Video Genre Categorization Using Audio Wavelet Coefficients

Phung Quoc Dinh[†], Chitra Dorai[‡], Svetha Venkatesh[†]

Department of Computer Science[†]
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
{phungquo, svetha}@cs.curtin.edu.au

IBM T. J. Watson Research Center[‡]
P.O. Box 704, Yorktown Heights
New York 10598, USA
dorai@watson.ibm.com

Abstract

In this paper, we investigate the use of a wavelet transform-based analysis of audio tracks accompanying videos for the problem of automatic program genre detection. We compare the classification performance based on wavelet-based audio features to that using conventional features derived from Fourier and time analysis for the task of discriminating TV programs such as news, commercials, music shows, concerts, motor racing games, and animated cartoons. Three different classifiers namely the Decision Trees, SVMs, and k-Nearest Neighbours are studied to analyse the reliability of the performance of our wavelet features based approach. Further, we investigate the issue of an appropriate duration of an audio clip to be analyzed for this automatic genre determination. Our experimental results show that features derived from the wavelet transform of the audio signal can very well separate the six video genres studied. It is also found that there is no significant difference in performance with varying audio clip durations across the classifiers.

1 Introduction

The rapid proliferation of multimedia on the Internet has led to abundant research targeted at automatically characterizing and categorizing the content of video streams. Classifying the streams and the scenes contained within into different categories such as news, commercials, sports, etc [11] enables efficient cataloguing and speedy retrieval during search with large video archives. When analysing the content of a video scene, audio is found to be as important as visual information [3]. This applies not only to speech information in the audio track, which provides semantic indices into what is being spoken about, but also to generic acoustic properties of the track [3]. For example, we can easily tell a weather report from a football game without actually watching the program or understanding the words spoken. Thus, analysis of audio for video genre identification has recently attracted the attention of many researchers [3][5][13]. A further advantage is that audio processing requires much less computation than image processing. Analysis of audio embedded in the video streams can enable not only video indexing and retrieval, but also potentially offer better tools for segmentation, abstraction, and summarisation of video content.

Liu *et al.* [5][4] were among the early researchers to use audio to enhance the understanding of the video content. Their work is based on the observation that different objects will naturally produce different sounds that potentially result in different aural signatures. Features such as silence ratio and subband energy ratio were extracted from the time and Fourier domain analyses of the audio signal. Using neural networks as the classifier, [5] reports an accuracy of 82.5% in separating four kinds of programs, namely commercials, basketball, football games, news reports, and weather forecasts. News reports, sport/games and advertisements in their experiments were well separated using audio; however the performance was poor in distinguishing between two different types of reports and two kinds of sports. In a more recent work by the same authors [4], HMMs were successfully used to distinguish between commercials, basketball/football games, news reports, and weather forecasts. For each of these programs, an ergodic HMM was built using audio clip-based features as observation vectors. The maximum likelihood estimator was then applied to classifying unseen data using the trained models. Their experimental results showed that HMMs were powerful for video content classification using audio information, compared with neural networks (with an improvement of 11.9%). In [6], audio characterisation was performed directly with compressed data (MPEG) for video indexing. The audio was classified into dialog, non-dialog, and silence intervals using features constructed from the energy, pitch, spectrogram, and pause rate domains. The authors report some misclassification between dialog and non-dialog segments.

The theory of wavelets has existed for decades, however its application to audio segmentation and classification is still very limited. In [10] the authors discuss the use of wavelet transforms for speech segmentation and classification. They show that with the information from the transform, locations of spectral changes in the speech signal can be detected accurately. Their experimental results indicate that fricative sounds can be easily identified and most of the beginning and ends of plosive sounds can be located. In [2], wavelet coefficients are utilised for audio information retrieval and indexing. Statistical properties of wavelet coefficients such as zero-crossing rate, mean and standard deviation at multiple scales are used as features to construct a hierarchical indexing scheme. Experimental results show a recall rate of more than 70% for sample queries with diverse audio characteristics.

In this paper, we investigate the application of the wavelet transform to program audio tracks to reliably distinguish different video genres. We compare the classification accuracy of our wavelet-based features to that based on conventional features from Fourier and time analyses of audio signals. As a demonstration of this approach, our system distinguishes between *news*, *commercials*, *music shows*, *concerts*, *sport games*, and *animated cartoons*. Further, we investigate estimating an effective clip duration for this analysis. Our results show that these categories can be well separated in the wavelet domain. The classification figures further show that there is no significant difference in performance with four different clip sizes: 0.5s, 1.0s, 1.5s, and 2.0s. The novelty of our work lies in the use of audio wavelet coefficients alone to discriminate video genres and in demonstrating their effectiveness. This differentiates our work from [2][9][10] that use wavelets exclusively for audio content management.

2 Audio Features for Video Genre Identification

Two sets of aural features are considered in this paper. The first set consists of features extracted from the wavelet transform among which we propose two new features along with many existing ones reported in the literature. The second feature set is designed for comparison purposes to include features from the time and Fourier domains.

2.1 Wavelet-Based Features

The discrete wavelet transform (DWT) is used in our work. In the discrete case, filters of different cutoff frequencies are used to analyse the signal at different scales [2]. In this work, we use Daubechies wavelet transform (DAUB4) [7] with 6 levels of decomposition. The audio is sampled at 44.1kHz and the following 7 subbands are used in our study.

Level	Subbands (in Hz)	Examples
1	11025 - 22050	noise, friction sounds
2	5513 - 11025	speech, lyrics
3	2756 - 5123	
4	1378 - 2756	lower harmomics, overtones
5	689 - 1378	
6	334 - 689	
7	0 - 334	lower tones, bass

We define two new features, namely centroid and bandwidth. In addition, we use standard features such as energy and variance computed in each subband.

Centroid and Bandwidth: Given wavelet coefficients of an audio clip, let w_i be the ith wavelet coefficient and N, the total number of coefficients. We define *centroid* and *bandwidth* as follows:

$$\mathsf{W}_c = \frac{\sum_{i=1}^{N} i\left|w_i\right|^2}{\sum_{i=1}^{N}\left|w_i\right|^2},\ \mathsf{W}_b^2 = \frac{\sum_{i=1}^{N}(i-\mathsf{W}_c)^2\left|w_i\right|^2}{\sum_{i=1}^{N}\left|w_i\right|^2} \quad (1)$$

In the kth subband ($1 \leq k \leq 7$), let N_k be the number of wavelet coefficients, μ_k the average amplitude and w_k^i the ith wavelet coefficient. We define the following features for each subband. Similar features have been proposed for audio indexing and retrieval in [2][9].

Subband Energy: This feature gives a rough measure of energy in each subband and is calculated as the sum of absolute value of coefficients in each band, normalised by the number of coefficients in that band.

$$\mathsf{W}_{\mathsf{E}}^{k} = \frac{1}{N_k}\sum_{i=1}^{N_k} |w_k^i| \tag{2}$$

Statistics of energy for different subbands can be very useful in discriminating different genres. With audio clips from a motor racing program, for example, particular subbands may correspond to noise generated by engines or friction of the vehicles with the road.

Subband Variance: The variance is defined as:

$$\mathsf{W}_{\mathsf{V}}^{k} = \frac{1}{N_k}\sum_{i=1}^{N_k} \left[w_k^i - \mu_k\right]^2 \tag{3}$$

Zero Crossing Rate: Analogous to its definition in time domain, a zero crossing is marked when successive samples have different signs and it describes how often the signal frequency has changed over a period of time within a subband [2]. The zero crossing rate for each subband is given as:

$$\mathsf{W}_{\mathsf{Z}}^{k} = \frac{1}{(N_k - 1)}\sum_{i=1}^{N_k - 1} \left|sign(w_k^i) - sign(w_k^{i+1})\right| \tag{4}$$

where $sign(x)$ takes a value of 1 if $x \geq 0$ and -1 otherwise.

In summary, the feature vector from the wavelet transform is obtained as $\mathsf{W}^{avelet} = \{\mathsf{W}_{\mathsf{E}}^{i}, \mathsf{W}_{\mathsf{V}}^{i}, \mathsf{W}_{\mathsf{Z}}^{i}, \mathsf{W}_{c}, \mathsf{W}_{b}\}$ where $i = 1, 2, \ldots, 7$.

2.2 Features from Time and Frequency Domains

For the sake of comparison, we also consider a set of features from the time and Fourier domains. These features have been investigated by other researchers and found effective in discriminating video scenes [3][5]. To compute these features, an audio clip is divided into a series of frames with each containing 512 sampling points and overlapping with the previous ones by 256 samples. Features are first extracted at the frame level, from which clip level features are derived usually by taking a weighted average. The following features are used in our work.

Loudness standard deviation and dynamic range $(\mathsf{L}_\sigma, \mathsf{L}_\Delta)$. Dynamic change in loudness is computed from the loudness contour as $\mathsf{L}_\Delta = \frac{\max_{\mathsf{L}} - \min_{\mathsf{L}}}{\max_{\mathsf{L}}}$, where $\min_{\mathsf{L}}$ and $\max_{\mathsf{L}}$ are the minimum and maximum loudness values respectively. L_σ is the standard deviation of the contour [3].

ZCR standard deviation and non-silence ratio $(\mathsf{Z}_\sigma, \mathsf{T}_{nsr})$. Similarly, the zero crossing (ZCR) contour is first computed, based upon which, Z_σ is derived. Information from loudness and ZCR contours is used to determine whether a frame is silent [5]. The non-silence ratio T_{nsr} is then determined as the ratio of total duration of non-silent frames to the whole clip [3].

From the Fourier domain, we use the mean of three subband ratios $(\mathsf{S}_\mu^1, \mathsf{S}_\mu^2, \mathsf{S}_\mu^3)$ proposed in [5] as features. In brief, the feature vector to be compared with the wavelet feature set is: $\mathsf{O}^{ther} = \{\mathsf{L}_\sigma, \mathsf{L}_\Delta, \mathsf{Z}_\sigma, \mathsf{T}_{nsr}, \mathsf{S}_\mu^1, \mathsf{S}_\mu^2, \mathsf{S}_\mu^3\}$.

3 Video Genre Classification

Our goal is to discriminate between *news, commercials, vocal music shows, concerts, motor racing game,* and *animated cartoons*. *News* is characterised by speech dominating the audio track. *Commercials* refer to advertising portions of broadcast video. *Shows* in our work refer to performances by music bands (e.g., Oasis, Scorpion, Pink Floyd) characterized by music playing in the background and people singing in the foreground. *Concert* refers to pure instrumental music produced by different musical instruments. As far as sports programs are concerned, [1] observes that there is almost nothing in common between videos of different kinds of sports, and therefore taking sport as a homogeneous genre is inappropriate. In this work we only consider *motor racing game* as an instance (or a sub-genre) of sports. *Animated cartoons* are characterised by cheerful and excited themes, as most of them are produced for children.

Approximately 90 minutes of digital video material was collected in MPEG-1 format containing six types of programs mentioned. The accompanying audio for each video class was then extracted and digitized at 44.1kHz with mono channel. The audio data is split into a series of clips with four different durations: 0.5s, 1.0s, 1.5s and 2.0s. With a clip size of 1.0s, our dataset contains a total of 1593 clips for news, 1062 for commercials, 1129 for cartoons, 1264 for motor racing games, 681 for shows and 897 for concerts. Feature extraction is then performed on the data to compute W^{avelet} and O^{ther}.

3.1 Feature Trends in Video Genres

To understand the discrimination capability of our features, for example, the energy in various wavelet subbands, we compute the energy for different subbands of each clip. Then, for each video genre we generate a histogram using these subband energies across all clips belonging to that genre. Figures 1,2,3 and 4 show the plots for subbands 1, 2, 4, and 6 respectively.

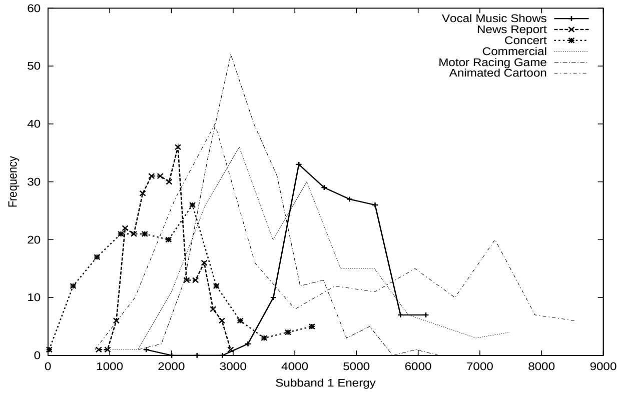

Figure 1. Energy histograms in subband 1.

¿From Figure 1, we can see that two types of music programs, namely vocal music *shows* (right curve) and *concert* (left curve) are almost completely separated with energy information in subband 1. The high energy in subband 1, which covers the high frequency band, for the *concerts* class can be explained by the variety in musical instruments being played, especially due to those producing very high tones (e.g., flute, violin).

In Figure 2, the histogram of the *motor racing game* relatively settles to the right, and thus energy in subband 2 distinguishes it from all other categories.

News and *commercials* classes are also fairly well separated using energy in subband 4 as shown in Figure 3.

Finally, Figure 4 shows that *animated cartoons* (the left curve) are very distinguishable from *concerts* and *shows* clips with the energy information in subband 6.

3.2 Genre Classification

¿From the feature vectors computed, we randomly selected $\frac{2}{3}$ of them for training and the rest as unseen data for testing with three different classifiers: Decision Trees, k-Nearest Neighbours ($k = 6$, number of video

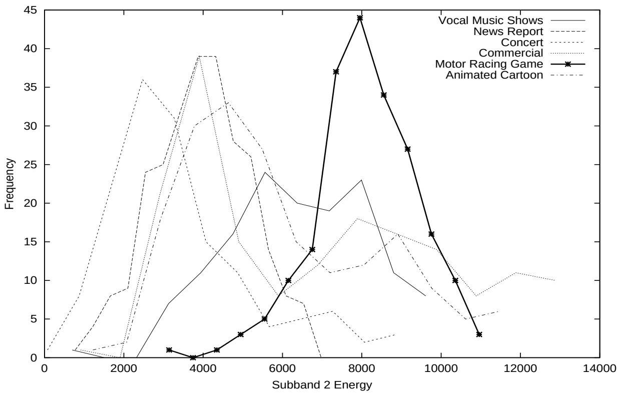

Figure 2. Energy histograms in subband 2.

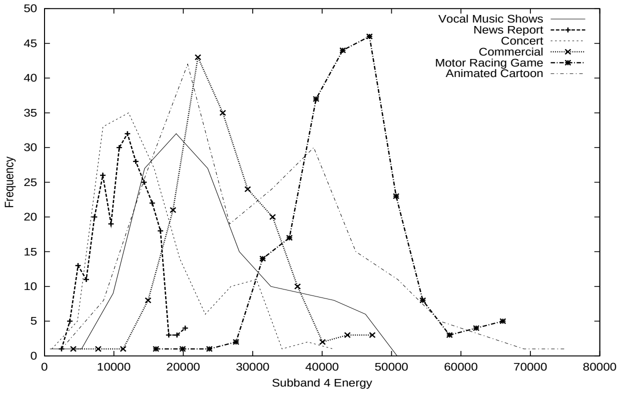

Figure 3. Energy histograms in subband 4.

genres to be classified) and Support Vector Machines with linear kernels. These classifiers are available in the Waikato Software Environment for Knowledge Analysis (WEKA) [12].

The Decision Tree classifier [8] recursively subdivides a set of data by using the concept of entropy from information theory. The feature which provides the most information gain, as defined by the difference in entropy, at each recursion is used to form a decision based on the values of the feature. The result is a tree where each node has a feature and a decision depending on its value. The k-Nearest Neighbours classifier generates clusters representing the classes of feature points and assigns a feature instance to the cluster that has k instances closest in Euclidean distance to it. The Support Vector Machine-based

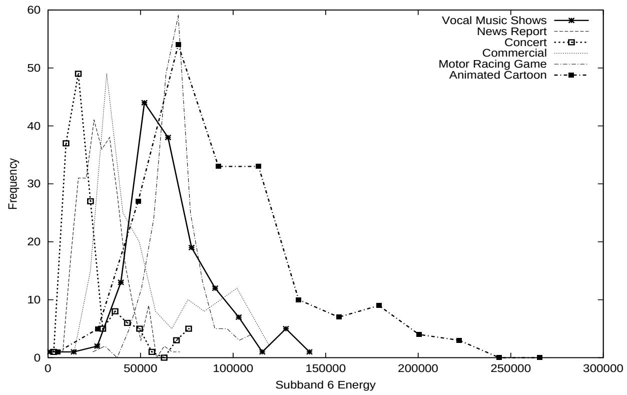

Figure 4. Energy histograms in subband 6.

classifier maps an input space into a high dimensional feature space through some mapping function and then constructs the optimal separating hyperplane in the high dimensional feature space [3]. A summary of our experimental results is provided in Table 1. It tabulates results for each clip size and each classifier in terms of the number of clips correctly and erroneously classified, and the resulting classification *accuracy*, which is measured as the ratio of number of audio clips correctly classified to the total number of clips. From the results, the following observations can be made.

- The performance of our proposed W^{avelet} based classification competes with that of the O^{ther} feature set based classification in most of the cases, indicating highly reliable performance with our approach using wavelet-based features.
- Surprisingly, the performance of k-Nearest Neighbours classifier is better than C4.5 and SVMs in all cases. This is possibly because the video classes are well clustered in the feature space (see Figures 1,2,3 and 4).
- As far as the impact of clip duration is concerned, there is no significant difference in the performance across classifiers. However, when computing the average performance across video genres and classifiers, 1.5s or 2.0s (88.15% and 89.25% accuracy respectively compared with 87.9% 88.15% for 0.5s and 1.0s) is recommended as the preferred length for each clip.

Table 2 shows the resulting confusion matrix for a clip duration of 1s when DTs are used as classifiers. A careful analysis of confusion matrices shows that *motor racing game* is the easiest one to be discriminated from others. The most misclassification occurs in distinguishing *news* from *commercials*, *commercials* from *shows* and *news* from *cartoons*. This may be due to the fact that there are many cases where *news* and *commercials* are very similar aurally. The fact that many show clips are misclassified as *commercials* can be anticipated by similar aural cues (speech in the foreground and music in the background) in these two categories.

Table 2. Confusion matrix from genre classification using Decision Trees with a clip duration of 1s.

Classified ⟶	A	B	C	D	E	F
A = *news*	489	18	0	0	0	12
B = *coms*	**24**	282	6	2	9	18
C = *MTgames*	3	15	414	6	0	12
D = *concert*	3	15	6	324	0	6
E = *shows*	6	**33**	6	0	249	9
F = *cartoons*	0	21	4	8	18	284

4 Conclusion

Using wavelet transform as the representational mechanism for the audio signal, we have demonstrated that video genres can be reliably determined automatically with only audio information. The features derived from wavelet coefficients of the audio signal are computationally simple, yet yield high genre classification accuracy, thus proving to be a robust feature set. Three different classifiers such as the Decision Trees, SVMs, and k-Nearest Neighbours are studied to analyse the reliability of the performance of our wavelet features based approach. Our study on the impact of clip duration has also shown that there was no significant difference in the classification performance with various durations.

References

[1] S. Fischer, R. Lienhart, and W. Effelsberg. Automatic recognition of film genres. In *The Third ACM International Multimedia Conference and Exhibition (MULTI-*

Table 1. Video genre classification accuracy with Decision Trees, k-Nearest Neighbours and SVM classifiers.

F. Set	C4.5, DTs			SVMs			k-NN			Clp. size
	N_{corr}	N_{wrong}	A_{crcy}	N_{corr}	N_{wrong}	A_{crcy}	N_{corr}	N_{wrong}	A_{crcy}	
W^{avelet}	4080	555	**88.0**%	3927	1098	**84.7**%	4320	315	**93.2**%	0.5s
O^{ther}	4176	459	**90.1**%	3774	861	**81.4**%	4179	456	**90.2**%	
W^{avelet}	2052	261	**88.7**%	1980	315	**86.4**%	2184	129	**94.4**%	1.0s
O^{ther}	1899	414	**82.1**%	1800	513	**77.8**%	1959	354	**84.7**%	
W^{avelet}	1032	126	**89.1**%	1050	108	**90.7**%	1098	60	**94.8**%	1.5s
O^{ther}	981	177	**84.7**%	957	201	**82.6**%	1008	150	**87.0**%	
W^{avelet}	1044	108	**90.6**%	1041	111	**90.4**%	1107	45	**96.1**%	2.0s
O^{ther}	993	153	**86.7**%	957	195	**83.1**%	1023	129	**88.8**%	

MEDIA '95), pages 367–368, New York, November 1995. ACM Press.

[2] G. Li and A. A. Khokhar. Content-based indexing and retrieval of audio data using wavelets. In *International Conference on Multimedia and Expo (ICME)*, August 2000.

[3] Z. Liu. *Adaptive and Multimodal Approach to Multimedia Content Analysis*. PhD thesis, Polytechnic University, September 2000.

[4] Z. Liu, J. Huang, and Y. Wang. Classification of TV programs based on audio information using hidden markov model. In *IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing*, pages 27–32, December 1998.

[5] Z. Liu, J. Huang, Y. Wang, and T. Chen. Audio feature extraction & analysis for scene classification. In *IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing*, New Jersey, June 1997.

[6] N. V. Patel and I. K. Sethi. Audio charactization for video indexing. In *Proceedings of SPIE*, volume 2670, pages 373–383, 1996.

[7] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Plannery. *Numerial Recipires in C: the Art of Scientific Computing*. Cambridge University Press, second edition, 1992.

[8] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, California, USA, 1993.

[9] S. Subramanya and A. Youssef. Wavelet-based indexing of audio data in audio/multimedia databases. In *Proceedings of the International Workshop on Multimedia Database Management Systems*, 1998.

[10] B. Tan, R. Lang, H. Schroder, and P. Dermody. Applying wavelet analysis to speech segmentation and classification. In *Proceedings of Wavelet Applications Conference*, Orlando, April 1994.

[11] B. T. Truong, C. Dorai, and S. Venkatesh. Automatic genre identification for content-based video categorization. In *Proc. 15th International Conference on Pattern Recognition*, volume II, pages 230–233, Barcelona, Spain, September 2000.

[12] I. H. Witten and E. Frank. *Data Mining: Practical Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers, 2000.

[13] T. Zhang and C. J. Kuo. Hierarchical system for content-based audio classification and retrieval. In *Proceeding of SPIE's Conferences on Multimedia Storage and Archiving System III*, volume 3527, pages 398–409, Boston, United States, November 1998.