 105 descriptors whose scores are summed and normalized by dividing the sum by 105.
- HASC (1):the Heterogeneous Auto-Similarities of Characteristics is applied to heterogeneous dense features maps.
- ELHF (22): this is an ensemble of variants of the Local Binary Pattern Histogram Fourier that is built from the following 7 descriptors (each trained by an SVM and with SVM scores summed and normalized by dividing the sum by 7):

[2] The MATLAB code we used is available so that misunderstandings in the parameter settings used for each method can be avoided.

  - FF: the original method, where from each discrete Fourier transform (DFT) the first half of the coefficients are retained;
  - DC: an approach where from each discrete Cosine transform (DCT) the first half of the coefficients are retained;
  - An approach where the histogram is decomposed by Daubechies wavelet before DFT and then FF is performed;
  - An approach where the histogram is decomposed by Daubechies wavelet before DCT and then the method DC is performed;
  - An approach where the histogram is decomposed by Daubechies wavelet before DFT and then the method FF is performed, with all coefficients retained;
  - An approach where the histogram is decomposed by Daubechies wavelet before DCT and then the method DC is performed, with all coefficients retained.
  - An approach where all the bins of the histogram are retained.

- GABOR: for the Gabor filter (GF) features extraction, several different values for scale level and orientation were experimentally evaluated with the best result obtained using 5 different scale levels and 14 different orientations. The mean-squared energy and the mean amplitude were calculated from each possible combination between scale and orientation. In this way, a feature vector of size $5\times14\times2$ is obtained.

### *2.4. Steps A2-A3/V4-V5: Classification and fusion*

We use a one versus all SVM in this work with a radial basis function (RBF) kernel for classification. To avoid the risk of overfitting due to the small training sets, we do not perform parameter optimization but rather set $C$=1000 and $\gamma$=0.1 for all experiments. Before the classification step, the features are normalized to [0, 1].

## 3. Acoustic features (step A1)

The following five acoustic descriptors were selected for evaluation:

1. Statistical Spectrum Descriptor (SSD) (36; 16): this descriptor is a set of statistical measures describing the audio content that are taken from the Sone, i.e. the moments on the Sonogram values of each of the 24 critical bands.
2. Rhythm Histogram (RH) (36; 16): the magnitudes of each modulation frequency bin of all the critical bands are summed up to form a histogram of “rhythmic energy” per modulation frequency. The histogram has 60 bins that reflect modulation frequency between 0 and 10 Hz. The feature set is the median of the histograms of each 6s segment extracted.
3. Modulation Frequency Variance Descriptor (MVD) (36; 16): this is a descriptor that measures variations over the critical frequency bands for a specific modulation frequency. MVD is obtained by computing the statistical measures for each modulation frequency over the 24 bands. The MVD descriptor for the audio file is the mean of the MVDs taken from the 6s segments and is a 420-dimensional vector.
4. Temporal Statistical Spectrum Descriptor (TSSD) (36; 16): this descriptor incorporates temporal information from the SSD, such as timbre variations and changes in rhythmic. Statistical measures are taken across the SSD measures extracted from segments at different time positions in an audio file.
5. Temporal Rhythm Histograms (TRH) (36; 16): this descriptor captures changes over time of rhythmic aspects of music.

For the music genre dataset, we make use of a commercial system that employs acoustic features for classifying audio files. This system is based on a method proposed in (18) that was improved in (26) (the version used here).

## 4. Experimental results

The proposed approach is evaluated on the following four databases using the recognition rate as the performance indicator:

- LMD(14): the Latin Music Database was originally designed to evaluate music information retrieval systems. This dataset contains 3,227 samples classified into ten musical genres: axe, bachata, bolero, forro, gaucha, merengue, pagode, salsa, sertaneja, and tango. The testing protocol for this dataset is the threefold cross-validation protocol, where artist filter restriction is applied (10) (i.e. where all the samples by a specific artist are included in only one fold). Since the distribution of samples per artist is far from uniform, only a subset of 900 samples is used for fold creation.

- ISMIR 2004 (2): it contains 1,458 (729 in test set and 729 in training set) samples assigned to six different genres: classical, electronic, jazz/blues, metal/punk, rock/pop, and world.

- GTZAN (39): this dataset was collected by G. Tzanetakis and represents ten genre classes (Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Popular, Reggae, and Rock). Each genre class contains 100 audio recordings. For a fair comparison with results reported in (18), we evaluate the performance using the same 10-fold split tested and shared by (18).

- BIRD songs dataset: we have used the publicly available dataset from (19)[3] . This dataset was built with bird sound samples taken from Xeno-Canto[4]. The developers of the dataset used the search tool available in Xeno-Canto to select species in a radius of up to 250 kilometers from the city of Curitiba, in the South of Brazil. All bird species with less than 10 samples were removed. We employ a stratified

[3] http://www.dainf.ct.utfpr.edu.br/ kaestner/brbird/SACBase/
[4] www.xeno-canto.org

10-fold cross validation strategy in the classification experiments. After these filters, 2,814 audio samples representing 46 bird species remained in the dataset. It will be made available at the author's homepages[5].

- Right whale calls dataset: we have used the dataset used in "The Marinexplore and Cornell University Whale Detection Challenge"[6]. The dataset is composed of 84,503 2-second audio clips which contain any mixture of right whale calls, non-biological noise, or other whale calls. Thirty thousand of the samples are available with class label. In this work, we used the first 20,000 labelled samples as training set and the last 10,000 samples were used to test. The results on this dataset are described using area under the ROC curve (AUC), we have used it as performance indicator since it is used in the challenge.

The aim of the experiments reported in Tables 1 to 4 was to compare the performance obtained by using the different texture descriptors as well as ensembles composed of different combinations of descriptors.

The following ensembles are evaluated:

- F2: LPQ+ELHF+LBP+RICLBP+HASC+GF
- F3: MLPQ+ELHF+LBP+RICLBP+HASC+GF
- FUSac: 4×SSD+RP+MVD+TSSD+TRH
- Full_light_W: F2+0.25×FUSac+OLDac1+OLDac2
- FULL_W: F3+0.25×FUSac+OLDac1+OLDac2

The labels OLDac1/OLDac2, refer to the acoustic features used in (26). OLDac1 is based on a SVM classifier, while OLDac2 is based on a random subspace of Adaboost. The labels $a \times X + b \times Y$ indicate the combination of approach $X$ and $Y$ by weighted sum rule, where the weight of $X$ is $a$ and the weight of $Y$ is $b$. Before combining the scores obtained by the ensemble of acoustic features (i.e. FUSac) and the scores obtained by the other ensemble (e.g. F2), the scores of each method are normalized to mean zero and standard deviation 1.

The columns in the following Tables are labeled as follows:

- $ME$ reports the performance obtained by extracting features from the spectrogram image created with lower limit set to -70 dBFS;
- $Sp$ reports the performance obtained by extracting features from the three spectrogram images created with lower limit set to -70 dBFS, -90 dBFS, and -120 dBFS. The features are extracted separately from each spectrum, and the scores obtained by the SVMs are summed;
- $Co$ reports the performance obtained by extracting features from the color spectrogram image. Since each color image is created with three bands, we used -70 dBFS coupled with the $R$ band, -90 dBFS coupled with the $G$ band, and -120 dBFS coupled with the $B$ band. Features are extracted separately from each image, and the scores obtained by the SVMs are summed;

**Table 1. Recognition accuracy (%) on the LMD dataset.**

| **Texture Descriptors** | $Me$ | $Sp$ | $Co$ | $3 \times Sp+$ $Co$ | $Ga$ | $3 \times Sp+$ $Co + Ga$ | $Ry$ | $3 \times Sp+$ $Co + Ga+$ $ry$ |
|---|---|---|---|---|---|---|---|---|
| LBP-HF | 82.8 | 82.2 | 81.4 | 82.8 | 76.0 | **83.3** | 41.4 | **83.3** |
| LPQ | 83.3 | 83.4 | 81.4 | **83.9** | 74.3 | 83.4 | 49.6 | 83.4 |
| ELHF | 84.7 | 85.1 | 81.9 | 85.1 | 76.1 | 85.0 | 48.9 | **85.3** |
| LBP | 84.9 | 85.1 | 82.2 | **85.8** | 76.4 | **85.8** | 47.7 | **85.8** |
| RICLBP | 84.3 | 84.9 | 82.5 | 84.8 | 76.2 | **85.6** | 43.5 | **85.6** |
| MLPQ | 83.7 | 84.4 | 82.0 | 84.4 | 81.2 | 84.7 | 63.6 | **84.8** |
| HASC | 83.9 | 84.0 | 83.8 | 84.3 | 80.3 | 84.9 | 53.2 | **85.0** |
| GF | 82.7 | 84.7 | 84.7 | 84.8 | 78.1 | 85.4 | 56.7 | **85.6** |
| **Non Texture Descriptor** | | | | | | | | |
| FUSac | 71.9 | | | | | | | |
| **Ensembles** | | | | | | | | |
| F2 | - | **86.2** | - | 86.0 | - | 85.8 | - | 86.1 |
| F3 | - | **86.3** | - | 86.2 | - | 86.0 | - | **86.3** |
| Full_light_W | - | 84.4 | - | 84.4 | - | 84.4 | - | **84.7** |
| Full_W | - | **84.6** | - | 84.3 | - | **84.6** | - | **84.6** |

**Table 2. Recognition accuracy (%) on the ISMIR 2004 dataset.**

| **Texture Descriptors** | $Me$ | $Sp$ | $Co$ | $3 \times Sp+$ $Co$ | $Ga$ | $3 \times Sp+$ $Co + Ga$ | $Ry$ | $3 \times Sp+$ $Co + Ga+$ $ry$ |
|---|---|---|---|---|---|---|---|---|
| LBP-HF | 78.8 | **80.2** | 75.6 | 79.5 | 60.4 | 79.1 | 60.6 | 79.0 |
| LPQ | 78.3 | **79.8** | 75.2 | 79.6 | 66.7 | 79.2 | 62.7 | 79.2 |
| ELHF | 77.4 | **80.9** | 76.4 | 80.5 | 63.9 | 79.0 | 63.1 | 79.0 |
| LBP | 78.1 | 79.8 | 74.9 | **80.2** | 64.6 | 79.4 | 63.2 | **80.2** |
| RICLBP | 76.7 | **79.0** | 75.2 | 78.3 | 63.8 | 77.6 | 60.7 | 75.6 |
| MLPQ | 78.3 | **78.8** | 72.2 | 77.0 | 69.1 | 76.6 | 68.1 | 76.6 |
| HASC | 80.5 | **81.8** | 79.7 | 81.6 | 63.4 | 80.7 | 65.5 | 80.7 |
| GF | 80.4 | **81.8** | 79.1 | 80.8 | 66.2 | 79.5 | 68.4 | 79.4 |
| **Non Texture Descriptor** | | | | | | | | |
| FUSac | 75.7 | | | | | | | |
| **Ensembles** | | | | | | | | |
| F2 | - | **82.2** | - | 81.5 | - | 80.5 | - | 80.1 |
| F3 | - | **81.9** | - | 81.4 | - | 80.7 | - | 80.1 |
| Full_light_W | - | **90.9** | - | 90.6 | - | 90.5 | - | 90.6 |
| Full_W | - | **90.9** | - | 90.7 | - | 90.5 | - | 90.6 |

- $Ga$ reports the performance obtained by extracting features from the Gammatonegram image;
- $Ry$ reports the performance obtained by extracting features from the Rythm image.

For each of the proposed ensemble methods, four values are reported: i) the result obtained by the descriptors extracted using the $Sp$ protocol; ii) the result obtained by the descriptors extracted using the $3{\times}Sp + Co$ protocol; iii) the result obtained by the descriptors extracted using the $3{\times}Sp+Co+Ga$ protocol; and iv) the result obtained by the descriptors extracted using the $3{\times}Sp+Co+Ga+$ $Ry$ protocol.

**Table 3. Recognition accuracy (%) on the GTZAN dataset.**

| **Texture Descriptors** | $Me$ | $Sp$ | $Co$ | $3 \times Sp+$ $Co$ | $Ga$ | $3 \times Sp+$ $Co + Ga$ | $Ry$ | $3 \times Sp+$ $Co + Ga+$ $ry$ |
|---|---|---|---|---|---|---|---|---|
| LBP-HF | 80.4 | 82.3 | 78.7 | 82.5 | 79.0 | **83.0** | 34.6 | **83.0** |
| LPQ | 81.9 | 83.5 | 83.4 | 84.1 | 80.1 | **84.6** | 49.7 | **84.6** |
| ELHF | 81.7 | 83.9 | 82.1 | 83.5 | 79.2 | 84.5 | 47.5 | **84.6** |
| LBP | 83.1 | **84.8** | 80.7 | 84.3 | 79.7 | **84.8** | 47.3 | **84.8** |
| RICLBP | 81.8 | 83.9 | 81.7 | 84.2 | 84.1 | **85.2** | 42.0 | **85.2** |
| MLPQ | 83.3 | 84.5 | 81.5 | 84.6 | 83.3 | **85.2** | 62.8 | **85.2** |
| HASC | 83.3 | 84.7 | 84.7 | 85.1 | 80.0 | 85.4 | 50.5 | **85.5** |
| GF | 83.6 | 85.4 | 83.6 | 85.2 | 74.9 | **85.5** | 53.7 | 85.3 |
| **Non Texture Descriptor** | | | | | | | | |
| FUSac | 80.2 | | | | | | | |
| **Ensembles** | | | | | | | | |
| F2 | - | 86.1 | - | 86.2 | - | 86.2 | - | **86.3** |
| F3 | - | 86.1 | - | **86.5** | - | **86.5** | - | **86.5** |
| Full_light_W | - | 89.6 | - | **90.7** | - | **90.7** | - | **90.7** |
| Full_W | - | **90.6** | - | 90.5 | - | 90.5 | - | 90.5 |

[5] http://www.din.uem.br/yandre/birds/bird_songs_46.tar.gz

[6] https://www.kaggle.com/c/whale-detection-challenge

**Table 4. Recognition accuracy (%) on the BIRD-46.**

| **Texture Descriptors** | $Me$ | $Sp$ | $Co$ | $3 \times Sp+$ $Co$ | $Ga$ | $3 \times Sp+$ $Co+Ga$ | $Ry$ | $3 \times Sp+$ $Co+Ga+$ $ry$ |
|---|---|---|---|---|---|---|---|---|
| LBP-HF | 80.6 | 82.0 | 86.3 | 84.2 | 80.9 | 86.5 | 30.8 | **86.6** |
| LPQ | 82.5 | 85.9 | 87.7 | 87.5 | 86.1 | **88.6** | 43.7 | **88.6** |
| ELHF | 84.7 | 86.1 | **88.6** | 87.3 | 86.0 | 88.3 | 41.3 | 88.4 |
| LBP | 80.0 | 85.3 | 87.0 | 87.1 | 85.0 | **88.3** | 40.7 | **88.3** |
| RICLBP | 84.4 | 85.7 | 88.6 | 87.5 | 86.2 | 88.6 | 45.7 | **88.7** |
| MLPQ | 86.4 | 86.6 | 87.3 | 87.7 | 86.9 | **88.7** | 45.2 | **88.7** |
| HASC | 84.1 | 85.9 | 88.3 | 87.3 | 84.1 | **88.8** | 60.3 | **88.8** |
| GF | 85.5 | 87.0 | 87.8 | 87.7 | 77.3 | **88.8** | 43.5 | **88.8** |
| **Non Texture Descriptor** | | | | | | | | |
| FUSac | 88.6 | | | | | | | |
| **Ensembles** | | | | | | | | |
| F2 | - | 89.2 | - | 89.8 | - | 90.8 | - | **90.9** |
| F3 | - | 89.2 | - | 89.8 | - | 90.8 | - | **90.9** |

*The ensembles FULL and Full_light are not reported since OLDac was based on a commercial system and we have not the features extracted for this dataset.

Examining the results reported in Tables 1 to 4, the following conclusions can be drawn:

- Using different spectrogram images improves the performance of each descriptor;
- The fusion between acoustic and visual features greatly improves the accuracy of a system based only on acoustic or visual features.
- The fusion among $Sp$, $Co$, $Ga$ and $Ry$ is useful in all the datasets except in ISMIR 2004, where it deteriorates the performance.
- F2 coupled with $Sp$ is best ensemble of visual features considering all the four datasets, while Full_W coupled with $Sp$ is best ensemble (considering both visual and acoustic features) among the three music genre datasets.

In addition, we have combined F3 with the acoustic features tested in (26) obtaining the following accuracy scores in the BIRD dataset: 94.5% ($Sp$), 94.7% ($3 \times Sp+Co$), 95.1% ($3 \times Sp+Co+Ga$), and 95.2% ($3 \times Sp+Co+Ga+Ry$).

We checked the error independence with Yule's Q-statistic (15). The values of $Q$ are bounded between [-1,1]. Classifiers that recognize the same patterns correctly have $Q > 0$; those which commit errors on different patterns have $Q < 0$. In Table 5, the Q-statistic of SMVs trained using LPQ features extracted from different images are reported. Results show that different descriptors train a set of partially uncorrelated classifiers.

**Table 5. Q-statistic among SVMs trained using LPQ.**

| | Me | Sp | Co | Ga | Ry |
|---|---|---|---|---|---|
| Me | - | 0.92 | 0.87 | 0.72 | 0.65 |
| Sp | - | - | 0.91 | 0.80 | 0.76 |
| Co | - | - | - | 0.83 | 0.84 |
| Ga | - | - | - | - | 0.65 |
| Ry | - | - | - | - | - |

We have tried to improve the performance (see Table 6) using all the 9 images (three values of dBFS and three bands), method named $CoFull$, to reduce the computation time we have run the experiments using only LPQ. $CoFull$ improve $Co$, anyway $3 \times Sp+CoFull$ obtains performance almost equal to $3 \times Sp+Co$. In future work we would to run new tests for better exploit the information of the color images (e.g. to use color descriptors).

**Table 6. Performance with Color images.**

| | LMD | ISMIR 2004 | GTZAN | Bird | Whale |
|---|---|---|---|---|---|
| Co | 81.4 | 75.2 | 83.4 | **87.7** | 91.5 |
| CoFull | 83.6 | 79.7 | 83.5 | **87.7** | 91.7 |
| 3×Sp+CoFull | **84.0** | **79.9** | **84.1** | 87.5 | **91.9** |

We try to improve the performance of the fusion using other combination approaches (Table 7), we have tested:

- The discriminative accumulation scheme (DAS) (30);
- Multiclass Multi Kernel Learning (MK) (30) (using RBF kernel), which Matlab code is available *http://dogma.sourceforge.net/*.

**Table 7. Ensemble F2 based on different fusion rules.**

| | LMD | ISMIR 2004 | GTZAN | Bird | Whale |
|---|---|---|---|---|---|
| SUM | 86.2 | 82.2 | **86.1** | 89.2 | 92.2 |
| DAS | 86.0 | 82.0 | 85.9 | **89.9** | 92.5 |
| MKL | **86.4** | **83.0** | 86.0 | **89.9** | **92.6** |

MKL permits to slightly improve the performance, but it's a trained method more complex that the sum rule.

## 5. Discussion and Related Work

In this section we present and discuss the existing (to the best of the authors knowledge) works that deal with audio classification tasks using hybrid approaches that use audio and visual features. In this work by visual features we are referring to features extracted from a visual time-frequency representation (i.e. spectrogram) taken from the original sound.

The first work to use visual features for an audio classification task was developed by Yu and Slotine (42). In that work, the authors reported a good performance in musical instrument classification experiments.

Inspired by the work of Yu and Slotine, Costa et al. (5) have used Gray Level Co-Occurrence Matrix features to classify music genres in the Latin Music Database. Their best reported result was of 67.2% using a SVM classifier.

Wu et al. (41) was the first to perform music genre classification using both visual and acoustic features. In that work, the acoustic features were created by using MFCCs and the visual features were extracted with GF. Wu et al. used an early fusion approach to combine the different kinds of features and they performed experiments on the GTZAN and on the ISMIR datasets. In the end, the best obtained accuracy was 86.1% on both datasets using this combined approach.

In a follow up work (6), Costa et al. have compared the performance of GLCM with LBP. They have also performed zoning, creating specific classifiers for each zone which improved the GLCM performance from 67.2% to 70.7%. Using zoning the LBP has outperformed GLMC with 80.3% recognition rate.

In another paper by Costa et al. (8), the authors performed additional experiments with LBP and for the first time they experimented a zoning scheme based on a psyco-acoustic scale (Mel scale). They have used the LMD and ISMIR 2004 databases. Their best obtained results were 82.3% and 80.6% respectively.

The use of other visual approaches such as GF and LPQ were investigated for the LMD by Costa et al. (7). Their best results were 74.6% for GF and 80.7% for LPQ.

In Nanni et al. (24), the authors evaluated different types of visual descriptors for the task of music genre classification on LMD and ISMIR 2004 dataset. Their best reported result were 86.1% and 82.9% respectively. Both results were obtained by combining classifiers made with features taken from different texture operators. An extended version of this paper was published in (26), where the authors performed experiments using three databases (LMD, ISMIR and GTZAN). The main contributions of this extended version was to show that by combining both visual and acoustic feature it is possible to get even better results on the music classification task.

In Nanni et al. (25) the combination of visual and acoustic features was evaluated for the task of bird species classification. The best reported result was 94.5% using f4+FUSac or f5+FUSac on the Bird Songs 46 (BSD 46) dataset, a dataset composed of bird vocalization clips obtained from Xeno-canto website.

Now that we have briefly presented the related work, let us discuss how the results in this paper contrast with the previously reported results.

For the LMD, the best obtained result in this work using an ensemble of different visual descriptors was of 86.3% using F3 (MLPQ + ELHF + LBP + RICLBP + HASC + GF) with $Sp$ and $3 \times Sp + Co + Ga + Ry$ and using an ensemble of audio + visual descriptors was of 84.7% using Full_light_W and $3 \times Sp + Co + Ga + Ry$. Current literature using audio only features have reported achieving 82.3% with Principal Mel-spectrum Components (12), 77.6% with Rhythmic Signatures and Deep Learning (34) and 77.0% with Time constrained sequential patterns (35).

For the ISMIR 2004 dataset, the best obtained results in this work using F2 was of 82.2% with $Sp$ and using the audio + visual descriptors was of 90.9% with Full_light_W and Full_W both with $Sp$. Current literature using audio only features have reported achieving 89.9% with Spectro-temporal features, 85.4% with JSLRR (32) and 79.0% with GVS-SVM with MFCC (3). In (31), the authors report an accuracy of 94.4% with LPNTF. However, as pointed out by Sturm (38), these results arise from a flaw in the experiment inflating accuracies from around 60%.

For the GTZAN dataset, the best obtained results in this work using the F3 was of 86.5% with $3 \times Sp + Co$, $3 \times Sp + Co + Ga$ and $3 \times Sp + Co + Ga + Ry$ and using the audio + visual descriptors was of 90.7% using Full_light_W with $3 \times Sp + Co$, $3 \times Sp + Co + Ga$ and $3 \times Sp + Co + Ga + Ry$. Current literature using audio only features have reported achieving 89.40% with JSLRR (32), 87.4% with Spectro-temporal features, and 82.1% with GVS-SVM with MFCC (3). Panagakis et al. reported 92.4% of accuracy using LPNTF (31), but this result was inflated as well as on ISMIR 2004 dataset.

For the BSD 46, the best obtained results in this work using both F2 and F3 was of 90.9% with $3 \times Sp + Co + Ga + Ry$ and using the audio+visual descriptors was of 95.2% with F3+($3 \times Sp + Co + Ga + Ry$). These results are better than the ones we obtained in previous work (25) where the best reported results were 89.5% using ICTAI-F3 (F3: ELHF + LPQ + RICLBP + HASC + GF) and visual+audio of 94.5% using ICTAI-F4 (ELHF + LPQ + RICLBP + HASC + LBP + GF) or ICTAI-F5 (ELHF + MLPQ + RICLBP + HASC + LBP + GF) + ICTAI-FUSac ($3 \times SSD + Rh + 3 \times Rp$).

**Table 8. Comparison with the state-of-the-art.**

| | | Dataset | | | | |
|---|---|---|---|---|---|---|
| **Method** | **Features** | **LMD** | **ISMIR** | **GTZAN** | **BIRD** | **Whale** |
| | | | **2004** | | **Song** | |
| Here, F2 (only visual features) | Visual | **86.2** | 82.2 | 86.1 | **89.2** | **92.2** |
| Best visual ensemble of (26) | Visual | 86.1 | 81.6 | 83.8 | 85.9 | 87.1 |
| Here, Full_W | Vis+Ac | 84.6 | **90.9** | **90.6** | - | - |
| Best visual + acoustic of (26) | Vis+Ac | 85.1 | 90.2 | 89.8 | - | - |
| LBP (8) | Visual | 82.3 | 82.1 | - | - | - |
| GLCM (6) | Visual | 70.7 | - | - | - | - |
| Gabor filter (41) | Visual | - | 82.2 | 82.1 | - | - |
| GSV + GF (41) | Vis+Ac | - | 86.1 | 86.1 | - | - |
| LPQ (7) | Visual | 80.8 | - | - | - | - |
| Gabor filter (7) | Visual | 74.7 | - | - | - | - |
| MARSYAS features (39) | Acoustic | - | - | 61.0 | - | - |
| GSV-SVM+ MFCC (3) | Acoustic | 74.7 | 79.0 | 82.1 | - | - |
| Spectro-temporal features (18) | Acoustic | - | 89.9 | 87.4 | - | - |
| Principal Mel-spectrum components (12) | Acoustic | 82.3 | - | - | - | - |
| Time constrained sequential patterns (35) | Acoustic | 77.0 | - | - | - | - |
| Rhythmic signatures + Deep learning (34) | Acoustic | 77.6 | - | - | - | - |
| Block-level (37) | Acoustic | 79.9 | 88.3 | 85.5 | - | - |
| JSLRR (32) | Acoustic | - | 85.45 | 89.40 | - | - |

On the right whale calls dataset, the obtained results confirm, one more time, that the classification protocol presented here performs better than those presented in (26). Table 8 shows the best results obtained here and the state-of-the-art on the five datasets used in this work.

In this section we have shown that the ensemble based on both acoustic and the visual features proposed here obtains state-of-the-art performance in all datasets. The only approach we are aware of that outperforms the proposed ensemble is (31), a method that greatly outperforms all the other published approaches. Moreover, our proposed ensemble based solely on visual features outperforms previous works based on visual features (i.e. texture descriptors).

It is important to highlight the fact that our approach obtained these excellent results across all the datasets without the use of ad hoc tuning (i.e. the same parameters for SVM, the same descriptors, and the same weights in the weighted sum rule were used on all datasets).

## 6. Conclusion

In this work we present a novel system for audio classification tasks that combines audio and visual features. Different texture descriptors were extracted and compared from different images derived from the audio file. Different acoustic feature vectors were also evaluated and compared. The experiments reported in this paper demonstrate that the fusion of visual and audio features improves performance compared with the state-of-the-art. Our proposed ensemble based only on texture features reaches results comparable with existing audio approaches, and the fusion of visual with audio features improves performance even further.

Many different types of images generated from audio files can now be explored, and deeper reviews of textural features can be investigated for maximizing the classification performance of audio images.

## References

[1] S. M. Biagio, M. Crocco, M. Cristani, S. Martelli, and V. Murino. Heterogeneous auto-similarities of characteristics (HASC): exploiting relational information for classification. In *IEEE International Conference on Computer Vision (ICCV), 2*, pages 809–816. IEEE, 2013.

[2] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, and N. Wack. ISMIR 2004 audio description contest. Technical report, Music Technology Group, Barcelona, Spain, 2006.

[3] C. Cao and M. Li. Thinkit's submissions for MIREX 2009 audio music classification and similarity tasks. In *MIREX abstracts, International Conference on Music Information Retrieval*, 2009.

[4] C. H. L. Costa, J. D. Valle Jr., and A. L. Koerich. Automatic classification of audio data. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 562–567, The Hague, Netherlands, 2004.

[5] Y. M. G. Costa, L. E. S. Oliveira, A. L. Koerich, and F. Gouyon. Music genre recognition using spectrograms. In *18th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 151–154. IEEE, 2011.

[6] Y. M. G. Costa, L. E. S. Oliveira, A. L. Koerich, and F. Gouyon. Comparing textural features for music genre classification. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1867–1872. IEEE, 2012.

[7] Y. M. G. Costa, L. E. S. Oliveira, A. L. Koerich, and F. Gouyon. Music genre recognition using gabor filters and LPQ texture descriptors. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 67–74. Springer, 2013.

[8] Y. M. G. Costa, L. E. S. Oliveira, A. L. Koerich, F. Gouyon, and J. G. Martins. Music genre classification using LBP textural features. *Signal Processing*, 92(11):2723–2737, 2012.

[9] D. P. W. Ellis. Gammatone-like spectrograms, 2009.

[10] A. Flexer. A closer look on artist filters for musical genre classification. *World*, 19(122):1–4, 2007.

[11] J. F. Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, and A. Toncheva. The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011. *IDC white paper, EMC sponsored*, 2008.

[12] P. Hamel. Pooled features classification. *Submission to Audio Train/Test Task of MIREX*, 2011.

[13] R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[14] C. N. Silla Jr., A. L. Koerich, and C. A. A. Kaestner. The latin music database. In *International Conference on Music Information Retrieval*, pages 451–456, Philadelphia, USA, 2008.

[15] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

[16] T. Lidy. Audio feature extraction, 2007.

[17] T. Lidy and A. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *ISMIR*, pages 34–41, 2005.

[18] S-C. Lim, J-S. Lee, S-J. Jang, S-P. Lee, and M. Y. Kim. Music-genre classification system based on spectro-temporal features and feature selection. *IEEE Transactions on Consumer Electronics*, 58(4):1262–1268, 2012.

[19] M. T. Lopes, L. Gioppo, T. Higushi, C. A. A. Kaestner, C. N. Silla Jr., and A. L. Koerich. Automatic bird species identification for large number of species. In *International Symposium on Multimedia (ISM)*, pages 117–122. IEEE, 2011.

[20] D. R. Lucio and Y. M. G. Costa. Bird species classification using spectrograms. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applic. p. 543-550*, 2015.

[21] A. Montalvo, Y. M. G. Costa, and J. R. Calvo. Language identification using spectrogram texture. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 543–550. Springer, 2015.

[22] L. Nanni, S. Brahnam, and A. Lumini. Combining different local binary pattern variants to boost performance. *Expert Systems with Applications*, 38(5):6209–6216, 2011.

[23] L. Nanni, S. Brahnam, A. Lumini, and T. Barrier. Ensemble of local phase quantization variants with ternary encoding. In *Local Binary Patterns: New Variants and Applications*, pages 177–188. Springer, 2014.

[24] L. Nanni, Y. M. G. Costa, and S. Brahnam. Set of texture descriptors for music genre classification. In *22th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 2014.

[25] L. Nanni, Y. M. G. Costa, D. R. Lucio, C. N. Silla Jr., and S. Brahnam. Combining visual and acoustic features for bird species classification. In *IEEE International Conference on Tools with Artificial Intelligence*, 2016.

[26] L. Nanni, Y. M. G. Costa, A. Lumini, M-Y. Kim, and S. R. Baek. Combining visual and acoustic features for music genre classification. *Expert Systems with Applications*, 45:108–117, 2016.

[27] R. Nosaka, C. H. Suryanto, and K. Fukui. Rotation invariant co-occurrence among adjacent LBPs. In *Computer Vision-ACCV*, pages 15–25. Springer, 2013.

[28] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[29] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In *Image and signal processing*, pages 236–243. Springer, 2008.

[30] F. Orabona, L. Jie, and B. Caputo. Multi kernel learning with online-batch optimization. *Journal of Machine Learning Research*, 13(Feb):227–253, 2012.

[31] Y. Panagakis, C. L. Kotropoulos, and G. R. Arce. Music genre classification using Locality Preserving Non-Negative Tensor Factorization and Sparse Representations. In *ISMIR*, pages 249–254, 2009.

[32] Y. Panagakis, C. L. Kotropoulos, and G. R. Arce. Music genre classification via joint sparse low-rank representation of audio features. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 22(12):1905–1917, 2014.

[33] R. D. Patterson. *Complex Sounds and Auditory Images*. Pergamon, Oxford, 1992.

[34] A. Pikrakis. Audio latin music genre classification: A MIREX submission based on a deep learning approach to rhythm modelling. 2013.

[35] J-M. Ren and J-S. R. Jang. Discovering time-constrained sequential patterns for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1134–1144, 2012.

[36] M. R. Schroeder, B. S. Atal, and J. L. Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *The Journal of the Acoustical Society of America*, 66(6):1647–1652, 1979.

[37] K. Seyerlehner, M. Schedl, T. Pohle, and 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.

[38] Bob L Sturm. Classification accuracy is not enough. *Journal of Intelligent Information Systems*, 41(3):371–406, 2013.

[39] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

[40] S. Umesh, L. Cohen, and D. Nelson. Fitting the mel scale. In *International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, volume 1, pages 217–220. IEEE, 1999.

[41] M-J. Wu, Z-S. Chen, J-S. R. Jang, J-M. Ren, Y-H. Li, and C-H. Lu. Combining visual and acoustic features for music genre classification. In *International Conference on Machine Learning and Applications*, pages 124–129. IEEE, 2011.

[42] G. Yu and J-J. Slotine. Audio classification from time-frequency texture. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1677–1680. IEEE, 2009.

[43] G. Zhao, T. Ahonen, J. Matas, and Matti M. Pietikäinen. Rotation-invariant image and video description with local binary pattern features. *IEEE Transactions on Image Processing*, 21(4):1465–1477, 2012.

[44] E. Zwicker, H. Fastl, and H. Frater. Psychoacoustics, facts and models, v. 22, springer series of information sciences, 1999.