Audio Meets Image Retreival Techniques

Dave Kauchak
Department of Computer Science
Univerisity of California, San Diego
La Jolla, CA 92037
dkauchak@cs.ucsd.edu

Abstract

In this paper we examine the problem of audio retrieval. We make a number of key contributions to this field. First, we examine artist recognition/retrieval as a problem instead of the traditional genre classification. This problem has the motivating benefit that there is a known, uncontraversial ground truth. Second, and more importantly, we suggest borrowing research from the image retrieval community. We provide results from one image retrieval technique ported over to audio retrieval. This technique consits of taking the discrete wavelet transform of the audio, histogramming the results and using statistical histogram comparison metrics to compare similarity. The results are not outstanding, but we do show that this sort of research can be done fairly easily and productively.

1 Introduction

The expanse of computer technology along with an increasing interconnectivity (i.e. the internet) has had a huge impact on a wide range of applications. One of the major effects of this technological revolution is that there is a vast amount of data available at the fingertips of anyone who owns a computer. This availability of information has brought a number of problems to the surface, all involving identifying or retrieving pertinent information within this vast amount of data.

Multimedia is a generic term for a broad range of different types of information ([13]). Multimedia information may consist of text, images, video and audio information. The task of processing multimedia has generally been broken down into appropriate sub-fields. Text retrieval methods have been fairly successful at identifying relevant information in documents and relevant documents themselves ([6]). Image retrieval techniques have come a long way since the first conception of querying a database containing images ([12] [3] [2]). Audio retrieval techniques have only recently been focused on ([4] [1] [9] [10].

Fundamentally, these tasks are not that different. The basic idea of a retrieval system is to examine various features of the training data, associate those features with some similarity metric or model and then compare those features against a set of possible solutions (i.e. documents, images, audio clips, etc.). There has been some sharing of techniques with some success between various retrieval techniques

([7]). However, there has not been enough sharing of techniques among these different fields.

In this paper, we are interested in examining audio. However, instead of developing new and novel strategies for processing audio documents, we propose that researchers in the audio field examine techniques from image retrieval and classification where the medium is similar to audio, but the problem has been analyzed more thoroughly. We do not suggest that current research for new techniques stop, however, what we do suggest is that researchers consider examining the techniques that have already been applied to images with success.

We have a number of goals in writing this paper. First, beyond just suggesting that researchers apply image processing techniques to audio, we provide a number of examples of this approach and show that not only can it be done without too much effort, but the results are also acceptable. Second, we propose artist or band recognition/retrieval as a problem for audio processing. One of the key advantages of this problem is that there is a known ground truth. Too often audio papers analyze techniques problems such as genre classification where there is a general opinion of what the correct genre might be, but by no means an absolute. Third, all of the techniques presented allow the system to train from multiple inputs. This has a key advantage over systems that only use a single input in that the results can be better generalized. Finally, as has been the emerging trend, this paper deals exclusively with raw audio ([11]) instead of midi format ([5]) or similar formats. Raw audio is advantageous mostly due to its accessible both in ease of obtaining and the shear quantity available.

The paper is laid out as follows. In the section 2, we present the basic analogy between the image domain and the audio domain. Within this section we present an algorithm from image retrieval and show how this algorithm can be used as a model for an audio retrieval system. In section 3, we explain in better detail the song data set that we worked with for experimenting and describe the setup for the testing. In section 4, we then show the results of these algorithms on the data set of popular music. Finally, in section 5, we summarize these results and hint at directions for future research.

2 Image Processing to Audio

Images are two dimensional signals sampled at some rate (i.e. pixels per inch) and represented as pixel values in some color scheme. Similarly, audio is simply a one dimensional signal sampled at some rate (i.e. samples per second or Hz) and represented as amplitude values. Processing techniques, such as filtering (i.e. applying some filter to the input source), can be applied to both images and audio. Given this, the following image processing techniques are provided with the appropriate conversion to audio.

2.1 Filter and Histogram

In [8], a fairly simple approach was taken to image retrieval. The basic idea was to use histogram comparison methods to compare the similarity of images. Previous histogramming approaches have used simple histograms of color or other metrics. In [8], the image was passed through a number of different filters. This filtering produced a number of different filtered images. These filtered images where then histogrammed and used as a model to represent that specific image for comparison purposes.

Once a histogram was obtained, the paper borrowed from a number of different fields such as statistics and information theory for various histogram comparison

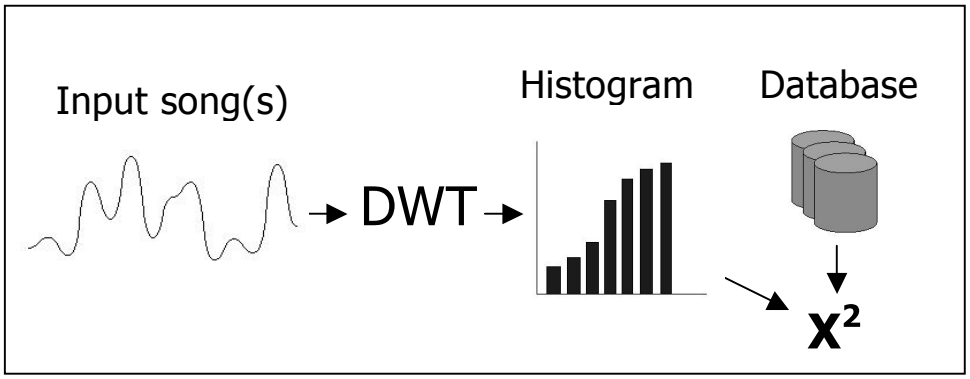

Figure 1: Audio retrieval method based on [8]. The coefficients of the different levels of the DWT are histogrammed and compared using X2 with the histograms from other songs in the database.

methods. The basic results from this paper showed that there was no absolute perfect measure under all circumstances, however, the chi-square statistic performed fairly well over the broad range of tests.

Given the relative simplicity of this approach and the ease by which it can be applied to audio, we chose this as a good example to demonstrate the key concept of porting image techniques over to audio. The basic idea for audio is basically the same as described above for images. A general view of the algorithm can be seen in Figure 1.

First, the song is split into a number of different frequency bands using the discrete wavelet transform (DWT). The DWT is used for a number of reasons. The DWT has been used in a number of audio applications successfully and seems to be the emerging technique for doing this sort of processing ([11]). Also, the DWT is very similar to the process used in [8] where a gabor filter is used to split the images. This is a good example showing how an image technique can be used as a guide, but not an exact manual, for developing new algorithms.

Next, the coefficients from the DWT are histogrammed. A number of different histogramming methods were examined in [8]. We examine a couple of different methods for histogramming in this paper. The simplest method is simply to bin each level with the same number of bins. This method is call "Normal" binning in this paper.

One problem arises with using the DWT. Because of the way in which the bands are broken up, the number of coefficients at each level is not the same. In fact, the number of coefficients halves at each progressive level. This leads to a fairly large discrepancy between levels. To try and account for this, another form of histogramming was used called "proportional" where the count per bin remained the same. This means that the lower level coefficients would be binned into a much larger number of bins than the higher levels.

Finally, after the sound files have been histogrammed, they are then normalized and compared with other histograms from the database of possible answers (i.e. other songs). As mentioned above, we chose to use the chi-square measure to determine the similarity between songs:

$$\mathrm{X}^2(I,J) = \sum_i \frac{\left(I(i) - J(i)\right)^2}{I(i) + J(i)}$$

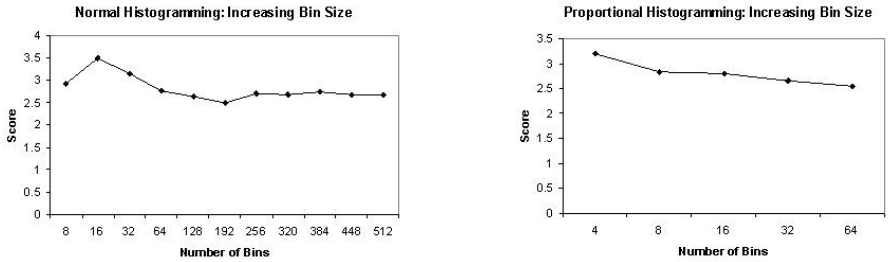

Figure 2: Point scores for the average of all the tests for the varying number of bins and for both the normal (left) and the proportional (right) histogramming methods.

I(i) is the ith bin of image I and J(i) is the ith bin of image J. As mentioned earlier, the chi-square method is fairly simple, but has been shown to work fairly well in a wide variety or circumstances.

For simplicity, the method above was only described with respect to a single input song. However, this method can easily be modified in a number of different ways to allow multiple songs to be used as input. The modification can happen in two stages, either the histogramming stage or the actual decision stage (i.e. after chi-square value is computed). In the histogramming stage, multiple input songs can simply be treated as a single song and integrated into a single histogram. Since the songs are normalized before comparison, the increase in number of coefficients will not affect the result. If the songs are histogrammed independently, they can still all be compared against each song in the database and then combine the scores in some manner to create an overall score for each song in the database (such as summing the scores together or a sum of squares to avoid outliers).

3 Experimental Setup and Data Set

The basic problem attacked by this paper is audio retrieval. Given a database of audio files and a number of input audio files, retrieve some number of audio files from the database that are similar to the input files. This framework has been applied on a wide range of data sets ranging from sound effects to full songs ([4] [11]). One common class separation used when dealing with actual songs is the genre of the song (such as classical, rap, rock, jazz, etc.). There have been a number of successes on this type of data set, however, the data set does not provide an uncontroversial ground truth.

The approach taken in this paper is slightly different. Instead of trying to recognize the genre of the input songs, we try and recognize the band or artist of the input songs. The main reason that this type of approach is better than a genre approach is that it provides a known ground truth for measurement purposes. This has been difficult to accomplish in both the realm of audio and image processing, but is an important factor in comparing algorithms and calculating the performance of these systems.

Given the size of most songs and the processing required to handle such a file, the database that we developed is relatively small. However, we feel that it still provides a good start. The database consists of 40 songs from 4 different bands: Dave Mathews Band, U2, Green Day and Blink 182 (10 songs from each band). The artists were chosen in a hope to present a problem that may be difficult for a person who is unfamiliar with popular music to accomplish. Although there are obvious discernable differences between these bands, likely a human might have

	1st	2nd	3rd	4th	5th	Score
Dave Mathews	.6	.8	.4	.3	.2	8.2
Blink 182	.3	.1	.1	0	.1	2.3
U2	0	0	0	.1	0	.2
Green Day	.2	.3	.2	0	.5	3.3
Average	.275	.3	.175	.1	.2	3.5

Table 1: Percentage correct for normal histogramming and 16 bins. The score and average percentage correct is also presented.

trouble with this problem. For computational purposes, the songs were all down sampled to 22 KHz and reduced to monophonic sound.

To test a wide range of parameters, a number of different bin sizes were used and the two different histogramming methods described above (normal and proportional) were used. For the normal histogramming method bins of size 8, 16, 32, 64, 128, 192, 256, 320, 384, 448 and 512 were used and for the proportional method, bins of size 4, 8, 16, 32 and 64 were used. The input consisted of 5 songs from a single band and the goal was to retrieve the other 5 songs of that band in the remaining database. For each band, 10 random samples of 5 songs were tested and the average was taken.

4 Results

A number of different rating methodologies are presented here to try and better examine the results. The first measure is similar to measures that have been used in image retrieval papers [3]. The first answer is given a score of 5 points, 4 for the second, 3 for the third, etc. For a query, the sum of these is taken, resulting in a perfect score of 5+4+3+2+1 = 15. A second measure is simply the percentage correct at each precision. This figure helps to get a good idea for where the correct answers are occurring. This measure is more precise than the point measure, but does not provide a good general view of the results.

A summary of the results can be seen in Figure 2. A few things should be noticed from these figures. First, note that, unfortunately, the average for both normal and proportional methods is around 3. Given that the best score that a perfect score is 15, 3 is not the most desirable score. Also, notice that for both methods, as the number of bins increases, the performance of the system does not really change. This is an interesting phenomenon. The actual choice of bin size seems to make little retrieval performance impact.

Table 1 presents answers for 16 bins. In general, the correct songs seemed to be located throughout the various query positions (as is seen in this example). Another interesting phenomenon also presents itself in this example. Notice that Dave Mathews Band performs considerably better than the other two algorithms. And there appears to be a difference between the other three also. Table 2 examines this difference a bit closer. Notice that there is a consistent difference between the performances of the four bands. This effect is somewhat surprising.

	Normal	Proportional
Dave Mathews	6.9	5.8
Blink 182	1.3	2
U2	.9	1.5
Green Day	2.1	2
Average	2.8	2.8

Table 2: Summary of point scores over the four bands for the two different methods tested.

bins	1	2	3	4	5
4	0.2	0.375	0.625	0.65	0.75
8	0.225	0.375	0.525	0.55	0.625
16	0.175	0.4	0.55	0.55	0.575
32	0.125	0.375	0.45	0.5	0.575
64	0.075	0.3	0.45	0.5	0.575

Table 3: Percentage of results that have at least one correct answer at that position or earlier for the proportional method.

If we analyze the results a little closer we will notice that not only does Dave Mathews Band appear to have better performance, but in fact, one of the reason for this better performance is that Dave Mathews Band appears to be selected more often as a result than other bands. Further analysis of this effect will hopefully reveal the cause.

Even though the results appear to be somewhat disappointing, one last result should be presented. Table 3 presents an interesting look at the recall of the system as the number of positions is increased. Notice that this is not the actual recall of the system, instead it is a measure of how often you'd get *at least one* result in those positions or higher. Given this framework, the results are not as bad as originally perceived. The results for the proportional method are only provided, but the results for the normal method are similar. Notice that the fifth entry shows the number of tests where there was at least one result recalled at all. Generally speaking, at least 60% of the time a correct answer at least appeared and, in the best case, a correct answer was in the top 5 75% of the time.

5 Conclusion and Future Research

The direction for future research in this area should be fairly clear at this point. Obviously we would still like to investigate a number of different parameters for the algorithm above, such as using adaptive binning instead of the simplistic binning used now. However, more importantly, as suggested numerous times, we would like to further investigate other image retrieval techniques, particularly those methods that do feature selection (since there has been little of this done in the audio domain, for example [12]).

The results from the experiments were not outstanding. However, they were not abysmal. The results were at least better than random. Although maybe optimal, 75% of the time a correct answer would appear in the top 5. To the credit of the

system, this is a problem that has not been previously analyzed. This type of band recognition is even difficult for human observers, so it is not surprising that we might have difficulty approximating this. Finally, we have no idea how hard this problem actually is. As we examine this further and more research is done using this as a test set, this should become more clear.

In this paper we have examined the problem of audio retrieval. Specifically, we examined the problem of trying to retrieve songs with the same artist as the input songs. This problem has the nice advantage that a known ground truth is known and therefore provides a stable testing platform for analysis. Instead of developing new techniques from scratch, this paper makes a strong suggestion to borrow results from other fields, particularly image retrieval, where the problem is similar and the has been investigated in more depth and detail than audio retrieval. I want to reiterate this point again to make sure that the goal of this paper is clear. We did not plan necessarily to find the best solution out there, what we did hope to do is to suggest to other researchers to examine the methods of image retrieval, borrow ideas and methodologies and apply those to audio processing. Doing this will save researchers time and effort and will allow the audio domain to benefit from similar research already conducted.

Acknowledgments

I'd like to thank Serge Belongie for his support and ideas along the way. I'd also like to thank the cs291 class for their helpful comments and feedback (I think you guys know who you are).

References

[1] Dannenberg, R., Thom, B. & Watson, D. A Machine Learning Approach to Musical Style Recognition.

[2] De Bonet, J. & Viola, P. (1997). Structure Driven Image Database Retrieval. Neural Information Processing 10.

[3] Faloutsos, C., Equitz, W., Flickner, M., Niblack, M., Petkovic, D. & Barber, R. (1994). Efficient and Effective Querying by Image Content. Journal of Intelligent Information Systems.

[4] Ghias, A., Logan, J., Chamberlin, D. & Smith, B. (1995). Query By Humming: Musical Information Retrieval in An Audio Database.

[5] McDonagh, J. & Smeaton, A. Multimedia Information Retrieval: MIDI as a format for Content Based Retrieval of Audio.

[6] Muslea, I. (1999). Extraction Patterns for Information Extraction: A Survey. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 1999.

[7] Piamsa-NGA, P., Subramanya, S., Alexandridis, N., Srakaew, S., Blankenship, G., Papakonstantinou, G., Tsanakas, G. & Tzafestas, S. (1998). Content-Based Audio Retrieval Using A Generalized Algorithm. Advanges in Intelligent Systems: Concepts, Tools, and Applications, Klewer Academic.

[8] Puzicha, J., Buhmann, J., Rubner, Y. & Tomasi, C. (1999). Empirical Evaluation of Dissimilarity Measures for Color and Texture.

[9] Soltau, H., Schultz, T., Westphal, M. & Waibel, A. Recognition of Music Types.

[10] Subramanya, S., Youssef, A., Bhagirath, Simha, R. Automated Classification of Audio Data and Retrieval Based on Audio Classes.

[11] Tzanetakis, G., Essl, G., Cook, P. (2001). Automatic Musical Genre Classification of Audio Signals.

[12] Viola, P. & Jones, M. (2001). Rapid Oject Detection using Boosted Cascade of Simple Features.

[13] Wang, Y., Liu, Z. & Huang, J. (2000). Multimedia Content Analysis: Using Both Audio and Visual Clues.