Image features in music style recognition

Kamil Behúň*
Supervised by: Michal Hradíš†

Department of Computer Graphics and Multimedia
Brno University of Technology
Brno / Czech Republic

Abstract

This work presents a novel approach to music style recognition inspired by feature extraction techniques used in image classification. To be able to utilize the image classification techniques, the 1D sound signal is transformed to its 2D representation — to a Mel-frequency spectrum. Small local areas of the spectrum are represented by 128-dimensional SIFT descriptors from which Bag of Words (BOW) representation of a whole signal is constructed. Dictionary used for translating the descriptors to the BOW representation is created by K-means algorithm. The BOW feature vectors are classified by non-linear Support Vector Machine classifier. The proposed approach was tested on publicly available music recognition data set GTZAN. The achieved results (86.4% classification rate) compare favorably to previous results on this dataset — 84.3% classification rate in (Panagakis et al., 2010), 74% classification rate in (Holzapfel and Stylianou, 2008). Further, performance of the proposed audio parametrization was tested on video semantic category detection task from TRECVID evaluations, where it provides results superior to standard audio parametrizations.

Keywords: Music genre recognition, Mel-frequency spectrum, Local features, SIFT descriptors, Bag of Words, Support Vector Machine

1 Introduction

Automatic music style recognition has potential applications in on-line, as well as personal music databases. It can provide music genre information in cases where it is not available, and thus support navigation and searching in such music database. Furthermore, the problem of music genre recognition is related to the task of automatically suggesting suitable songs for users based on their particular taste or personalized song ratings.

A music genre or a music style is a conventional category of music works which all share some common attributes or origin (e.g. the period during which a musical composition was written, its instrumentation and treatment of those instruments, its form). Although the terms music genre and music style can be interpreted as being different, sometimes, they are used interchangeably. In the rest of the text, only the term genre is used, as the experiments are performed on a dataset of well established music genres (e.g. jazz, rock and country). Music genre recognition is a task, where the goal is to identify the attributes (features) specific for a particular genre, and estimate the genre of a given music piece based on these features.

Music genre recognition can be treated as a standard classification task which has two main parts - feature extraction and classification. Classification algorithms are usually task-independent and standard of-the-shelf classifiers can be used for most tasks. On the other hand, feature extraction is task-specific. It has to extract information from the classified data which is relevant for the particular task while suppressing effects non-informative sources of variation (e.g. lighting conditions in visual object detection, and compression or microphone quality in speaker recognition). For these reasons, feature extraction is usually hand-crafted for particular application, as is the case of music genre recognition [39]. However, several approaches have been recently suggested which learn features [28, 16]. These feature learning approaches were shown to be able to learn good representations for various data and achieve state-of-the-art results in wide range of classification tasks.

In this work, we explored different idea. We attempted to transfer well established features in one area to completely different task and data. Inspired by success of local image features [25] and Bag of Word representation [13] in image classification, we applied these methods to audio data. We show that the image features can be applied to audio data by transforming the 1D signals to spectrograms (2D representation), and that this type of feature extraction is able to achieve state-of-the-art results in music genre recognition, and that it surpasses standard audio features in semantic class detection as well.

Next section summarizes previous work in music genre recognition. Section 3.1 introduces the proposed feature extraction method and Section 3.2 discusses classification methods used in experiments. Section 4 presents datasets on which the approach has been evaluated and Section 5 describes the experiments and discusses the achieved re-
results. Finally Section 6 concludes the paper and outlines possible future work.

2 Previous work

When the music genre recognition is approached as a pattern recognition task, it consists of two parts. These parts are feature extraction and classification. This section describes previously proposed methods for classification and feature extraction for music genre recognition.

According to Tzanetakis at al. [39] features for music recognition can be divided into three basic types. These types are Pitch content features, Rhythmic content features and Timbral texture features. This classification is based on the type of information extracted by the features from audio signal.

Pitch content features characterize audio signals in terms of energy of different frequency bands. For example, in [39] the authors propose to detect few dominant pitches in a short time window from which histograms of the dominant pitches across a whole song are computed. From the histograms, a small number of feature is extracted (e.g. amplitude of maximum peak, pitch interval between the two most prominent peaks).

Rhythmic content features represent rhythmic structure of the music. Method of Tzanetakis et al. [39] is based on detecting the most salient periodicities of the signal from which a beat histogram is created. Based on the beat histogram, several features are computed (e.g. relative amplitude of the two highest histogram peaks, frequency of the highest peaks). Another example of Rhythm content features is proposed in [20, 19, 22] where a 24-band psycho-acoustically modified spectrogram which reflects human loudness sensation is computed. Fourier transform is applied to each of the 24 frequency bands to obtain spectra of loudness changes. Several types of features are then extracted from the loudness change spectra. The loudness change spectra is called by the authors Rhythm Pattern. The extracted features are average spectrum across the 24 frequency bands, statistical descriptors of each frequency band (mean, median, variance, skewness, kurtosis, min- and max-values), variance of each loudness change frequency across all 24 frequency bands, and temporal variation of the Rhythm Patterns extracted from short (six second) time windows.

Timbral texture features should exhibit properties related to general timbre of the sound. They are based on short-time Fourier transform and they are calculated on short time frames of sound. This group includes many features, for example Spectral Centroid, Spectral Rolloff, Spectral Flux, Time Domain Zero Crossings [34], discrete wavelet transform coefficients [17, 23, 10], histogram of the wavelet coefficients [18], octave-based spectral contrast [40], and Mel-Frequency Cepstral Coefficients (MFCC). Mel-Frequency Cepstral Coefficients are very general features traditionally used in speech recognition [31, 29], and they provide good results in music genre recognition [19, 15, 8, 40, 39] as well.

Bergstra et al. [4], use combination of several Timbral texture features. After computing different frame-level Timbral texture features, they group non-overlapping blocks of consecutive frames into segments from which means and variances over a whole music signal are computed (means and variances are used as input to weak learners in AdaBoost).

The mentioned types features can be used individually; however, the best results are obtained by their combination. In [19], the authors combine features of different types (Pitch content features, Rhythmic content features, Timbral texture features) into hybrid features.

Many approaches which can not be clearly assigned to the above described groups exist. An example is the Bio-Inspired Joint Acoustic and Modulation Frequency Representation of Music [30]. In this method, an auditory temporal modulation representation is computed from an auditory spectrogram. This representation discards much of the spectro-temporal details, and focuses on the underlying slow temporal modulations of the audio signal.

Panagakis et al. [30] describe feature extraction methods called Multilinear Subspace Analysis Techniques. These techniques are Multilinear principal components analysis (MPCA), Non-negative tensor factorization (NTF), and Higher-order singular value decomposition (HOSVD) which are computed from tensors. Linear counterparts of these methods are Non-negative matrix factorization (NMF), Singular value decomposition (SVD) and Principal component analysis (PCA), which can be viewed as a special cases for first-order tensors (vectors).

Any type of classifier can be used for music genre recognition. In literature, the most commonly used classifier is Support Vector Machine (SVM) [20, 19, 22, 17, 18]. Compared to other classifiers, SVM was shown to give superior results [20, 17, 18]. Other classification methods used for music genre recognition are Gaussian mixture models [40, 39, 34], K-nearest neighbor classifier [20, 10, 39, 34] (K-NN), Round Robin ensemble [10] and Adaboost [4].

3 Method

As mentioned earlier, music genre recognition has two main parts — feature extraction and classification. The proposed feature extraction approach relies on transformation of 1D audio signal to 2D representation, and on consequent application of local feature extraction methods from the field of image classification. As a classifier, we selected SVM, because it provides state-of-the-art results in music genre recognition [20, 19]. Structure of our approach is shown in Figure 1, and a detailed description of each part of this structure is presented in the following text.
Figure 1: The processing pipeline: Audio signal is segmented, and Mel-frequency spectrograms are computed for the segments. Local features (SIFT) are extracted from the spectrograms on a regular grid. The local descriptors are translated to codewords by codebook, and Bag of Words representation is computed as a occurrence histogram of the codewords. The Bag of Words representation is used as an input to classification by SVM.

### 3.1 Feature extraction

This section presents a novel approach to music genre recognition inspired by feature extraction techniques used in image classification. Common approach in image classification [36] is to represent local parts of an image by a high-dimensional descriptors. Such descriptors encode appearance of the image patches, for example, spatial and directional distribution of gray-scale gradients as in SIFT descriptor [25, 24]. Such local descriptors are also called local image features in computer vision [33, 26]. To create a more compact representation, the local features can be assigned numerical identifiers based on their similarity to a set of prototypes [9]. The set of prototypes is called a visual codebook and the resulting image patches with assigned prototype IDs are called visual words.

Similarly to a text document, which can be described by the counts of individual words it contains, an image can be represented by counts of visual words [12]. Such representation is called Bag of Visual Words (BOW). BOW discards information about spatial relations. It discards any information about positions of the visual words, and retains only information of the local appearances.

To be able to use the local feature techniques from image classification, a 1D sound signal has to be transformed to its 2D representation. A natural way to create such 2D representation is to describe the sound signal in terms of energies of different frequency bands in short time windows. Methods for represent 1D signals in this way include short-term Fourier transform [2], wavelet transform [2] and many others. For the purpose of music genre recognition, we have chosen Mel-frequency spectrogram [38] as the 2D representation. The Mel-frequency spectrogram is similar to short-term Fourier transform, except the frequency scale is logarithmic which is closer to how humans perceive sounds (e.g. music intervals are multiples of frequencies instead of fixed additions).

In our work, the Mel-frequency spectra were obtained by segmenting the audio track into 100 ms segments with 25 ms overlap. Mel-frequency spectra with 128 frequency bands and maximum frequency 12 KHz was computed from each of the segments. In order to be able to use existing local image feature methods, dynamic range and contrast of the spectrograms were reduced by

$$x = \left( \frac{\log(e + 1)}{\log(\maxV)} \right) \times 255,$$  \hspace{1cm} (1)

where $e$ is a value from the original Mel-frequency spectrum, $\maxV$ is logarithm of the maximal value in the original Mel-frequency spectrogram and $x$ is the resulting value in the lower dynamic range spectra. The transformation from equation 1 assures that the resulting values are in interval $< 0, 255 >$, and that they correspond to how humans perceive sound intensity (perception of acoustic intensity is logarithmic). Example of the obtained spectrogram is show in Figure 2.

Then the spectrograms were handled as images and local features were extracted from them. As the spectrograms do not exhibit any stable and distinct areas which could be detected by interest region detectors [27], we sampled the spectrograms on a regular grid with cell size $8 \times 8$ pixels and SIFT descriptors [24] were computed from the sampled small circular areas. The SIFT descriptor computes 16 Histograms of Oriented Gradients (HOG) on a 4 by 4 grid in the area of interest. Each of the histograms has 8 orientational bins, and magnitude of image gradient in each pixel is distributed between the closest bins according to its orientation and between the spatially closest histograms. The resulting SIFT descriptor is a vector of 128 values created by concatenating the 16 Histograms of Oriented Gradients.

To create a feature vector for a whole audio recording, extracted local features are aggregated to a Bag of Word (BOW) representation. In order to obtain the BOW representation, local features are first translated to visual words by codebook transform. To create a codebook, it is possible use to any clustering or quantization algorithm. The most commonly used is the k-means algorithm [9]. We
used k-means algorithm with Euclidean distance to obtain the set of prototypes which constitute the codebook - cluster centers become the prototypes. In the experiments, we used 4096 codewords (clusters) and initial positions of cluster centers were chosen randomly from training data set. Note that the goal of the k-means clustering in this case is not to find tight clusters of feature descriptors, rather, it provides good coverage of all possible descriptors, and it achieves low reconstruction error when the descriptors are quantized by the prototypes.

When assigning local features to the single closest codeword, quantization errors occur and some information is lost. This is especially significant in high-dimensional spaces, as is the case of the local patch descriptors where distances to several nearest codewords tend to be very similar. In the context of image classification, this issue was discussed for example by Gambert et al. [9], who propose to distribute the local descriptors to several close codewords according to codeword uncertainty. Computation of BOW with codeword uncertainty is defined for each codeword \( w \) from a codebook \( B \) as

\[
\text{UNC}(w) = \sum_{p \in P} \frac{K(w, p)}{\sum_{v \in B} K(v, p)},
\]

where \( P \) is a set of local image features and \( K \) is a kernel function. We use Gaussian kernel

\[
K(w, w') = \exp\left(-\frac{\|w - w'|^2}{2\sigma^2}\right),
\]

where \( \sigma \) defines the size of the kernel. In our experiments, value of \( \sigma \) was set as the average distance between two closest neighboring codewords from a codebook. Equation 2 computes contributions of all local features to a particular word \( w \) and it sums the individual contributions. The BOW representation is a vector of \( \text{UNC}(w) \) values for all words from a codebook normalized to unit length.

### 3.2 Classification

Our experiments were performed with Support Vector Machine classifier (SVM) [6] which is often used for various tasks in image and video classification [14, 33, 9, 36, 35]. SVM has four main advantages. It generalizes well, it can use kernels, it is easy to work with, and good-quality SVM solvers are available. Although non-linear kernels have been shown to perform better over the linear kernel in image recognition [32], we use the linear kernel in some experiments due to computational reasons. In addition to the linear kernel, Gaussian kernel

\[
K(x, x') = \exp\left(-\gamma\|x - x'\|^2\right),
\]

was used in the experiments as well. Optimal value of the SVM regularization parameter \( C \) and the Gaussian kernel scale \( \gamma \) were estimated by grid search with 10-fold cross-validation with stratified sampling of training dataset.

Classifiers with the Gaussian kernel were trained using LIBSVM\(^1\) [5] implementation of a solver for the standard soft-margin SVM formulation [6] which has a single regularization parameter \( C \). Linear classifiers were trained using LIBLINEAR [7] which is able to handle large linear SVM problems efficiently.

### 4 Data sets

The approach, created for music genre recognition, was tested on GTZAN Genre collection\(^2\). This dataset was collected by G. Tzanetakis [39]. We chose GTZAN Genre collection, in order to be able to compare to previously published result on this dataset [30, 11, 3, 21, 4, 18, 39]. GTZAN Genre collection contains 1000 audio tracks, where each track is 30 seconds long. These tracks are divided into 10 music genres — namely Blues, Classical,

\(^1\)Experiments with Gaussian kernel SVM were concluded in RapidMiner which is a open-source tool for data mining, machine learning, text mining and other classification operations (http://rapid-i.com/).

\(^2\)http://marsyas.info/download/data_sets
Country, Disco, HipHop, Jazz, Metal, Pop, Reggae, and Rock. Each genre is represented by 100 tracks. The tracks are all 22050Hz mono 16-bit audio files.

Further, performance of the proposed audio parametrization was tested on data from semantic indexing task from TRECVID evaluations. The purpose of the semantic indexing task is to automatically detect semantic categories in video. Detection of semantic categories can be understood as video tagging which is, for example, performed by users when uploading videos to Internet archives (e.g. YouTube). The videos in semantic indexing task are tagged at the level of video shots.

In our experiments, we used a subset of TRECVID 2011 training data. First 20,000 shots from the TRECVID dataset were chosen as a training set, and following 50,000 shots were used for testing.

The TRECVID data was collaboratively annotated in an active learning setup\(^3\) [1] which resulted in 346 usable semantic classes with enough annotations. From the annotated classes, we chose 55\(^4\) the most common classes in our training set. The training set contains 29478 positive and 312398 negative class annotations. The testing set contains 86091 positive and 820057 negative annotations.

In the TRECVID data, some shots are very short while others are several minutes long. In order compensate for these difference, we chose a representative audio segment for a shot to be centered on the shot with duration restricted to be between 10 s and 20 s.

### Table 1: Confusion matrix on the GTZAN genre collection achieved by parametrization for 32 × 32 local features (classification accuracy 86.4%).

<table>
<thead>
<tr>
<th>Genre</th>
<th>Blues</th>
<th>Metal</th>
<th>Classical</th>
<th>Rock</th>
<th>Disco</th>
<th>Hiphop</th>
<th>Country</th>
<th>Jazz</th>
<th>Reaggae</th>
<th>Pop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blues</td>
<td>92</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Metal</td>
<td>0</td>
<td>98</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Classical</td>
<td>2</td>
<td>0</td>
<td>83</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Rock</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>88</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Disco</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>86</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Hiphop</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>91</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Country</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>94</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Jazz</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>77</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Reggae</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>83</td>
<td>5</td>
</tr>
<tr>
<td>Pop</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>72</td>
</tr>
</tbody>
</table>

5 Experiments and results

This section contains description of experiments and results in genre recognition and semantic indexing. Our main experiment for both tasks was aimed to explore the effect of local feature size, in order to estimate what size of the local features is the best for the solved tasks (small size is capture detail and bigger sizes represent larger context). Results achieved on the GTZAN dataset were compared to results of other music genre recognition approaches, and result achieved on TRECVID 2011 dataset were compared to two standard audio parametrizations.

Following the experimental setup used in [30, 39], stratified tenfold cross validation is employed to estimate performance in experiments conducted on the GTZAN dataset. That means, each training set contained 900 tracks (9 parts of validation), testing set contained 100 tracks (1 part of validation) from GTAZAN, and the parts were gradually alternated for training and testing in the experiment. This experimental setup allows for comparison to other published approaches.

We did experiments for 8 × 8, 16 × 16 and 32 × 32 pixels sizes of the local features extracted from spectograms. SVM with RBF kernel function was used. Results for these different sizes are shown in Table 4. The best classification accuracy (86.4%) was achieved by parametrization for 32 × 32 local features. Confusion matrix for this result is shown in Table 1.

### Table 4: Classification accuracy achieved on GTZAN Genre collection for different sizes of local features.

<table>
<thead>
<tr>
<th>Size of loc. features</th>
<th>Classification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 × 8</td>
<td>83.4%</td>
</tr>
<tr>
<td>16 × 16</td>
<td>84.2%</td>
</tr>
<tr>
<td>32 × 32</td>
<td>86.4%</td>
</tr>
</tbody>
</table>

\(^3\)http://mrim.imag.fr/tvca/

\(^4\)The selected classes were: person, outdoor, face, vegetation, single person, male person, adult, indoor, trees, daytime outdoor, overlaid text, female person, computer or television screens, plant, entertainment, sky, building, vehicle, suburban, singing, ground vehicles, landscape, streets, road, celebrity entertainment, actor, walking running, instrumental musician, scene text, sitting down, animal, child, waterscape waterfront, car, doorway, news studio, walking, politics, cityscape, two people, hand, city, quadruped, mammal, carnivore, explosion fire, reporters, mountain, domesticated animal, female-human-face-closeup, dark-skinned people, graphic, crowd, chair, teenagers
Table 2: Classification accuracies achieved by our approach and other published approaches for GTZAN Genre collection.

<table>
<thead>
<tr>
<th>Approach (features + classifier)</th>
<th>Classification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>mel-spectrogram - SIFT + SVM (this work)</td>
<td>86.4%</td>
</tr>
<tr>
<td>non-negative MPCA + SVM [30]</td>
<td>84.3%</td>
</tr>
<tr>
<td>aggregate features + AdaBoost [4]</td>
<td>82.5%</td>
</tr>
<tr>
<td>wavelet histograms + SVM [18]</td>
<td>78.5%</td>
</tr>
<tr>
<td>audio and symbolic features + SMV [21]</td>
<td>76.8%</td>
</tr>
<tr>
<td>many features + NTF [3]</td>
<td>75.0%</td>
</tr>
<tr>
<td>spectrogram - NMF + GMM [11]</td>
<td>74.0%</td>
</tr>
<tr>
<td>wavelet histograms + LDA [18]</td>
<td>71.3%</td>
</tr>
<tr>
<td>pitch, rhythmic and timbral features + GMM [39]</td>
<td>61.0%</td>
</tr>
</tbody>
</table>

Table 3: Mean Average Precisions achieved in semantic indexing task on TRECVID 2011 training set.

<table>
<thead>
<tr>
<th>Features</th>
<th>Mean Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>mel-spectrogram - SIFT 8 × 8 + SVM</td>
<td>0.144</td>
</tr>
<tr>
<td>mel-spectrogram - SIFT 16 × 16 + SVM</td>
<td>0.147</td>
</tr>
<tr>
<td>SPEC + SVM</td>
<td>0.115</td>
</tr>
<tr>
<td>MFCC + SVM</td>
<td>0.113</td>
</tr>
</tbody>
</table>

6 Conclusions

This paper presented a novel feature extraction approach for audio data which is based on SIFT local features and Bag of Word representation which were previously used with good results in image classification. The experiments show that the proposed feature extraction provides very good results in music genre recognition task. Specifically, its performance is superior to previously published results on the GTZAN dataset. In semantic indexing, the proposed approach provides better results compared to MFCC and spectral features. Overall, the results are very promising and we plan to extend the work to other classification tasks which process audio data. Moreover, image classification provides wide variety of existing features which could all be applied to spectrograms as well, and which could possibly surpass the presented approach. Previous work in image classification suggests that no particular type of features is optimal in wide variety of tasks, and that combining different types of features improves results [36] (at the expense of higher computational cost). It is probable that combining multiple features inspired by image classification in music recognition would improve results even further. Furthermore, the features could be combined with traditional audio features.

References


Proceedings of CESC 2012: The 16th Central European Seminar on Computer Graphics (non-peer-reviewed)


