AN EMPIRICAL COMPARISON OF THREE AUDIO FINGERPRINTING METHODS IN MUSIC AND FEATURE-LENGTH FILM

Thanh Pham, Matthew Giamou, and Gerald Penn
Dept. of Computer Science, University of Toronto; Toronto, Ontario, Canada, L4J 7P7
{mpham,mattgiamou,gpenn}@cs.utoronto.ca

1. INTRODUCTION

A classic problem in content-based audio retrieval is to identify the title of a song in a song database given a short audio query from one of the songs. Many techniques have been proposed to solve this problem (*inter alia*, Li et al., 2008; Mapelli and Lancini, 2003). A more recent trend has been to construct audio fingerprints by applying computer vision techniques to spectrogram images from short-term Fourier transforms of the audio data, rather than from the audio data directly (Ke et al., 2005). This work, however, generally takes the form of a proof of concept that the proposed method works at all, rather than a comparison with an appropriate non-vision-based baseline.

We have chosen one commercially very successful algorithm, Shazam (Wang, 2003), and compared it against two vision-based algorithms, the original CMU algorithm (Ke et al., 2005), and also Google's Waveprint algorithm (Baluja and Covell, 2008), an improvement upon the CMU algorithm which introduced the use of wavelet features as well as a few innovative hashing techniques to improve the CMU algorithm's time efficiency. We evaluate these three on two different datasets: one of 6000 proprietary CD-quality songs with an average length of 228 seconds, ranging from 4.4 seconds to 2051 seconds, with 12000 queries of 30 seconds each, the other of the mixed audio layers (talking, music, sound effects, pauses etc.) of 223 feature-length Hollywood films, together with 6690 queries of 30 seconds each. The task is to identify the song or film, respectively, of each of the queries corresponding to the respective dataset in turn. There are no negative answers --- every query occurs in exactly one song/film.

2. DATASETS

The feature-length film dataset was transcoded and down-sampled from 48KHz to 16KHz mono-channel PCM with 256 kbps bitrates. The dataset contains all of the popular movie genres. The video data were discarded and never used. On average, each film lasts about 6960 seconds, with lengths ranging from 2118 seconds to 12444 seconds.

From each transcoded film, we sampled 30 different 30-second queries at random positions. To emulate the audio of pirated films recorded in theatres, we re-recorded the queries using an inexpensive omni-directional microphone that has a frequency response from 50Hz to 13KHz, an impedance of 650 ohms and a sensitivity of -58dB +/- 3dB at 1 KHz. All the recordings took place in the same room, which had a baseline noise level of -42db. In total, we gathered 30*223=6690 queries.

The music dataset was sampled from 450 licensed audio CDs, with variety of genres and singers. Each piece was down-sampled to 11025Hz mono-channel with 352kb/s bitrates. From each down-sampled piece, we extracted two 30-second queries from random positions, for a total of 12000 queries. Each audio query was passed through an MP3 encoder, then an MP3 decoder, to introduce noise.

3. EXPERIMENTS AND RESULTS

For the music data experiment, we used 1000 songs and their corresponding 2000 audio queries for the threshold tuning. The remaining 5000 songs and their corresponding 10000 queries were used for evaluation testing only. For the film experiment, we used 100 films and their corresponding 3000 audio queries for threshold tuning, and the remaining 123 films and 3690 queries for testing. The experiments were conducted on a machine with a single 3.0GHz Intel Xeon CPU with a 4MB cache and 16GB RAM.

A comparison of per-query execution speed is shown in Table 1. The F-measures on the two datasets are shown in Tables 2 and 3. In brief, Waveprint's optimizations for speed pay very high dividends on the music data, but not on film audio, and its optimizations for quality pay very high dividends on film audio, but not on music data. Nevertheless, Shazam handily outperforms both vision-based algorithms, in both time and quality.

Dataset	Shazam	Waveprint	CMU
Music	Is	6s	21s
Movie	4s	89s	11s

Table 1. Speed Comparison per query on testing datasets.

3.1 Discussion of Speed Performance

Waveprint spends most of its time in the full-comparison and wavelet decomposition steps. To fingerprint a short segment of an audio snippet, the algorithm needs to perform multi-level wavelet decompositions on its spectral image. For silent or short pause moments, the corresponding spectral images usually have very low energy, and thus are highly similar to each other across all films. For these moments, the algorithm must spend extra time performing full comparisons. To achieve the top speed, we did not include dynamic temporal warping (DTW) in our implementation. The authors indicate that non-DTW system accuracy was only 0.76% faster.

Algorithm	Precision	Recall	F-measure
Shazam	0.9836	0.9990	0.9913
CMU	0.9631	0.9940	0.9786
Waveprint	0.9020	0.9952	0.9463

Table 2. Comparison on 5000 songs and 10000 queries.

Unlike Waveprint, instead of performing a full comparison between 2 spectral images, the original CMU algorithm only compares two descriptors extracted from the two spectral images. This step is implemented by applying direct indexing on descriptors extracted from spectral images. To cope with noise, the CMU algorithm sacrifices speed by including an EM temporal model and extra descriptors within a hamming distance of 2 of the original descriptors.

The Shazam algorithm is the fastest among the three because of its efficient hash signature design. Each hash is associated with an identified instance, so for each query call, matched hashes of every instance in the database can be quickly extracted by looking only once in the hash table. By making the strong assumption that distorted noise is linear, the algorithm forsakes DTW and simply computes the time difference of 2 anchor points coupled from each matched hash to accumulate scores across time. This score accumulation process takes O(n log n) in our implementation with n the number of matched hashes.

3.2 Discussion of F-measure performance

To account for the relative quality of Shazam's performance, we examined some of our audio queries in which there were: (1) lots of speech, (2) low-noise backgrounds, (3) pauses between speaker turns, or (4) silent moments. We also analyzed spectral images in these queries and the corresponding images reconstructed from their top 200 wavelets. In these reconstructed images, we can notice that a lot of important information from speech is neglected at high frequency bands by the CMU and Waveprint algorithms. This is due to a low-pass filtering stage with cut-off frequencies at 2kHz. The fundamental frequencies of musical instruments are usually below 2khz, and it is reasonable (or, at least, more reasonable than with speech) to assume that the important information above 2kHz derives primarily from harmonics of the fundamental frequency. Speech is different, i.e., it is more bursty, and has many different formants. By applying low-pass filters at 2khz to speech signals, the CMU algorithm and Waveprint neglect distinctive information of these higher formants. Increasing the cut-off frequency to an appropriately higher value such as 8khz would address this in principle, but was found to be computationally so inefficient as not to be testable. A literary analysis of the choice of 2kHz traces back to the original work of Haitsma and Kalker (2002), in which it is apparently selected because the ``the range from 300Hz to 2000Hz [is] the most relevant spectral range for the human auditory system.'' In terms of classifier performance, focussing on this range does not work well.

Also, the film dataset contains high energy at low frequencies due to low frequency background noises and our microphone. The Waveprint algorithm extracts the top

Algorithm	Precision	Recall	F-measure
Shazam	0.9862	0.9871	0.9867
CMU	0.2586	0.4650	0.3324
Waveprint	0.6279	0.9716	0.7628

Table 3. Comparison on 123 movies and 3690 queries.

200 wavelets by magnitude, thus, if the top wavelets happen to lie at low frequency bands its fingerprints are not likely to have enough distinctive power. Low noise background, pauses and silent moments do not improve the F-measure, and in fact, they could worsen the performance because they occur in most films, leading to many false positive hits.

The CMU algorithm's low precision could be due to using only descriptors within a hamming distance of 2 of the original descriptors.

4. CONCLUSION AND FUTURE WORK

Our comparison shows that the provenience of the audio data (music vs. voice and background noise from films) is enough to foil even simple design decisions that some algorithms (notably the two vision-based algorithms) use to improve speed and/or F-measure performance in other domains. Also, while these vision-inspired approaches are indeed competitive with Shazam on music data, the same cannot be said about their performance on film data. The F-measures are not even close. The spectral characteristics of the data seem to be responsible for this in great part.

It is also interesting that Shazam can perform so well on film identification --- with only the audio layer --- even though it was designed for music identification.

REFERENCES

Baluja, S. and M. Covell (2008). Waveprint: efficient wavelet-based audio fingerprinting. *Pattern Recognition* 41(11), 3467-3480.

Haitsma, J. and Kalker, T. (2002). A highly robust audio fingerprinting system. Proc. Intl. Sym. Music Info. Retrieval.

Ke, Y., D. Hoiem and R. Sukthankar (2005). Computer vision and music identification. Proc. Comp. Vision and Pat. Rec., 597-604.

Mapelli, F. and R. Lancini (2003). Audio hashing technique for automatic song identification. Proc. Intl. Conf. on Info. Tech., Research and Education, 84-88.

Li, Q. and Wu, J. and He, X. (2008). Content-based audio retrieval using perceptual hash. Proc. Intelligent Info. Hiding and Multimedia Sig. Proc., 791-794.

Wang, A. L.-C. (2003). An industrial-strength audio search algorithm. Proc. Intl. Sym. Music Info. Retrieval.

ACKNOWLEDGEMENTS

This research was funded by the GRAND NCE as part of the NEWS project.