AUDIO CONTENT DESCRIPTION WITH WAVELETS AND NEURAL NETS

Stephan Rein[*]*, Martin Reisslein*[†]*, and Thomas Sikora*

Technical University Berlin and Arizona State University
srein@uni.de, reisslein@asu.edu, sikora@nue.tu-berlin.de

ABSTRACT

Precision audio content description is one of the key components of next generation internet multimedia search machines. We examine the usability of a combination of 39 different wavelets and three different types of neural nets for precision audio content description. More specifically, we develop a novel wavelet dispersion measure that measures obtained ranks of wavelet coefficients. Our dispersion measure in conjunction with a probabilistic radial basis neural network trained by only three independent example sets obtains a success rate of approximately 78% in identifying unknown complex classical music movements.

1. INTRODUCTION

Due to the immense and growing amount of audiovisual data that is available on the world-wide web (WWW), techniques for multimedia content retrieval and classification are becoming increasingly important. Next generation internet search machines are expected to be able to understand and process multimedia content. More precisely, a user query can be a mixture of multimedia data including text, voice, picture, and video content. The search machine will give a reasonable answer providing content that is highly related to the query and of relevance to the user. An audio content description and retrieval methodology for implementation in internet search machines should allow for a very compact content representation since there is an immense volume of audio data on the WWW. In addition, the methodology should allow for an efficient computation of these descriptors.

In this paper we develop and evaluate an audio content description and retrieval methodology that is based on a novel wavelet dispersion measure and is readily applicable for next generation internet search machines. We find that our measure efficiently describes the wavelet patterns corresponding to the audio content. We find that the wavelet dispersion data can be processed by a neural net to realize a computationally effective mapping and classification technique. We examine the performance of our wavelet dispersion measure for 39 different wavelets, different wavelet scales, and three different types of neural nets. In our performance evaluation we consider the following identification problem: the search machine is provided with a performance of a classical music movement (piece of a composition) and the task is to find the same movement in a different performance/recording, whereby the performances differ in time, frequency, sound environments, and recording quality. We consider four different performances/recordings of the same 32 movements in our evaluation. By combining the biorthogonal wavelet with the order numbers 3 (for reconstruction) and 9 (for decomposition) with the scales $1,\ 3,\ 5, \ldots, 47$ with a probabilistic radial network trained with three different performances, our methodology achieves a mean success rate of 78% for identifying the movements of a performance that is not in the search system's data base. The identification success rate for a performance known to the system is approximatively 100%.

1.1. Related Work

There exists a large body of literature on audio content description, sound classification, and audio retrieval. This literature includes audio fingerprinting systems for identification of audio songs *known* to the search system's data base, see for instance [1] [2]. Our system differs from these works in that it identifies *unknown* complex audio with a high success rate.

The existing body of literature also includes retrieval systems for the categorization of different sounds. Generally, the system is trained by a number of example sounds for classification of novel sound segments into *elementary* content based classes, see for instance [3][4][5][6][7][8][9][10]. Our system differs from these classification systems in that it identifies highly *complex* classical compositions even if they are from a different performance/recording and in that it employs a novel wavelet dispersion measure to obtain this goal.

To the best of our best knowledge, this is the first paper

[*] S. Rein performed this work while visiting Arizona State University, Tempe.

[†] Supported in part by the National Science Foundation through Grant No. Career ANI-0133252 and Grant No. ANI-0136774, as well as the state of Arizona through the IT301 initiative.

to propose a methodology for identifying highly complex musical audio recordings that are not part of the search system's data base.

2. THE WAVELET DISPERSION VECTOR: A NOVEL WAVELET DISPERSION MEASURE

In this section we outline our novel wavelet dispersion measure for extracting the characteristic features of an audio file. We refer to our measure as *wavelet dispersion vector*. To obtain this measure, a wavelet transform with S scales is performed on the audio data. The obtained coefficients are stored in a $S \times T$ matrix, where T denotes the number of audio samples. For each scale (represented by a row in the coefficient matrix), a rank histogram is calculated as explained by the following illustrative example. Let $\mathbf{C}$ denote the wavelet coefficient matrix. For illustration $\mathbf{C}$ has only 3 scales and 5 samples:

$$\mathbf{C} = \begin{array}{ccccc|c} 1 & 2 & 3 & 4 & 5 & \\ \hline 0.43 & 0.22 & 0.14 & 0.76 & 0.33 & 1 \\ 0.10 & 0.32 & 0.11 & 0.28 & 0.90 & 2 \\ 0.54 & 0.49 & 0.34 & 0.18 & 0.91 & 3 \end{array}. \quad (1)$$

The row and column indices do not belong to $\mathbf{C}$. We now construct the rank histogram for scale 1. For every value of scale 1 the corresponding rank within its column is estimated:

$$\begin{array}{ccccc|c} 1 & 2 & 3 & 4 & 5 & \\ \hline 0.43\ (2) & 0.22\ (3) & 0.14\ (2) & 0.76\ (1) & 0.33\ (3) & 1 \\ 0.10 & 0.32 & 0.11 & 0.28 & 0.90 & 2 \\ 0.54 & 0.49 & 0.34 & 0.18 & 0.91 & 3 \end{array}. \quad (2)$$

The number in brackets represents the rank within a column. For example, the first value of scale 1 ($\mathbf{C}(1,1) = 0.43$) obtained the second rank in the first column. This process is repeated for all scales:

$$\begin{array}{ccccc|c} 1 & 2 & 3 & 4 & 5 & \\ \hline 0.43\ (2) & 0.22\ (3) & 0.14\ (2) & 0.76\ (1) & 0.33\ (3) & 1 \\ 0.10\ (3) & 0.32\ (2) & 0.11\ (3) & 0.28\ (2) & 0.90\ (2) & 2 \\ 0.54\ (1) & 0.49\ (1) & 0.34\ (1) & 0.18\ (3) & 0.91\ (1) & 3 \end{array}. \quad (3)$$

We now only retain the ranks:

$$\begin{array}{ccccc|c} 1 & 2 & 3 & 4 & 5 & \\ \hline 2 & 3 & 2 & 1 & 3 & 1 \\ 3 & 2 & 3 & 2 & 2 & 2 \\ 1 & 1 & 1 & 3 & 1 & 3 \end{array}. \quad (4)$$

There is some redundancy in this matrix. The third row can be calculated from the other two rows. We keep this redundancy for illustration.

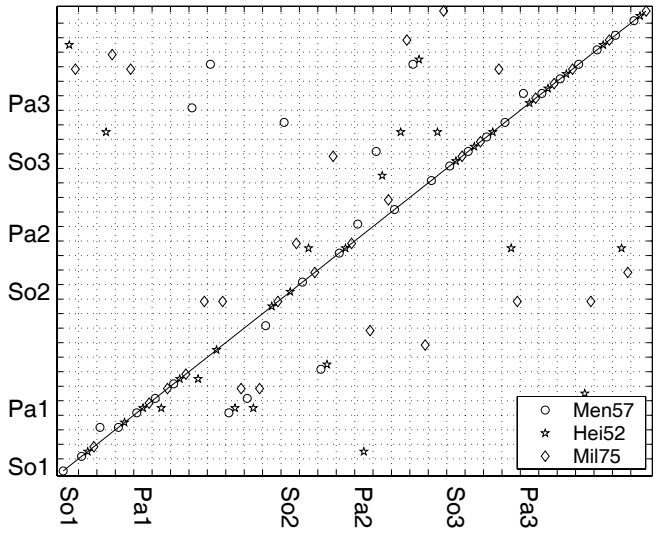

Fig. 1. Performance of wavelet dispersion vector in conjunction with a maximum correlation measure: Approximately 60% of the audio pieces are correctly identified.

Now, for each scale (for each row) a rank histogram is constructed. Thereby the ranks within a row are counted to obtain the values of the wavelet dispersion measure:

$$C_{\text{disp}} = \begin{array}{ccc|c} 1 & 2 & 3 & \\ \hline 1 & 2 & 2 & 1 \\ 0 & 3 & 2 & 2 \\ 4 & 0 & 1 & 3 \end{array}. \quad (5)$$

For example, the first row obtained one time the first rank, two times the second rank, and two times the third rank. There is again some redundancy in this matrix, i.e., one column could be left out. We perform a low–complexity redundancy reduction which discards the lowest and highest ranks of the wavelet dispersion histogram, as these represent outlying wavelet coefficients, see [11] for details.

The wavelet rank dispersion data is now stringed to be stored in a vector. For our illustrative example this vector is given as

$$\vec{v} = [\,1 \quad 2 \quad 2 \quad 0 \quad 3 \quad 2 \quad 4 \quad 0 \quad 1\,]. \quad (6)$$

We call this vector *wavelet dispersion vector*. For every audio file such a vector can be constructed to represent the characteristic audio features.

2.1. Performance Evaluation of Wavelet Dispersion Vector

In this section we give an overview of the evaluation of the identification and generalization properties of our wavelet rank dispersion measure. We employ an audio data base

containing 128 different audio files. Specifically, we employ six pieces—containing a total of 32 movements—composed by Johann Sebastian Bach, the Sonatas and Partitas for Solo Violin, Bachwerkeverzeichnis (BWV) 1001–1006. We consider four different performances of these 32 movements; specifically, the performances Menuhin 1934–6 (Men36), Menuhin 1957 (Men57), Heifetz 1952 (Hei52), and Milstein 1973 (Mil75). These audio recordings were chosen because they have consistent relevance, represent different levels of qualities, and have polyphonic and not separable phenomena.

We calculate the 128 different wavelet dispersion vectors for the first five seconds of each of the considered pieces. We store these vectors in a so-called *wavelet classifier matrix*. This matrix has always 128 columns. The number of rows depends on the number of employed wavelet scales and the dimension reduction technique. Thus there exist different classifier matrices for different wavelet mother functions, different wavelet scales, and different dimension reduction parameters.

In Figure 1 we show the results for a search scenario where the 32 descriptors of the Men36 recording are entered as the user query and the 32x3 descriptors from the remaining 3 recordings are employed by the search machine data base. Each individual Men36 classifier query (x-axis) is assigned an answer (y-axis) employing one of three classifier sets. Therefore, on the plot, there are 3 points in each column, whereby each column represents one of the 32 movements of the Men36 recording. The three points in a given column represent from left to right the matched query results in the Men57, Hei52, and Mil75 recordings, respectively. We initially obtain an answer to a user query by calculating the maximum correlation between the single query-classifier and the 32 different classifiers of one of the 3 search machine classifier sets. For example, observe in Figure 1 how a user query containing the second movement of Partita 1 of the Men 36 recording (Men36Pa1ii) is answered. When employing the 32 column classifier matrix of Hei52, the search machine would identify this piece as the first movement of Partita 1. This identification is obtained by the maximum of 32 different correlations between the Men36Pa1ii classifier and the 32 Hei52 classifiers. The maximum correlations between Men36So1ii and the 32 Men57 and Mil75 classifiers give the correct answers. If all points are on the line in Figure 1, then all pieces were correctly identified. We observe that approximately 60% of the movements are correctly identified.

Each point in Figure 1 represents an audio retrieval of only one recording. In the related literature, search and retrieval systems are proposed that are trained by many audio files describing the same content. For example, in [4], more than 48 sound clips of laughter are employed to construct a laughter classifier. In this work, we are interested in employing a minimum number of audio files to construct a classifier that already allows for reasonable results. Therefore, our data base only contains 4 different recordings of the same pieces. Thus, a piece unknown to the identification system can be identified by a classifier that has been constructed from 3 different recordings, as detailed in the following section.

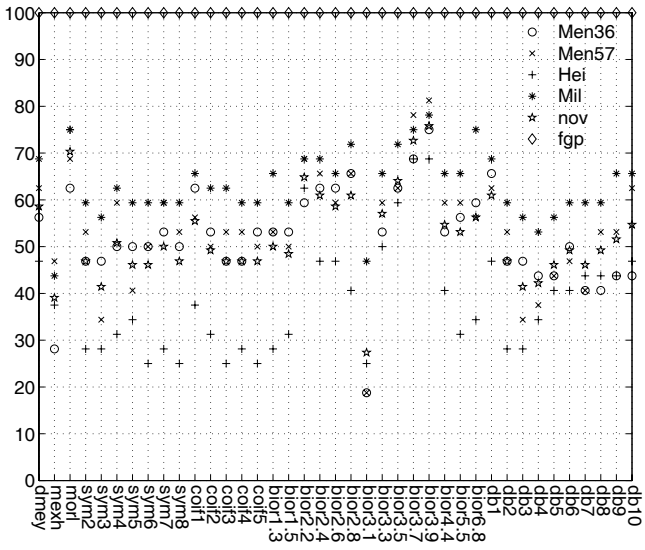

Fig. 2. Percentage performance of the probabilistic radial network for the scales 1:2:48. The best performance of 78% is achieved by the bior3.9 wavelet.

3. EVALUATION OF NEURAL NET CLASSIFICATION

In the preceding section we measured the similarities between the different descriptors by a correlation measure. Generally, there are a lot of other possibilities for comparing the different content description vectors. In this section we consider three different types of neural nets to process the wavelet dispersion vectors to answer an audio query. We combine the wavelet dispersion measure obtained with 39 different wavelets and three different types of neural networks. Specifically, we consider the following wavelets and wavelet families: Meyer wavelet, Mexican hat wavelet, Morlet wavelet, 7 types of symlets (modified Daubechies wavelets), 5 types of coiflets, 14 types of biorthogonal wavelets, and 10 types of Daubechies wavelets. We consider the single-layer perceptron network, the backpropagation network, and the probabilistic radial basis network. We refer the reader to [11] for details on these wavelets and networks.

For each of the 128 (4 performers $\cdot$ 32 pieces) audio files we perform a wavelet decomposition (wavelet scales $1, 3, 5, \ldots, 47$) and construct the wavelet dispersion vectors

as detailed in Section 2. Each vector describes a segment of the first 5 seconds of an audio file. As we examine 39 different wavelets, we calculate $128 \cdot 39$ classifier vectors, each of the length $24 \cdot 24$ elements. Thus we obtain a 576x($128 \cdot 39$=4992) wavelet classifier matrix. We perform a dimension reduction as detailed in [11], reducing the dimension of the wavelet classifier matrix from 576x4992 to 440x4992. For processing this matrix with the neural nets, each vector of this matrix is normalized to zero mean and unit standard deviation.

The neural nets were trained with a minimum of epochs and neurons to allow an identification of example vectors known to the identification system with a success rate of approximatively 100% over the entire range of different wavelets. Each neural net is trained by example vectors of 3 different players to allow a user query of one novel performer, i.e., a performer not known to the search and retrieval system. Thus, there are 4 possible neural net classifier constellations.

Figure 2 reports the retrieval results for the radial basis network, which gives the best performance among the three considered networks, see [11]. The 4 different classifier constellations are indexed by *Men36, Men57, Hei52,* and *Mil75*. One Men36 point reflects the mean success rate of 32 different user audio queries on the Men36 recording. In this case the search system has been trained on the recordings Men52, Hei52, and Mil75. If all 32 pieces of the Men36 recording are correctly identified, the mean success rate is 100%. The points *nov* (novel) give the mean success percentage rate (piece identification) across the four search constellations. In addition, the points *fgp* (fingerprint) give the success rates for scenarios where all four performances have been used for the training, i.e., the audio query piece is known to the system.

We observe from the figure that the success rates for novel data range from 62% to 75% for the Morlet wavelet. The best mean identification success rate of 78% is achieved by the bior3.9 wavelet. These results indicate that our methodology of combining the novel wavelet dispersion vector with a neural network achieves good generalization, i.e., identification of movements that are unknown to the system.

4. CONCLUSION

We have proposed a novel methodology to solve a highly complex classification problem. Our methodology may form the basis for an identification service of classical audio movements that are not part of the retrieval system's data base, a problem that is likely to arise in next generation Internet search machines.

We have evaluated the proposed system with 32 different movements using a very small training set of 96 movements. Thus one class of the identification system is only constructed by three different example audio clips using small extracts with a duration of 5 seconds. The system achieves a mean success rate of up to 78%.

5. REFERENCES

[1] S. Sukittanon and L.E. Atlas, "Modulation frequency features for audio fingerprinting," in *Proc. of ICASSP '02*, Orlando, FL, May 2002, vol. 2, pp. 1773–1776.

[2] C.J.C. Burges, J.C. Platt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *IEEE Trans. Speech and Audio Proc.*, vol. 11, May 2003.

[3] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. on Speech and Audio Proc.*, vol. 10, October 2002.

[4] M. Casey, "MPEG-7 sound recognition tools," *IEEE Trans. Circuits and Systems for Video Techn.*, vol. 11, no. 6, pp. 737–747, June 2001.

[5] T. Zhang and C.-C.J. Kuo, "Hierarchical classification of audio data for archiving and retrieving," in *Proc. of ICASSP' 99*, Phoenix, AZ, March 1999, vol. 6, pp. 3001–3004.

[6] G. Guo and S.Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, January 2003.

[7] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang, "Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification," in *Proc. of ICASSP '03*, Hong Kong, April 2003, vol. 5, pp. 628–631.

[8] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, and A. Linney, "Classification of audio signals using statistical features on time and wavelet transform domains," in *Proc. of ICASSP' 98*, Seattle, WA, May 1998, vol. 6, pp. 3621–3624.

[9] G. Li and A. A. Khokhar, "Content-based indexing and retrieval of audio data using wavelets," in *Proc. of ICME*, August 2000, pp. 885–888.

[10] S. R. Subramanya and A. Youssef, "Wavelet-based indexing of audio data in audio/multimedia databases," in *Proc. Multi-Media Database Management Systems*, Aug. 1998, pp. 46–53.

[11] S. Rein and M. Reisslein, "Audio content description with wavelets and neural nets," Tech. Rep., Arizona State University, Dept. of Electrical Eng., Oct. 2003, avail. at `http://www.fulton.asu.edu/~mre`.