KDD-Based Approach to Musical Instrument Sound Recognition

Dominik Ślęzak, Piotr Synak, Alicja Wieczorkowska, and Jakub Wróblewski

Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland

Abstract. Automatic content extraction from multimedia files is a hot topic nowadays. Moving Picture Experts Group develops MPEG-7 standard, which aims to define a unified interface for multimedia content description, including audio data. Audio description in MPEG-7 comprises features that can be useful for any content-based search of sound files. In this paper, we investigate how to optimize sound representation in terms of musical instrument recognition purposes. We propose to trace trends in evolution of values of MPEG-7 descriptors in time, as well as their combinations. Described process is a typical example of KDD application, consisting of data preparation, feature extraction and decision model construction. Discussion of efficiency of applied classifiers illustrates capabilities of further progress in optimization of sound representation. We believe that further research in this area would provide background for automatic multimedia content description.

1 Introduction

Automatic extraction of multimedia information from files is recently of great interest. Usually multimedia data available for end users are labeled with some information (title, time, author, etc.), but in most cases it is insufficient for content-based searching. For instance, the user cannot find automatically all segments with his favorite tune played by the flute in the audio CD. To address the task of automatic content-based searching, descriptors need to be assigned at various levels to segments of multimedia files. Moving Picture Experts Group has recently elaborated MPEG-7 standard, named "Multimedia Content Description Interface" [8], that defines a universal mechanism for exchanging the descriptors. However, neither feature (descriptor) extraction nor searching algorithms are encompassed in MPEG-7. Therefore, automatic extraction of multimedia content, including musical information, should be a subject of study.

All descriptors used so far reflect specific features of sound, describing spectrum, time envelope, etc. In our paper, we propose a different approach: we suggest observation of feature changes in time and taking as new descriptors patterns in trends observed for particular features. We discuss how to achieve it by applying data preprocessing and mining tools developed within the theory of rough sets introduced in [13].

M.-S. Hacid et al. (Eds.): ISMIS 2002, LNAI 2366, pp. 28–36, 2002.

The analyzed database origins from audio CD's MUMS [12]. It consists of 667 samples of recordings, divided onto 18 classes, corresponding to musical instruments (flute, oboe, clarinet, violin, viola, cello, double bass, trumpet, trombone, French horn, tuba) and their articulation (vibrato, pizzicato, muted).

2 Sound Descriptors

Descriptors of musical instruments should allow to recognize instruments independently on pitch and articulation. Sound features included in MPEG-7 Audio are based on research performed so far in this area and they comprise technologies for musical instrument timbre description, sound recognition, and melody description. Audio description framework in MPEG-7 includes 17 temporal and spectral descriptors divided into the following groups (cf. [8]):

- basic: instantaneous waveform, power values
- basic spectral: log-frequency power spectrum, spectral centroid, spectral spread, spectral flatness
- signal parameters: fundamental frequency, harmonicity of signals
- timbral temporal: log attack time and temporal centroid
- timbral spectral: spectral centroid, harmonic spectral centroid, spectral deviation, spectral spread, spectral variation
- spectral basis representations: spectrum basis, spectrum projection

Apart from the features included in MPEG-7, the following descriptors have been used in the research ([6], [10], [17], [18]):

- duration of the attack, quasi-steady state and ending transient of the sound in proportion to the total time
- pitch of the sound
- contents of the selected groups of harmonics in spectrum, like even/odd harmonics Ev/Od

$$Ev = \frac{\sqrt{\sum_{k=1}^{M} A_{2k}^2}}{\sqrt{\sum_{n=1}^{N} A_n^2}} \qquad Od = \frac{\sqrt{\sum_{k=2}^{L} A_{2k-1}^2}}{\sqrt{\sum_{n=1}^{N} A_n^2}} \tag{1}$$

  and lower/middle/higher harmonics $Tr_1/Tr_2/Tr_3$ (Tristimulus parameters [14], used in various versions)

$$Tr_1 = \frac{A_1^2}{\sum_{n=1}^{N} A_n^2} \quad Tr_2 = \frac{\sum_{n=2,3,4} A_n^2}{\sum_{n=1}^{N} A_n^2} \quad Tr_3 = \frac{\sum_{n=5}^{N} A_n^2}{\sum_{n=1}^{N} A_n^2} \tag{2}$$

  where A_n denotes the amplitude of the n^{th} harmonic, N – the number of harmonics available in spectrum, $M = \lfloor N/2 \rfloor$ and $L = \lfloor N/2 + 1 \rfloor$
- vibrato amplitude
- statistical properties of sound spectrum, including average amplitude and frequency deviations, average spectrum, standard deviations, autocorrelation and cross-correlation functions ([2])
- descriptors based on wavelet analysis and numerous other features