Audio Classification and Retrieval Using Wavelets and Gaussian Mixture Models

Ching-Hua Chuan, School of Computing, University of North Florida, Jacksonville, FL, USA

ABSTRACT

This paper presents an audio classification and retrieval system using wavelets for extracting low-level acoustic features. The author performed multiple-level decomposition using discrete wavelet transform to extract acoustic features from audio recordings at different scales and times. The extracted features are then translated into a compact vector representation. Gaussian mixture models with expectation maximization algorithm are used to build models for audio classes and individual audio examples. The system is evaluated using three audio classification tasks: speech/music, male/female speech, and music genre. They also show how wavelets and Gaussian mixture models are used for class-based audio retrieval in two approaches: indexing using only wavelets versus indexing by Gaussian components. By evaluating the system through 10-fold cross-validation, the author shows the promising capability of wavelets and Gaussian mixture models for audio classification and retrieval. They also compare how parameters including frame size, wavelet level, Gaussian components, and sampling size affect performance in Gaussian models.

Keywords: Audio Classification, Compact Vector Representation, Gaussian Mixture Models, Retrieval, Wavelets

INTRODUCTION

Content-based analysis for audio classification and retrieval has been studied for more than a decade, but the issues of efficiency, scalability, and accuracy require more in-depth research as the amount of multimedia on the Web increases. Although various approaches have been proposed for such tasks, these approaches usually lack information about the robustness of the algorithm, particularly if the algorithm is applied to a different dataset or for a different classification/retrieval task. Classification and retrieval tasks can be done in various ways. For example, we can classify a song based on its musical genre or its artist. Similarly, we can retrieve a song that is an exact match to a request or that simply sounds similar. It is usually difficult to predict the performance of a classification/retrieval approach based on its result. The problem is further complicated if the approach involves parameter turning, such as deciding frame size, which is unavoidable for analysis of audio content. It is also important to discuss the reuse of a technique, or processing
the data once and using it for multiple purposes, as opposed to implementing different algorithms for individual tasks and subtasks.

In this paper, we focus on wavelet transforms and Gaussian mixture models (GMMs), demonstrating the manner in which these techniques are applied in audio classification and retrieval. We examine performance factors in details by systematically conducting a series of experiments. Generally speaking, an audio classification and retrieval system contains two major processing steps. In the first step, the system extracts low-level features from audio signals. Most of the existing systems use methods based on Fourier transform or mel-frequency cepstral coefficients for low-level spectral analysis. In the second step, a machine-learning component captures the characteristics of an audio recording or a type of audio. Commonly used learners include decision trees, neural networks, support vector machines, GMMs, and hidden Markov models. In this paper, we use wavelet transform for low-level feature extraction and then apply GMMs as the learner for audio classification and retrieval.

Compared with Fourier transform, wavelet transform is a relatively new tool for signal processing, and it is not as widely adopted. However, it has been shown mathematically that wavelet transform has several advantages over Fourier transform for analyzing signals. With this study, we demonstrate the capability of wavelet transform for audio classification and retrieval. We apply wavelet transform to extract low-level features from audio signals and build a feature vector as a compact representation. We perform a multiresolution wavelet analysis and experiment with different levels of wavelet decomposition of signals. The compact multi-resolution feature vector is then used and tested in various classification and retrieval tasks.

GMMs are used as learners for audio classification and retrieval in this paper. The advantages of using GMMs include computational efficiency and the flexibility of modeling arbitrary probability densities. In the classification task, several Gaussian components are used to learn the probability densities of a type of audio in the training phase. In the test phase, the learned GMMs for individual audio types are then used to calculate the likelihood that a test audio presents an instance of a particular type. In the retrieval task, the GMMs are built for individual audio examples and the learned parameters in GMMs are stored to index the examples. Given an audio input as the query, its learned GMMs are compared with the GMMs of the examples in the dataset using Monte Carlo sampling. A symmetric and normalized distance measure is adapted for comparing GMMs using similar calculations as in the classification task. A simple $k$-nearest neighbor search is performed to retrieve examples that report the shortest distances.

To evaluate the system, we design three audio classification and retrieval tasks with different levels of difficulty. The first task involves a binary classification in which the system attempts to separate an audio recording of speech from a musical one. In the second task, which consists of subclasses of the first task, the system attempts to identify the speaker’s gender. In the third task, the most challenging one, audio recordings of four different musical genres (classical, pop, jazz, and electronic) are used for classification and retrieval. For each type of audio, we collect 100 examples and conduct 10-fold cross validation for evaluation. The results of the classification task are displayed as the average correct classification rate for each audio type. The results of the retrieval task are displayed as the precision in the top-$k$ retrievals for $k = 1, 3, 5, 10$. We systematically analyze the effect of the following parameters on the classification and retrieval outcomes: frame size = 0.1, 0.5, 1, 2, and 3 seconds; the level of multi-resolution wavelet analysis = 4, 5, 6, 7, and 8; the number of Gaussian components = 2, 4, 6, and 8; and the number of samples in Monte Carlo sampling = 1, 5, 10, 50, 100, 200, and 300.
18 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:

www.igi-global.com/article/audio-classification-and-retrieval-using-wavelets-and-gaussian-mixture-models/78745?camid=4v1


www.igi-global.com/e-resources/library-recommendation/?id=2

Related Content

Design and Performance Evaluation of Smart Job First Multilevel Feedback Queue (SJFMLFQ) Scheduling Algorithm with Dynamic Smart Time Quantum

Amit Kumar Gupta, Narendra Singh Yadav and Dinesh Goyal (2017). International Journal of Multimedia Data Engineering and Management (pp. 50-64).

War and Play: Insensitivity and Humanity in the Realm of Pushbutton Warfare
www.igi-global.com/chapter/war-play-insensitivity-humanity-realm/50733?camid=4v1a

Interface Design, Emotions, and Multimedia Learning
www.igi-global.com/chapter/interface-design-emotions-multimedia-learning/27153?camid=4v1a

Location-Aware Caching for Semantic-Based Image Queries in Mobile AD HOC Networks
www.igi-global.com/article/location-aware-caching-semantic-based/64629?camid=4v1a