INTERACTIVE VIDEO RETRIEVAL USING EMBEDDED AUDIO CONTENT

Tahir Amin, Mehmet Zeytinoglu and Ling Guan

Department of Electrical and Computer Engineering
Ryerson University, Toronto, Ontario, M5B 2K3, Canada
tamin, mzeytin, lguan@ee.ryerson.ca

Qin Zhang
Sino-German Joint Software Research Institute
Beijing, China 100000
zhangq@bsw.gov.cn

ABSTRACT

Audio is a rich source of information in the digital videos that can provide useful descriptors for indexing the video databases. In this paper, we model the shape of the distribution of wavelet coefficients of embedded audio with a Laplacian mixture. The distributions of wavelet coefficients are very peaky in nature. The shape of these distributions can be modeled with only two components in the Laplacian mixture with low computational complexity. The parameters of this mixture model form a low dimensional feature vector representing global similarity of the audio content of the video clips. An interactive approach involving the feature vector updating scheme is used to adapt the retrieval system to the users' needs. This Relevance Feedback (RF) increases the retrieval ratio substantially. A comprehensive experimental evaluation using the CNN news database has been performed.

1. INTRODUCTION

The advances in the digital and network technology have produced a flood of multimedia information. A large amount of multimedia content consists of digital videos, giving rise to an unprecedented high volume of data. The provision of an interactive access to this huge digital video information repository currently occupies researchers' minds in several fields. The visual information is traditionally used for video indexing. Here, we consider using embedded audio because it's a rich source of content-based information. Users of digital video are often interested in certain action sequences. While visual information may not yield useful "action" indexes, the audio information often directly reflects what is happening in the scenes and distinguishes the actions.

Although image-based approaches are common, a few studies have also considered audio analysis. However, it remains an area of basic research. Since we must deal with mixed sound sources, existing speech-recognition algorithms generally don't suit ordinary videos. Most studies on music and speech detection are aimed at improving speech recognition systems. The difficulty in handling mixed audio sources has, until now, hindered the use of audio information in handling video. Therefore, few have attempted to deal with the type of videos we come across in everyday situations. In this paper, a televised newscast from CNN including commercials, anchor person clips, reporters and environmental/machine noise excerpts is used as subject because it contains various mixed audio sources.

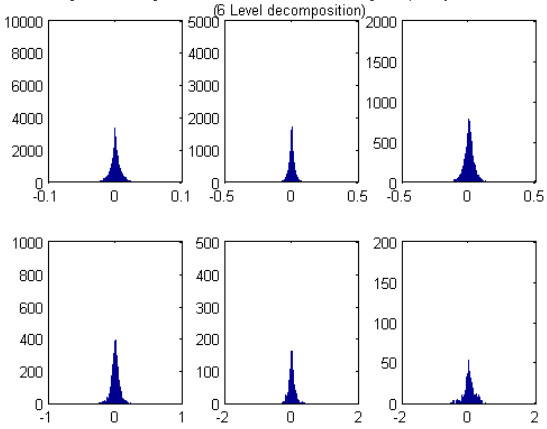
Wavelet coefficient distributions are very peaky in nature due to their energy packing property. This type of peaky distributions is usually modeled using mixture of Gaussians. We can model any arbitrary shaped distribution using mixture of Gaussians if we have an infinite number of components in the mixture. Modeling the wavelet coefficients is a typical example in which a large number of components in the Gaussian mixture may be required to catch the peakedness in the distribution. This is however practically infeasible. In Figure 1 we plot the histograms of wavelet coefficients at different scales for the embedded audio information in a video clip. The peaky nature of the distributions is clearly observed from this figure.

Taking into account the peaky distributions of wavelet coefficients, they can be modeled with only a few components in the Laplacian mixture. Here we use a mixture of only two Laplacians for modeling the shape of the distribution. The model parameters of this mixture are used as features for indexing the video clips based on the audio information only. It is observed that the resulting features possess high discriminatory power for classification of the video clips into different classes. The low dimensionality of the resulting feature vector makes

the system response very fast enhancing the user experience while interacting with the system.

The semantic gap between the low level feature representation and the high level similarity as perceived by human auditory system contributes to low performance. Although the discriminatory power of the features is highly important for an efficient retrieval system yet it is very difficult to model the human notion video similarity with the low level features alone. This serves as a motivation for implementation of relevance feedback (RF) to learn the user requirements and adapt the system accordingly. We have shown the power of RF based on the simple feature vector updating scheme.

Figure 1: Histograms of wavelet coefficients in high frequency subbands (6 Level decomposition)



2. RELATED WORK

In the first step, the digital video is decomposed into smaller structural units called shots, scenes and clips by video parsing algorithms. In [1], Arman et al. proposed to automatically extract a reference frame from each shot segment to facilitate efficient video browsing comparable to the fast-forward and fast-rewind functionality of a conventional video cassette player. Naphade and Huang attempted the fusion of multimodal features such as visual and audio in [2]. Saunders classified the audio into speech and music based on zero crossing rate and audio energy [3]. Wold et al. classified the audio into 10 different classes [4]. The techniques to classify the audio or the embedded audio may not be useful for the retrieval of news video. In this type of application, music, speech, noise and crowd voice may be found together in the same video clip. Hence we need features that represent the global similarity of the audio content.

3. THE PROPOSED METHOD

3.1. Feature Extraction

The video is segmented manually into small clips and then is demultiplexed to separate the embedded audio. We compute the 1-D Discrete Wavelet Transform of all the audio clips using Daubechies-2 (db2) and Daubechies-4 (db4) wavelet kernels. The 1-D wavelet transform decomposes the clips into 2 subbands at each wavelet scale representing the low frequency and high frequency information of the audio segments.

A statistical model based on the Laplacian mixture is developed to catch the peaky shape of the wavelet coefficient distributions. It is noted that the parameters of this model are a good representation of the global content of the audio content. We model the wavelet coefficients in each wavelet subband as a mixture of two Laplacians:

$$P(w_i) = P_s . l(w_i, 0, b_s) + P_l . l(w_i, 0, b_l) \quad (1)$$

$$P_s + P_l = 1 \quad (2)$$

where the class of small coefficients is represented by subscript "*s*" and the class of large coefficients by subscript "*l*". P_s and P_l are the *a priori* probabilities of the two classes. The Laplacian component:

$$l(w_i, 0, b_s) = \frac{1}{2b_s} exp\left(-\frac{|w_i|}{b_s}\right) \quad (3)$$

corresponding to the class of small coefficients has relatively small value of parameter *b*. The shape of the Laplacian distribution is determined by this single parameter *b*. The value of this parameter is a good representation of the contents of the audio clip.

EM algorithm is applied to estimate the parameters of the model [5]. The EM algorithm is iterative and consists of two steps, E-step and M-step, for each iteration. For the *n-th* iterative cycle, the E-step computes two probabilities for each wavelet coefficient:

$$P_{si} = \frac{P_s(n) l(w_i, 0, b_s(n))}{P_s(n) l(w_i, 0, b_s(n)) + P_l(n) l(w_i, 0, b_l(n))} \quad (4)$$

$$P_{li} = \frac{P_l(n) l(w_i, 0, b_l(n))}{P_s(n) l(w_i, 0, b_s(n)) + P_l(n) l(w_i, 0, b_l(n))} \quad (5)$$

In the M-step, the parameters *[bs, bl]* and *a priori* probabilities *[Ps, Pl]* are updated.

$$P_s(n+1) = \frac{1}{K} \sum_{i=1}^{K} P_{si}(n) \quad (6)$$

$$P_l(n+1) = \frac{1}{K} \sum_{i=1}^{K} P_{li}(n): \quad (7)$$

$$b_s(n+1) = \frac{\sum_{i=1}^{K} |w_i| P_{si}(n)}{K P_s(n+1)} \quad (8)$$

$$b_l(n+1) = \frac{\sum_{i=1}^{K} |w_i| P_{li}(n)}{K P_l(n+1)} \quad (9)$$

All the audio clips are decomposed using 1-D wavelet transform. The EM algorithm is applied to the detailed sub-bands at each wavelet scale. The model parameters *[P_l b_S, bl]* calculated for each subband are used as features. The mean and variance of the wavelet coefficients in the approximate subband are also chosen as features.

3.2. Normalization

The value of each component in the feature vector has different dynamic range because it represents a different physical quantity. Therefore the features are normalized before the application of the similarity measure. The normalization process puts equal emphasis to each component of the feature vector. We use the Gaussian normalization [6]. If we take a sequence V to be a Gaussian sequence, we can compute the mean μ and standard deviation σ of the sequence. We then normalize the original sequence as follows:

$$v = \frac{v - \mu}{\sigma} \qquad (10)$$

3.3. Similarity Measure

After normalization of the features, the Euclidian distance measure or L2-norm is used in the initial search setting equal weight to all the features.

$$d(\boldsymbol{x}, \boldsymbol{q}) = \sum_{i=1}^{N} B_i \sqrt{(x_i - q_i)^2} \qquad (11)$$

where $\boldsymbol{x}$ and $\boldsymbol{q}$ are the feature vectors of the audio clip and the query clip respectively. The $\boldsymbol{B}$ is a weighting factor based on the relevance of the feature to the perceptual similarity measure. The criterion to select the value of B is discussed in the following section.

3.4. Feature Weights

Let the training set of K images be $T = (x_k, y_k)_{k=1}^{K}$, where $\boldsymbol{x}_K$ denotes the feature vector of the k-th image and y_K is the label of the image given by the user. The value of the label is 1 for the relevant images and 0 for the non-relevant images. Now we form two matrices $R = [x_{mi}]$ and $D = [z_{ni}]$ of dimension M x P and N x P respectively, where $m = 1,...,M$, $n = 1,...N$. Here R is the matrix containing the feature vectors of video clips labeled as relevant *($y = 1$)* and D is the matrix containing the feature vectors of clips marked as irrelevant *($y = 0$)*. M is the number of relevant video clips and N is the number of irrelevant clips. P is the dimension of the feature vector and $i = 1,..., P$.

The density of a feature x_{mi} around q_i is related to the relevancy of the i-th feature, which is inversely proportional to the length of the interval. A large density usually indicates high relevancy for a particular feature, while a low density implies that the corresponding feature is not critical to the similarity characterization. Hence we update the weighting factors B_i in equation 11 based on the standard deviation method:

$$B_i = \frac{1}{\sigma_i} \qquad (12)$$

where σ_i is the standard deviation of the i-th feature of the video clip labeled as relevant by the user.

4. EXPERIMENTAL RESULTS

A comprehensive experimental evaluation has been performed using CNN news video database. The database used in these experiments consists of 302 video clips of around 3 second duration each. The video has been segmented manually. The video clips thus created are then demultiplexed to separate the audio and visual information. These video clips have been classified into 5 broad classes representing a typical news broadcast. The 5 classes are Male Anchor person, Male reporter, Female reporter, Environmental Noise (such as plane and car sounds) and commercials. The detailed description of the database is given in the following table 1. The commercials category consists of a variety of video segments containing male/female voices, music, noise and crowd voice. The retrieval ratio is calculated for each of the class as well as the overall ratio for the whole database.

Table 1: Number of Clips in each class

Anch (M)	Rep (M)	Rep (Fe)	Noise	Comm
47	79	63	14	97

In Table 2, we summarize the results obtained by performing a 5-level decomposition of audio clips using db2 wavelet kernel. The results obtained by decomposing the audio clips upto 7-level using db2 are given in Table 3. These results indicate that the retrieval ratio is improved in case of 7-level decomposition. This improvement is achieved because more features are used for indexing. The dimension of the feature vector is 17 in case of 5 level decomposition and 23 in case of 7-level decomposition. The overall retrieval ratio is increased from 70% to 77.2% by taking more features. The improvement is very significant, especially in the case of *Noise* class where the retrieval ratio is increased from 64.4% to 83.7%.

The results also emphasize the importance of the relevance feedback in improving the accuracy of the system. It's observed that the retrieval ratio improves significantly after first iteration when it is increased from 59.1% to 68.7%. However we observe very slight

improvement after the 2nd and 3rd iterations. In Figure 2, we have plotted the retrieval ratio after each of the iteration. These plots become almost flat after the 2nd iteration indicating that the system has achieved its optimum retrieval performance.

Table 2: Average retrieval rate (%) for top 16 video clips retrieved (5-level decomposition using db2)

Cat./ Iter.	Anch (M)	Rep (M)	Rep (Fe)	Noise	Comm	Over-all
Initial	52.2	60.2	50.4	61.5	71.2	59.1
Iter 1	67.2	68.5	59.1	64.4	81.7	68.7
Iter 2	69.7	70.5	60.4	64.4	83.3	69.7
Iter 3	70	70.6	61.1	64.4	83.6	69.9
Iter 4	70	70.7	61.1	64.4	83.7	70

Table 3: Average retrieval rate (%) for top 16 video clips retrieved (7-level decomposition using db2)

Cat./ Iter.	Anch (M)	Rep (M)	Rep (Fe)	Noise	Comm	Over-all
Initial	53.1	60.8	46.1	79.8	70.6	62.1
Iter 1	69.1	78	64.7	83.7	83.2	75.7
Iter 2	70	79.7	67.1	83.7	84.7	77
Iter 3	70.3	79.8	67.3	83.7	84.8	77.2
Iter 4	70.3	79.8	67.3	83.7	84.8	77.2

In Tables 4 and 5 below, we have presented the results using the db4 wavelet kernel for decomposition of embedded audio clips. It is observed that db4 wavelet kernel performs better than the db2 wavelets. The overall retrieval ratio after 4th iteration is 79.6% in case of 7-level decomposition using db4 wavelets. We observe an overall improvement of 1.1% in case of 5-level decomposition. An overall improvement of 2.4% in results is attained with 7-level decomposition using db4 wavelet kernel in comparison with db2 wavelets.

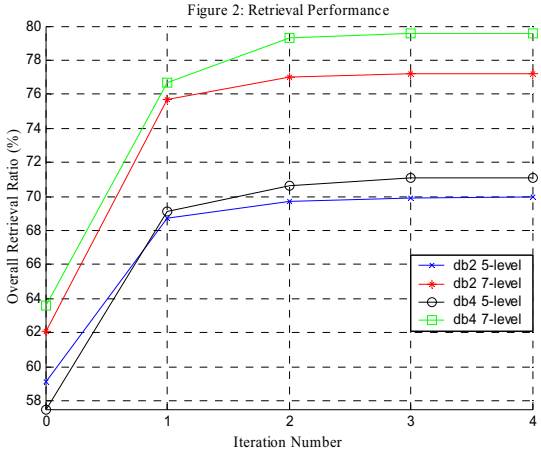
Table 4: Average retrieval rate (%) for top 16 video clips retrieved (5-level decomposition using db4)

Cat./ Iter.	Anch (M)	Rep (M)	Rep (Fe)	Noise	Comm	Over-all
Initial	51.8	63.5	45.5	57.2	69.3	57.5
Iter 1	70.3	79.4	54.5	58.2	82.9	69.1
Iter 2	72.5	81.6	56.9	58.2	83.9	70.6
Iter 3	73.1	82.2	57.4	58.2	84.5	71.1
Iter 4	73.1	82.2	57.4	58.2	84.6	71.1

Table 5: Average retrieval rate (%) for top 16 video clips retrieved (7-level decomposition using db4)

Cat./ Iter.	Anch (M)	Rep (M)	Rep (Fe)	Noise	Comm	Over-all
Initial	56.6	66.4	45	79.8	70.1	63.6
Iter 1	71.9	84.9	64.7	80.3	82	76.7
Iter 2	75	88	69.5	80.3	83.5	79.3
Iter 3	75.3	88.4	70.1	80.3	83.9	79.6
Iter 4	75.3	88.4	70.2	80.3	84	79.6

Figure 2: Retrieval Performance



5. CONCLUSION

We presented a new feature extraction method for video retrieval based on the embedded audio content. The video clips are indexed using a low dimensional feature vector that is a good representation of the global similarity of the audio contents. We demonstrate the ability of the embedded audio content of the digital video for searching the databases based on the auditory information. This may be particularly useful for finding the action sequences in the videos. A comprehensive experimental evaluation of the system is presented using a news video database.

6. REFERENCES

[1] F. Arman, R. Depommier, A. Hsu, and M.Y. Chiu: *Content-based Browsing of Video Sequences*, Proc. 2nd ACM intl. conference on Multimedia, 1994, pp. 97-103.

[2] M.R. Naphade, T. Kristjansson, B. Frey, and T.S. Huang, *Probabilistic Multimedia Objects (Multijects): A Novel Approach to Video Indexing and Retrieval in Multimedia Systems*, Proc. 1998 International Conference on Image Processing, Volume: 3, 1998, pp. 536-540.

[3] J. Saunders, *Real-Time Discrimination of Broadcast Speech /Music*, ICASSP 1996, vol. 2, pp. 993-996, Atlanta, May 1996.

[4] E. Wold, T. Blum, D. Keislar and J. Wheaton, *Content-Based Classification, search and Retrieval of Audio,*IEEE Multimedia, vol.3, No. 3, pp. 27-36, Fall 1996.

[5] J. Bilmes, "*A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models"*, Technical Report ICSI-TR-97-021, University of Berkeley (1998).

[6] Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra, *Relevance feedback: a power tool for interactive content-based image retrieval*, IEEE Transactions on Circuits and Systems for Video Technology, vol. 8, pp. 644--655, 1998.