WAVELET-BASED FEATURE EXTRACTION FOR MUSICAL GENRE CLASSIFICATION USING SUPPORT VECTOR MACHINES

[1]Liming Chen, [2]Eugene Bovbel[*], [2]Maxim Dashouk

[1]Dept. Mathématiques Informatique, Ecole Centrale de Lyon, France
[2]Belarussian State University
Department of radiophysics and electronics, Minsk, Belarus
bovbel@bsu.by

Abstract

Musical genre classification task falls into two major stages: feature extraction and classification. The latter implies a choice of a variety of machine leaning methods, as support vector machines, neural networks, etc. However, the former stage provides much more creativity in development of musical genre classification system and it plays crucial part in performance of the system as a whole. In this paper we present initial study of wavelet-based feature extraction in the task of musical genre classification. A new type of feature vector, based on continuous wavelet transform of input audio data is proposed. The method of feature extraction was tested using support vector machine as a classifier. The results of our experimental study are shown.

1. Introduction

Significant progress in network, data storage and retrieval technologies resulted in fact that there is a huge amount of musical recordings data available for users all over the world. These places are first of all commercial musical databases and popular commercial "mp3 download" sites in the World Wide Web. For usability's sake, these musical collections are typically sorted into different musical genres. So far, such an operation is performed manually.

Also, interest in algorithms of musical genre classification emerges from their possible deployment in multimedia indexing systems. Very often audio data encapsulated much useful information about content of some video stream and thus must be used to describe a video scene along with visual images. In numerous cases, it becomes critical to determine genre of background music for thorough description.

These two examples of possible applications of musical genre classification explain why it is desirable to have an automatic failsafe system of musical genre classification. Fortunately, there have been several algorithms of genre classification proposed.

Burred *et al* [1] developed the system of automatic musical genres classification. The signals are recognized as speech, background noise and one of 13 musical genres. The authors evaluated audio features for their suitability in such a classification task, including well-known physical and perceptual features, audio descriptors defined in the MPEG-7 standard, and features, such as timbre, rhythm, etc. A 3-component Gaussian Mixture Model was used as classifier.

Shao *et al* in [2] proposed an unsupervised clustering method based on a given measure of similarity provided by Hidden Markov Models. The music dataset for each genre contained 50 music pieces. The genres are Pop, Country, Jazz and Classic.

Xu *et al* [3] used Support Vector Machines for classification and MFCC, beat spectrum, LPC-derived cepstrum, zero-crossing rate as features to classify music into four genres: rock, pop, jazz and classic.

Methodology of automatic musical genre classification described by Tzanetakis *et al* in [4] represents an up-to-date system, based on advances feature extraction. Their proposed features are timbral texture features and rhythmic content features. In fact, timbral texture features include several features, which were used in earlier works on speech recognition: Mel-frequency cepstral coefficients (MFCC) and time-domain zero crossings [5]. In addition to these classical audio processing features more sophisticated features were added, such as spectral rolloff, spectral centroid and spectral flux. Also the authors point out such additional "high-level" features as pitch content features and beat histogram features. This comprehensive set of features was accompanied by Gaussian Mixture Model classifiers. As a result, this system was reported to have 39% error rate for nonreal-time and 56% error rate for real time classification of ten genres. This performance was concluded to be

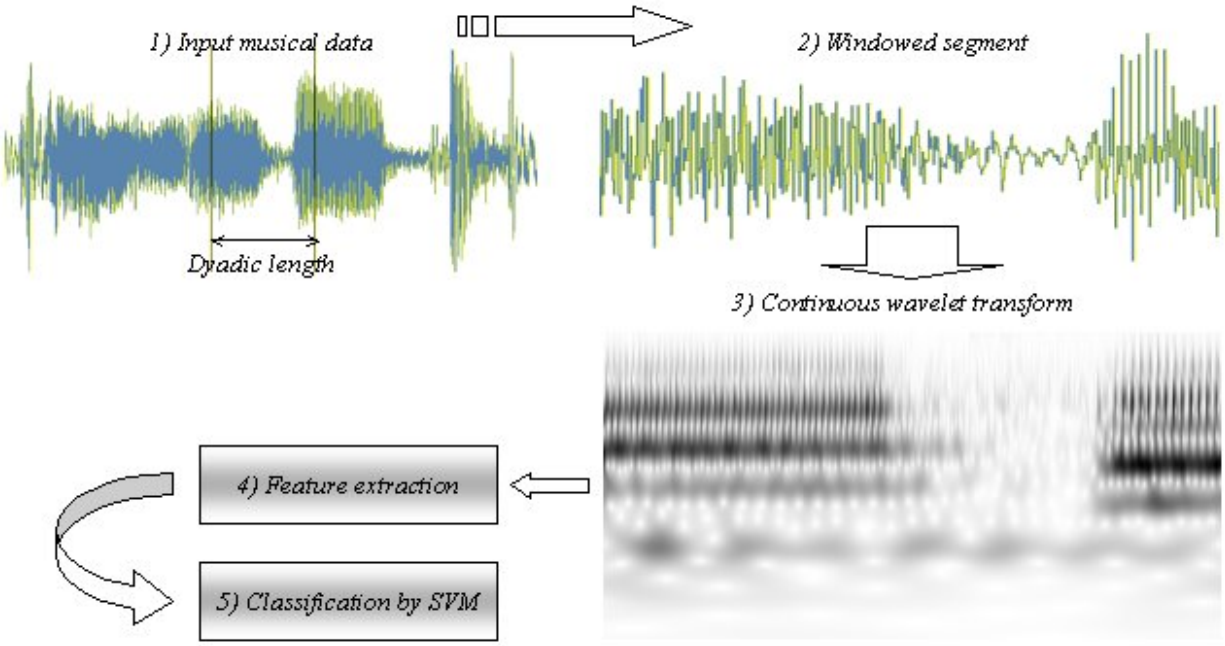

Fig. 1 Block diagram of musical genre classification method

comparable with performance of human (manual) musical genre classification.

In this paper, we propose features based on continuous wavelet transform of musical signal data.

2. Wavelet-based features

The idea to build feature vector on wavelets for audio classification was previously reported by Tzanetakis *et al* in [4] and Li *et al* [6]. The authors used discrete wavelet transform (DWT) coefficients for their method of feature extraction for content-based audio classification. Unlike DWT, continuous wavelet transform allows much more flexibility due to its arbitrary time-scale resolution. The primary aim of our paper is to show that CWT-based features can serve as one of the basic features for automatic musical genre classification.

2.1. Fast algorithm of continuous wavelet transform

Continuous wavelet transform *Wf(a,b)* of some signal *f(t)* can be interpreted as a convolution product:

$$Wf(a,b) = \int_{-\infty}^{+\infty} f(t)\overline{\psi}\left(\frac{b-t}{a}\right)dt,$$

were $\overline{\psi}(z) = \psi(-z)$ and $\psi(z)$ is a basic wavelet function, but modified by two parameters *a* (scale or "frequency" parameter) and *b* (time shift). Once *f(t)* is known one can easily obtain a time-frequency representation *Wf(a,b)* of the given signal.

In this paper, we used an algorithm of CWT presented at [7]. It uses Gabor function as a wavelet function. This fact makes it the only closed-form version of CWT and, consequently, very practical for its use in computations. The description of this version of CWT goes beyond the scope of the paper and further details can be obtained in [7], but the authors would like to mention some details relevant to our study.

First, we used a dyadic version of the algorithm due to its enhanced efficiency comparing to nondyadic version. Thus, all the experimental audio data has dyadic length in our study.

In our experiment, the scale parameter a changes as $a = 2^s 2^{j/J}$, where s is a current octave, J denotes a number of voices per octave; j is specifies current voice such that is $0 \le j < J$. In our study J = 8.

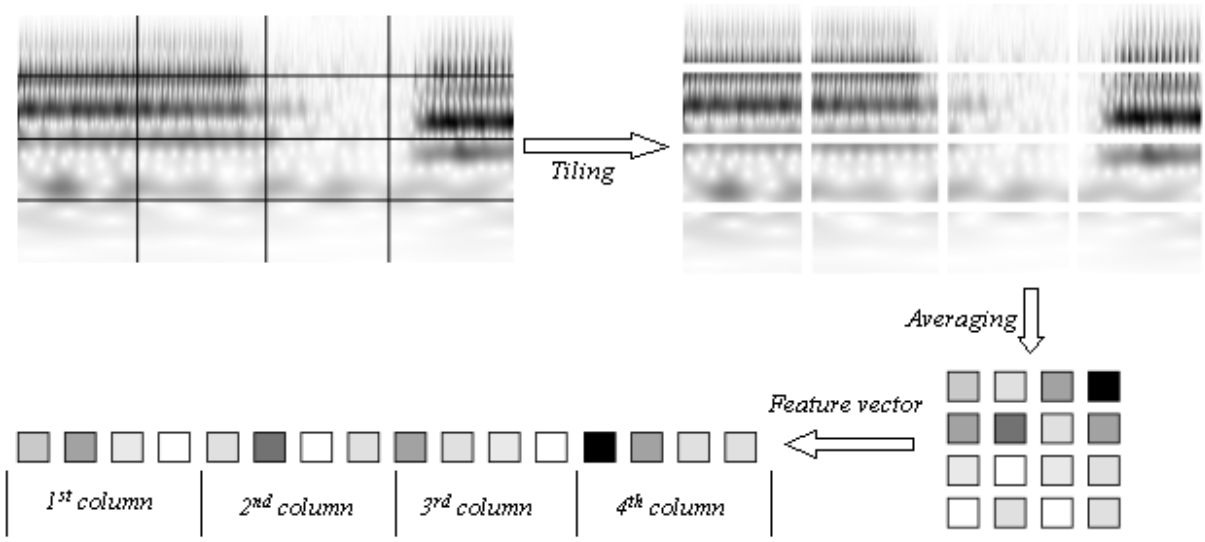

Fig.2 First type of feature vector

2.2. Feature extraction

Block diagram of our classification method is depicted on **Fig. 1**. A window of constant dyadic length is applied to input audio data picking out an audio segment. Then continuous wavelet transform is applied to this audio segment. The result of this transform is time-frequency representation of the given signal. Now wavelet image of the segment is a source for further feature extraction. In fact, there are plenty of ways to extract information from CWT representation of a signal. However, at the beginning of our study of CWT-based features we decided to limit ourselves by two simplest ways of feature extraction from CWT of a signal.

Either way of feature extraction uses reduction of wavelet information by averaging of neighboring wavelet coefficients on time-frequency plane. That is, the whole time-frequency plane is cut into subbands along the scale axis and subsegments along the time axis. The width of subbands and subsegments is equal and this results in uniform tiling of CWT time-frequency plane.

The first type of feature vector (**Fig.2**) simply averages all the coefficients in every tile resulting in one mean value for every tile. Then these mean values form a feature vector as it is shown on **Fig.2**.

The second type of feature vector slightly differs from the previous one. In this case, the first column of tiles is averaged as in type I. But beginning from the second column of tiles, each mean value results from averaging over the current tile as well as over all the previous tiles of the same subband as illustrated on **Fig.3**.

3. Support Vector Machines

For classification, a statistical learning algorithm called support vector machine (SVM) is used. SVMs which were proposed by Vapnik [8], have become an acknowledged classification method in the task of musical genre recognition. Their usage in this task was already justified by works of Li et al. [6] and Xu et al. [3], were SVMs outperformed other commonly used classification methods as Gaussian Mixture models, k-Nearest neihbour classifier, hidden Markov models (see also **Table 2**).

We will consider the basic theory of SVM in this section.

Given a set of training vectors belonging to two separate classes, $(\mathbf{x}_1, y_1),...,(\mathbf{x}_l, y_l)$, where $\mathbf{x}_i \in R^n$ and $y_i \in \{-1,...,1\}$, one wants to find a hyperplane $\mathbf{wx} + b = 0$ to separate the data. In fact, there are many possible hyperplanes, but there is only one that maximizes the margin (the distance between the hyperplane and the nearest data point of each class).

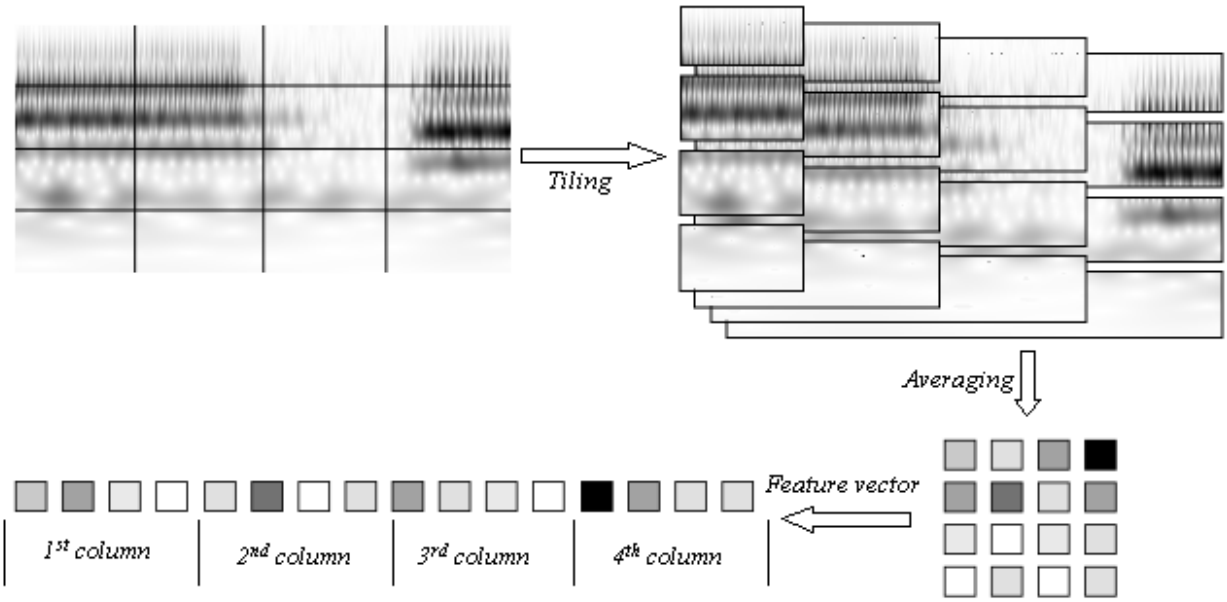

Fig. 3 The second type of feature vector

The solution to the optimization problem of SVM is given by the saddle point of the Lagrange functional

$$L(\mathbf{w},b,\alpha)=\frac{1}{2}\|\mathbf{w}\|^2-\sum_{i=1}^{l}\alpha_i\{y_i[(\mathbf{w}\cdot\mathbf{x}_i)+b]-1\}\ (1)$$

where α_i are the Lagrange multipliers. Classical Lagrangian duality enables the primal problem (1) to be transformed to its dual problem, which is easier to solve. The solution is given by

$$\overline{\mathbf{w}}=\sum_{i=1}^{l}\overline{\alpha}_i y_i\mathbf{x}_i,\ \overline{b}=-\frac{1}{2}\overline{\mathbf{w}}\cdot[\mathbf{x}_r+\mathbf{x}_s]\ (2)$$

where $\mathbf{x}_r$ and $\mathbf{x}_s$ are any two support vectors with $\overline{\alpha}_r$, $\overline{\alpha}_s>0,\ y_r=1,\ y_s=-1$.

To solve the nonseparable problem slack variables $\xi_i\geq 0$ and a penalty function, $F(\xi)=\sum_{i=1}^{l}\xi_i$, where the ξ are a measure of the misclassification error. The solution is identical to the separable case except for a modification of the Lagrange multipliers as $0\leq\alpha_i\leq C,\ i=1,...l$. The choice of C is not strict in practice.

The SVM can realize nonlinear discrimination by kernel mapping [8]. In **Fig. 4**, the samples in the input space can not be separated by any linear hyperplane, but can be linearly separated in the nonlinear mapped feature space.

There are three typical kernel functions for the nonlinear mapping [8]:

1) *Polynomial function:*

$$K(\mathbf{x},\mathbf{y})=((\mathbf{x}\times\mathbf{y})+1)^d,$$

where parameter d is the degree of the polynomial;

2) *Gaussian radial basis function:*

$$K(\mathbf{x},\mathbf{y})=\exp(-((\mathbf{x}-\mathbf{y})^2/2\sigma^2)),$$

where parameter σ is the width of the Gaussian function;

3) *Multilayer perception:*

$$K(\mathbf{x},\mathbf{y})=\tanh(s\cdot(\mathbf{x}\times\mathbf{y})-t),$$

where s and t are scale and offset accordingly.

4) *Exponential radial basis function (ERBF) :*

$$K(\mathbf{x},\mathbf{y})=\exp\left(-\frac{|\mathbf{x}-\mathbf{y}|}{2\sigma^2}\right)\ (3).$$

We used only ERBF in all our experiments because it showed best performance in our task.

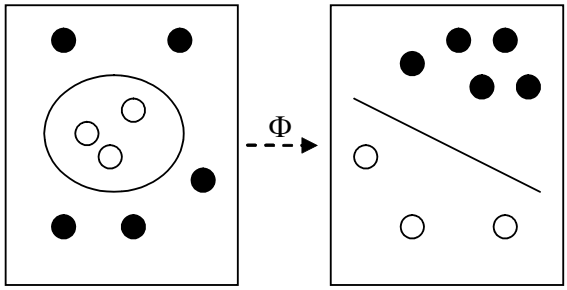

Fig. 4. Feature space is related to input space via a nonlinear map Φ, causing the decision surface to be nonlinear in the input space. By using a nonlinear kernel function, there is no need to do mapping explicitly.

	Subsegments	FV type	Subbands	Err. Rate
classic/jazz	4	0	64	19.7
	2	0	64	21.1
	4	1	64	17.7
	2	1	64	21.6
	1	0	32	21
	1	0	16	22.3
classic/pop	4	0	64	19.4
	2	0	64	17.6
	4	1	64	17.2
	2	1	64	17.6
	1	0	32	17.1
	1	0	16	16.3
jazz/pop	4	0	64	26.9
	2	0	64	26.4
	4	1	64	28.8
	2	1	64	24.7
	1	0	32	26.2
	1	0	16	26.7

	Subsegments	FV type	Subbands	Err. Rate
classic/rock	4	0	64	7.4
	2	0	64	7.1
	4	1	64	6.4
	2	1	64	6.4
	1	0	32	6.7
	1	0	16	6.7
rock/jazz	4	0	64	9.5
	2	0	64	10.5
	4	1	64	11.8
	2	1	64	13.6
	1	0	32	12.1
	1	0	16	14.1
rock/pop	4	0	64	17.1
	2	0	64	16.3
	4	1	64	17.8
	2	1	64	17.1
	1	0	32	14.4
	1	0	16	19

Table 1. Error rates of classification between 6 genre combinations. The window length is constant and equal to 4096.

For a given kernel function, the classifier is given by

$$f(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{l} \overline{\alpha}_i y_i K(\mathbf{x}_i, \mathbf{x}) + \overline{b} \right) \quad (4).$$

4. Experimental results

In our study we were not focused in spanning as many genres as possible and limited ourselves by four general genres: classic, jazz, pop, and rock.

4.1. Database

Our experimental database consisted of 400 musical records in 16 bit mono PCM format, each 30 seconds long, digitized at 22050 samples per second. Among these 400 records there were not a pair of records as fragments of the same composition. 100 records represented each genre. The database was recorded from different sources: compact disks, mp3 databases and radio.

4.2. Experiment

We used half of the base for training and the other half for testing, namely 50 records from each genre. The kernel function of SVM was a radial-basis kernel function. Five-fold cross-validation was used in a procedure of grid search for optimal parameters of SVM.

We tested our method of feature extraction for each type of feature vector. In turn, for every type of feature vector we independently varied window length, number of subbands and subsegments. We tested all 6 genre combinations for two-class classification by SVM. The results are presented on **Table 1**.

5. Conclusion and future work

Table 2 represents summary of some published methods of musical genre recognition. As comparison with our results shows, the presented method performs with similar accuracy. But, obviously, correct comparison can be made only using the same database. Nevertheless, our features based on continuous wavelet transform have potential for future usage in musical genre classification task. Our methodology of wavelet feature extraction seems to work well with some genre pairs as for example *classic/rock*. Although it may be possible to try to build a complete multiclass classification system with an hierarchy of support vector machines, we suppose that further search for more sophisticated feature extraction from continuous wavelet transform coefficients must be performed. Also, some optimal configuration of feature extraction must be found, namely, an optimal window length, number of subbands and subsegments, constant for every SVM in a multiclass classification system. And evidently four-genre classification is not enough for real-world applications.

Article	Number of genres	Number of original pieces in database	Features	Classifiers	Error rate
Soltau *et al*, [9]	4	360	Cepstrum	HMM, ETM-NN	21% 14%
Jiang *et al*, [10]	5	1500	Spectral contrast	GMM	18%
Tzanetakis *et al*, [4]	10	1000	Timbral texture, beat histogram, pitch content	GMM	39%
Burred *et al*, [1]	13	850	Timbral, Beat histogram, MPEG-7 LLD, other	GMM	48%
Li *et al*, [6]	10	1000	Daubechies wavelet coefficient histograms	GMM k-NN LDA SVM	36% 38% 29% 21%
Xu *et al*, [3]	4	100	MFCC, LPC-derived cepstrum, spectrum power, ZCR, beat spectrum	GMM HMM k-NN SVM	12% 12% 21% 7%

Table 2. Summary of genre recognition systems' performance [11] (k-NN – Explicit time modeling with neural network; k-NN – k-Nearest neighbour classifier, HMM – hidden Markov model, GMM – Gaussian mixture model, LDA – Linear discriminant analysis, ZCR – Zero crossing rate; MFCC – Mel frequency cepstral coefficients, LPC – Linear predictive coding).

Thus, our genre hierarchy should be greatly expanded in both width and depth in our future research.

6. References

[1] J. Burred and A. Lerch, "A Hierarchical approach to automatic musical genre classification," pp. 308-311 *Int. Conf. on Digital Audio Effects (DAFx-03), London, UK, September 2003*.

[2] X. Shao, C. Xu, M. Kankanhalli, "Unsupervised Classification of Music Genre Using Hidden Markov Model", *in Proc. Of IEEE International Conf of Multimedia Explore (ICME04), Taibei, Taiwan, China, 2004*.

[3] G. Xu, N. C. Maddage, X. Shao, F. Cao and Q. Tian. "Musical Genre Classification Using Support Vector Machines," vol. 5, pp. 429-432. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2003*.

[4] G. Tzanetakis, and P. Cook "Musical genre classification of audio signals*", IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, July 2002, pp. 293 – 302.

[5] L. Rabiner and B. H. Juang, "Fundamentals of speech recognition".*Englewood Cliffs, NJ: Prentice-Hall, 1993*.

[6] T. Li, M. Oginara and Q. Li, "A comparative study on content-based music genre classification," in *Proc. of the 26th annual int. ACM SIGIR conf.on Research and development in information retrieval,* pp. 282–289. ACM, ACM Press, July 2003.

[7] S. Mallat, "A wavelet tour of signal processing", *San Diego, CA: Academic, 1998*.

[8] V. N. Vapnik, "Statistical Learning Theory", *New York, Wiley, 1998*.

[9] H. Soltau, T. Schultz, M. Westphal, and A. Waibel, "Recognition of music types", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seattle, 1998.*

[10]D.-N. Jiang, L. Lu and H.-J. Zhang, "Music Type Classification by Spectral Contrast Features," vol. 1, pages 113-116, *IEEE International Conference on Multimedia and Expo (ICME), Lausanne, Switzerland, August 2002*.

[11]Toni Heittola, "Automatic Classification of Music Signals", Master of Science Thesis, *Tampere University of Technology, Department of Information Technology*.