Audio classification method based on machine learning

Feng Rong
(Liaoning Jianzhu Vocational College; Liaoyang,Liaoning,111000,China)
279049844@qq.com

***Abstract*—Audio classification has very large theoretical and practical values in both pattern recognition and artificial intelligence. In this paper, we propose a novel audio classification method based on machine learning technique. Firstly, we illustrate the hierarchical structure of audio data, which is made up of four layers: 1) Audio frame, 2) Audio clip, 3) Audio shot, and 4) Audio high level semantic unit. Secondly, three types of audio data feature are extracted to construct feature vector, including 1) Short time energy, 2) Zero crossing rate and 3) Mel-Frequency cepstral coefficients. Thirdly, we discuss how to classify audio data using the SVM classifier with Gaussian kernel. Finally, experimental results demonstrate that the proposed method is able to achieve higher audio classification accuracy.**

Keywords- Audio classification, Machine learning, SVM, Audio frame, Audio shot.

I. INTRODUCTION

The development of multimedia technology and Internet has brought huge multimedia information to the people, and further resulted in the production of ultra large scale multimedia information database [1][2]. It is very difficult to make the description and retrieval of multimedia information, which needs a kind of effective retrieval method for multimedia. How to effectively help people to quickly and accurately find the needed multimedia information has become the core issue in the multimedia information retrieval [3]. Content based information retrieval is proposed under such a background, it is a kind of new retrieval technology [4].

Audio information refers to one of the most important sources of human perception. It is urgent to organize mass audio files in terms of the semantic description with the rapid increasing of audio and video files [5]. Due to the complexity of audio retrieval, it is more difficult than the image and video retrieval. As the original audio data is made up of non-semantic and non-structured binary stream, it lacks of semantic description and structured organization [6][7]. Moreover, the audio data has the characteristics of complex structure, massive data, and high requirement of data processing [8][9]. Therefore, it brings great difficulties to deeply process and analyze audio information, and the audio retrieval and content filtering applications are hard to design. It is of great importance to extract structured information and semantic contents from audio data [10][11][12]. To solve this task, the audio classification should be tackled well.

The rest of the paper is organized as follows. We introduce description of audio data in Section 2. In Section 3, we propose a novel audio classification algorithm based on SVM. To prove the effectiveness of our proposed algorithm, experimental results are conducted in section 4. Section 5 concludes the whole paper.

II. DESCRIPTION OF AUDIO DATA

In this section, we analyze main features of audio data, and the hierarchical structure of audio data is illustrated in Fig. 1.

As is shown in Fig. 1, definitions of audio structural units with different time granularity is proposed, and the hierarchical structure of audio data contains 1) Audio high level semantic unit, 2) Audio shot, 3) Audio clip and 4) Audio frame.

(1) Audio frame: Audio is a non-stationary random process, and its characteristics change with time varying. However, its change is very slow. Therefore, the audio signal can be divided into a number of short periods of processing. These short segments are generally 20-30ms (called as the audio frame) and audio frame is the smallest units in the audio processing.

(2) Audio clip: As the audio frame time granularity is too small and it is difficult to extract meaningful semantic contents, we should define a new audio unit with larger time granularity (named as Audio clip). Audio clip is made up of a number of frames, of which the length of time is fixed. The characteristics of the audio segment are calculated on the basis of audio frame.

(3) Audio shot: This concept is generated from the video footage. As audio segment are too short, it is not suitable to be used for semantic content analysis. An audio structural unit containing the same audio class is defined as an audio shot. The audio shot is composed of several audio segments within a same class

(4) Audio high level semantic unit: It refers to an audio structural unit with rich semantic contents which are formed by the different combinations of the audio shots.

978-1-5090-6061-0/17 $31.00 © 2017 IEEE
DOI 10.1109/ICITBS.2016.98

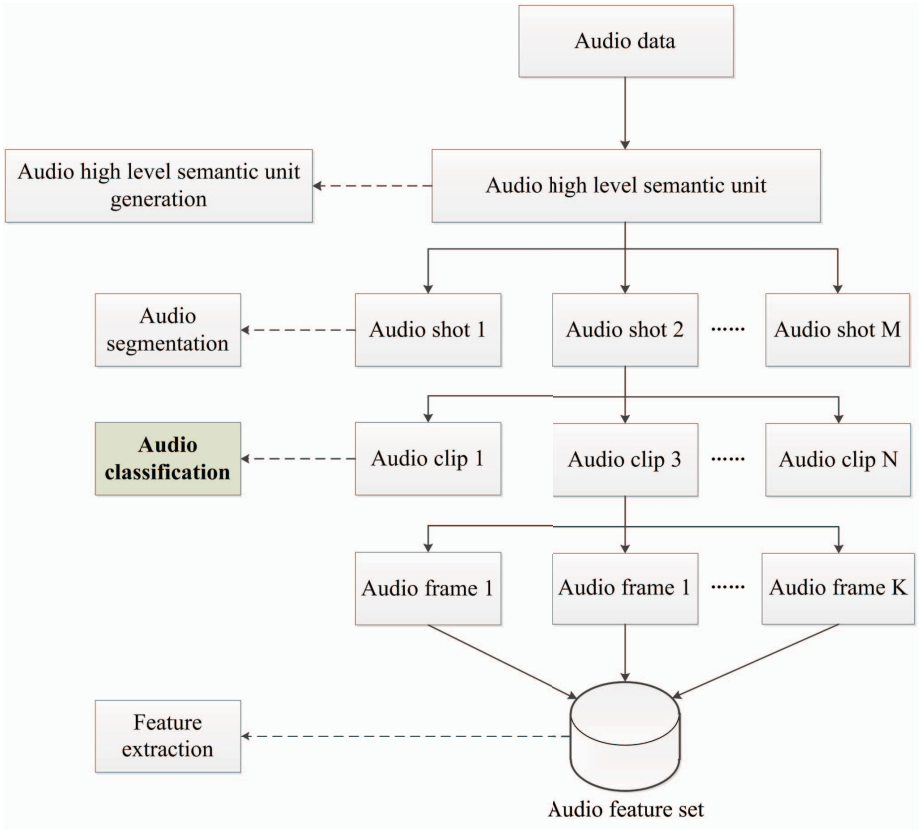

Fig. 1 Hierarchical structure of audio data

III. THE PROPOSED AUDIO CLASSIFICATION ALGORITHM BASED ON SVM

In this section, we discuss how to classify audio data by the SVM classifier. The task of audio data classification is implemented based on audio data feature extraction. In our work, the following three types of audio features are utilized.

Feature 1: Short time energy (denoted as STE) is a discrete signal $x[n]$ is defined as the sum of squares of the signal samples N.

$$STE = \frac{1}{N} \cdot \sum_{n=0}^{N-1} x^2[n] \tag{1}$$

Feature 2: Zero crossing rate (denoted as ZCR). For a given frame, ZCR is defined as the number of times the audio from positive to negative in the given frame as follows.

$$\mathrm{ZCR} = \frac{1}{2} \sum_{n=0}^{N-1} \left| \mathrm{sgn}\left(x[n]\right) - \mathrm{sgn}\left(x[n-1]\right) \right| \tag{2}$$

Feature 3: Mel-Frequency cepstral coefficients (denoted as MFCC), it is defined based on discrete cosine transform as follows.

$$c_n = \sqrt{\frac{2}{K}} \cdot \sum_{k=1}^{K} \left(\log S_k\right) \cdot \cos\left(\frac{1}{K} \cdot \left(n\left(k-0.5\right)\pi\right)\right) \tag{3}$$

where K denotes the number of band pass filters.

Assume that training samples are defined as (x_i, y_i), and $i \in \{1,2,\cdots,l\}$, $i \in \{1,2,\cdots,l\}$, $x_i \in R^n$, $y_i \in \{1,\text{-}1\}$ are satisfied. SVM classifier is designed to tackle an optimization problem as follows.

$$\min_{w,b,\xi} \left(\frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \right) \tag{4}$$

s.t.

$$y_i\left(w^T z_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0, i \in \{1,2,\cdots,l\} \tag{5}$$

where x_i is mapped by function $\varphi(\ )$, and C refers to a penalty parameter.

Afterwards, the audio data classification problem can be solved as follows.

$$\operatorname{sgn}\left(w^T\phi(x)+b\right)=\operatorname{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K(x_i,x)+b\right) \quad (6)$$

Next, in order to enhance performance of SVM algorithm, we utilize Gaussian kernel in our proposed classifier, which is defined as follows.

$$K(x_i,x)=\exp\left(-\frac{1}{\sigma^2}\|x-x_i\|^2\right) \quad (7)$$

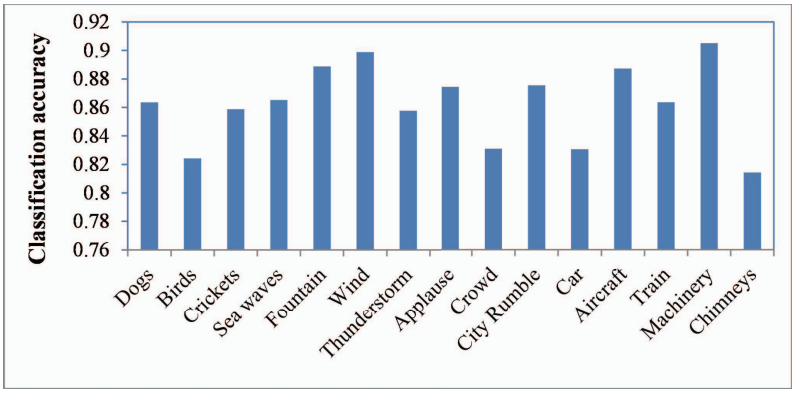

Fig. 2 Audio classification result using the GS database

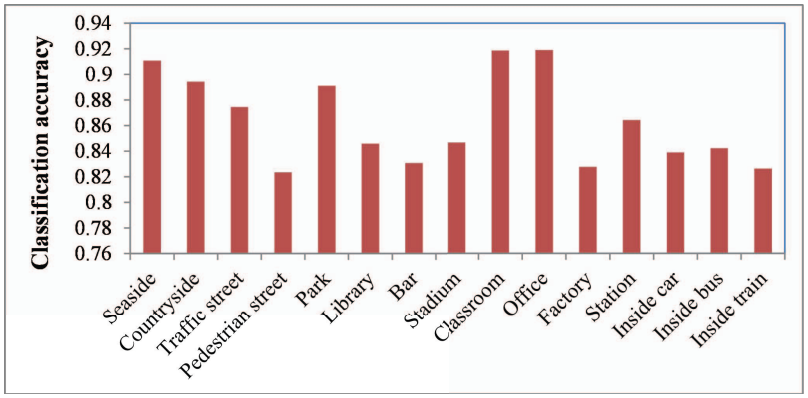

Fig. 3 Audio classification result using the AS database

Combining experimental results in Fig. 2 and Fig. 3 together, it can be observed that our proposed audio classification can achieve higher classification accuracy. Furthermore, the average classification accuracy of the database GS and AS are 0.876 and 0.863 respectively.

IV. Experiment

In this section, we choose two standard audio databases to make performance evaluation, that is, 1） General Sounds (GS) (e.g., applause, birds, cars, etc.) and 2） Audio Scenes (AS) (e.g., park, bar, station, etc.). Particularly, the audio samples of the above two audio databases are collected from various origins to guarantee data variability. Furthermore, each audio database contains nearly 4 hours of audio, and each class is made up of audio samples with the number ranged between 150 and 300.

V. Conclusion

This paper aims to design an efficient audio classification algorithm based on support vector machine method. The hierarchical structure of audio data contains four layers: 1) Audio frame, 2) Audio clip, 3) Audio shot, and 4) Audio high level semantic unit. Next, three types of audio data feature are extracted to establish feature vectors.

Afterwards, we propose a novel audio classification algorithm based on SVM with Gaussian kernel. In the future, we will try to test the performance of our algorithm in other audio database.

REFERENCE

[1] X. K. Yang, L. He, D. Qu, W. Q. Zhang and M. T. Johnson, Semi-supervised feature selection for audio classification based on constraint compensated Laplacian score, Eurasip Journal on Audio Speech and Music Processing, 2016, 1-10

[2] B. K. Baniya and J. Lee, Importance of audio feature reduction in automatic music genre classification, Multimedia Tools and Applications, 2016, 75(6): 3013-3026

[3] S. Zahid, F. Hussain, M. Rashid, M. H. Yousaf and H. A. Habib, Optimized Audio Classification and Segmentation Algorithm by Using Ensemble Methods, Mathematical Problems in Engineering, 2015,

[4] Z. Q. Shi, J. Q. Han and T. R. Zheng, Soft Margin Based Low-Rank Audio Signal Classification, Neural Processing Letters, 2015, 42(2): 291-299

[5] A. Rakotomamonjy and G. Gasso, Histogram of Gradients of Time-Frequency Representations for Audio Scene Classification, Ieee-Acm Transactions on Audio Speech and Language Processing, 2015, 23(1): 142-153

[6] M. Benatan and K. Ng, Cross-covariance-based features for speech classification in film audio, Journal of Visual Languages and Computing, 2015, 31(215-221

[7] S. Fagerlund and U. K. Laine, Classification of audio events using permutation transformation, Applied Acoustics, 2014, 83(57-63

[8] A. Chen and M. A. Hasegawa-Johnson, Mixed Stereo Audio Classification Using a Stereo-Input Mixed-to-Panned Level Feature, Ieee-Acm Transactions on Audio Speech and Language Processing, 2014, 22(12): 2025-2033

[9] S. Zubair, F. Yan and W. W. Wang, Dictionary learning based sparse coefficients for audio classification with max and average pooling, Digital Signal Processing, 2013, 23(3): 960-970

[10] L. Xing, Q. Ma and M. Zhu, Tensor semantic model for an audio classification system, Science China-Information Sciences, 2013, 56(6):

[11] Z. Q. Shi, J. Q. Han and T. R. Zheng, Audio Classification with Low-Rank Matrix Representation Features, Acm Transactions on Intelligent Systems and Technology, 2013, 5(1):

[12] M. A. Haque and J. M. Kim, An analysis of content-based classification of audio signals using a fuzzy c-means algorithm, Multimedia Tools and Applications, 2013, 63(1): 77-92