Segment Related Classification for Automatic Genre Detection

Alexander Wankhammer, Alois Sontacchi
Inst. of Electronic Music and Acoustics, Inffeldgasse 10/3, 8010 Graz, Austria, Email: {wankhammer,sontacchi}@iem.at

Introduction

Most genre classification systems are based on feature vectors which are either computed from the whole audio file or short arbitrary excerpts. However, to the best of our knowledge, structural information related to the musical form of songs has not been considered so far. To account for this musically relevant information, we propose to perform an additional segment detection stage prior to the final classification, allowing to focus the audio-analysis to distinct representative song segments.

Music Structure Discovery (*MSD*)

In the field of Music Information Retrieval (MIR), several approaches to solve the problem of musical structure discovery have been developed [1]. Typically, multiple features that have been found to be adequate descriptors for either one or several different aspects of human cognition of music, are initially derived from a short time spectral representation of the audio file. In many approaches, the feature extraction is followed by the calculation of a so called Self Distance Matrix (SDM) [1], allowing to visualize and intuitively analyze the temporal structure of the underlying song.

Our proposed MSD system is based on the approach initially presented in [2]. We use Normalized Cross Correlation (NCC), a template matching algorithm, to find repetitive parts inside a combined, bar level SDM. Exploiting the inherent symmetry of SDMs, the template matching algorithm compares different similarity profiles (vertical SDM slices) and makes the detection of repetitions widely independent of any distinct structures in the matrix. An example for this template matching stage is illustrated in Figure 1.

The detection matrix in Figure 1 manifests all segment pairs detected by the NCC algorithm. To get a measure for the quality of the detected pairs, the mean of all correlation values related to each pair is mapped according to the average correlation value of all calculated correlation vectors. Pairs with a resulting quality measure below zero are likely to represent misdetections and are not further processed. As pairs related to similar starts and stops represent different repetitions of the same song segment, they are grouped into subsets. These subsets will typically reveal overlapping parts and a final segment selection step is needed to find the optimal combination of segments.

Segment Selection and Combination

We have modified the final segment selection stage of the initial publication as follows. As the algorithm aims to exploit all relations of repetitions throughout the song, changes to one segment of a subset should have a direct impact on all subset members. Therefore, each subset is only represented by one prototype segment and its related starting positions. Changes to this prototype segment automatically modify all occurrences of the related subset.

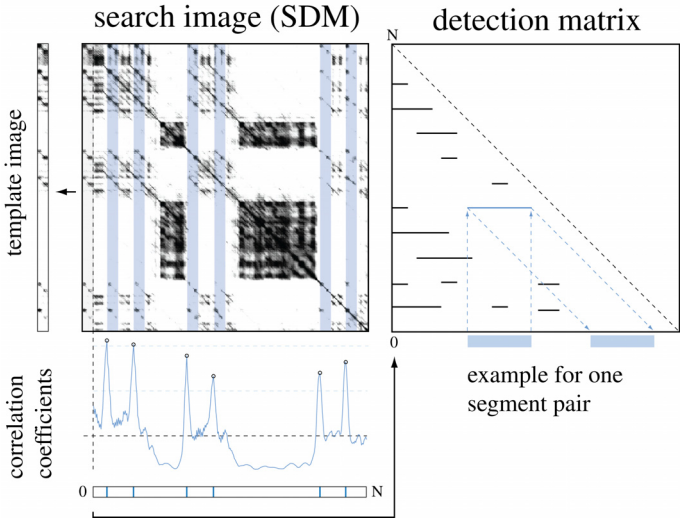

Figure 1: As indicated with light blue areas, maxima in the cross correlation vector show possible repetitions of the template image. By marking these maxima in binary vectors for all possible templates, a detection matrix can be created.

Starting with the subset related to the highest quality measure, all repetitions are iteratively compared. As depicted in Figure 2, two main operations (besides the definition of new segments, if no overlap is detected) can be performed for each pair of repetitions: splits and updates. Adjustable *skip in* and *skip out* values for each occurrence (see Figure 2, line 6) offer additional flexibility regarding different lengths of repetitions related to the same segment.

After an initial segmentation has been found, a post processing step helps to further investigate the segmentation by searching for repeated sequences in the preliminary segmentation (see Figure 2, line 7). An optimization function based on the length as well as the inner and inter similarity of all possible sub-sequences (found by correlating the related SDM segments) is maximized to find the optimal segmentation.

The algorithm has been evaluated on the well known data set *TUT Beatles*[1] consisting of 174 songs from 13 *Beatles* albums. The results of the evaluation are summarized in

[1] http://www.cs.tut.fi/sgn/arg/paulus/structure.html

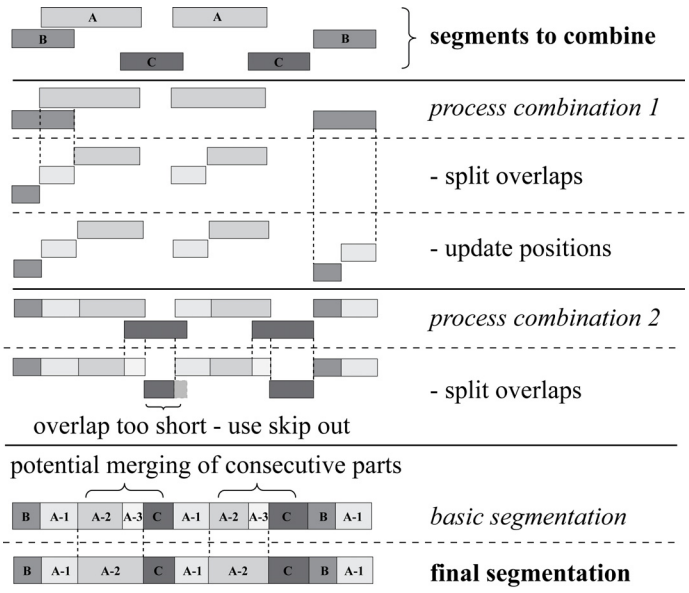

Figure 2: Exmeplary processing steps for three segments.

Table 1. The first row shows the quality of the automatically detected segment boundaries and was calculated using *classical F-measure* (allowing a deviation of $\pm 3s$ from reference boundaries). The second row shows the results related to the full segmentation performance of the algorithm including lengths and relations of all segments and was calculated using *pairwise F-measure*. To the best of our knowledge, the F-measures are the highest ever reported for the used data set and proof the stability of the developed algorithm.

Table 1: Segmentation Results

	R	P	**F**
Boundaries	69.8	62.8	**64.2**
Segments	72.3	67.5	**67.9**

Segment Related Genre Classification

To analyze the influence of incorporating knowledge about the internal structure of songs when trying to automatically find adequate genre labels, we compared the performance of a state of the art genre classification system [3][2] on different selections of test and training data: representative thumbnails *(Set 1)* and arbitrary song parts including the whole song *(Set 2)*. The songs in the test set are randomly selected from 8 genres (Alternative & Punk, Classical, Dance, Easy Listening, Hip-Hop, Jazz & Blues, R&B, Rock & Pop) of a publicly available dataset[3], using 40 songs per genre (320 songs in total).

The representative thumbnails of *Set 1* have been created by beat synchronously concatenating the most relevant repetitions of each song. Besides a thumbnail consisting of all complete repetitive segments, we additionally created shorter thumbnails (lengths: $5s, 10s, 30s$), to investigate the influence of focusing feature extraction to representative parts, when only a very small amount a data from a song is used. Although classical song preview generation was not the aim of this experiment, in particular the $30s$ thumbnails turned out to be excellent, compact previews of songs. *Set 2* contains the full songs, and a set of shorter song parts (lengths: $5s, 10s, 30s$), all starting $30s$ after the beginning of each song.

The tests on these datasets showed a significant difference in classification performance for the test sets $\leq 10s$ (marked with $*$ in Table 2). The results indicate that the classification performance of genre classification systems is not very sensitive to the selection of test and training data, as long as the "hit by chance" of a representative section is relatively large (e.g. $30s$). At the same time it could be shown that focusing feature extraction to representative parts may be an important strategy when aiming to reduce the area selected for feature extraction.

Table 2: Classification Accuracies(%) | *mean*(*std*)

Section	*Set* 1	*Set* 2
Full Song	59.16 (1.39)	60.13 (1.32)
30 sec	58.69 (1.35)	58.13 (1.44)
10 sec	56.91 (1.55)*	52.75 (1.28)
5 sec	51.06 (1.69)*	47.63 (1.15)

In a second test, we classified the individual parts of each song and evaluated the percentage of song parts labeled according to their reference genre. While the majority of parts related to well defined genres like *Classical* or *Hip-Hop* is labeled correctly, the individual results of vague genres like *Pop & Rock* or *Easy Listening* show no clear distribution of subparts with respect to their individual genres. For example, only 27% of the individual parts related to the category *Rock & Pop* are marked accoring to their reference genre. This observation may help to offer a better understanding of certain genre ambiguity problems when analyzing classification results and may particularily be of interest when trying to identify *non representative* songs in a large database.

References

[1] Paulus, J. and Müller, M. and Klapuri, A.: Audio-based music structure analysis. Proc. of the 11th Int. Soc. for Music Information Retrieval Conference (2010), 625-636

[2] Wankhammer, A. and Clarkson, I. Vaughan L. and Bradley, Andrew P.: Music Structure Discovery based on Normalized Cross Correlation. Proc. of the 13th Int. Conf. on Digital Audio Effects (2010), 88-493

[3] Seyerlehner, K. and Widmer, G. and Pohle, T.: Fusing block-level features for music similarity estimation. Proc. of the 13th Int. Conf. on Digital Audio Effects (2010), 225-232

[2]We would like to thank the authors for offering their prototype system to perform our experiments.

[3]http://www.seyerlehner.info