Design of a Content Based Multimedia Retrieval System

YUK YING CHUNG, KA YEE NG, LIWEI LIU, XIAOMING CHEN, MOHD AFIZI MOHD SHUKRAN
*ERIC H.C.CHOW, *DAVID YU SHI, *FANG CHEN
School of Information Technologies, University of Sydney, NSW 2006, AUSTRALIA
* Multimodal User Interface Program, National ICT for Australia, NSW 1435, AUSTRALIA

Abstract: - The large and growing amount of digital data, and the development of the Internet highlight the need to develop sophisticated access methods that provide more than just simple text-based queries. Many programs have been developed with complex mathematics algorithms to allow the transformation of image or audio data in a way that enhances searching accuracy. However it becomes difficult when dealing with large sets of multimedia data. This paper proposes and demonstrates a Content Based Multimedia Retrieval System (CBMRS). The proposed CBMRS includes both video and audio retrieval systems. The Content Based Video Retrieval System (CBVRS) based on DCT and clustering algorithms. The audio retrieval system based on Mel-Frequency Cepstral Coefficients (MFCCs), the Dynamic Time Warping (DTW) algorithm and the Nearest Neighbor (NN) rule.

Key-Words: - video retrieval, audio retrieval, DCT, wavelet, MFCC, DTW, NN, clustering, greedy tree

1 Introduction

Multimedia Retrieval is one of the most important and fastest growing research areas in the field of multimedia technology. Large collections of scientific, artistic and commercial data comprising image, text, audio and video abound in the present information based society. There must be an effective and precise method of assisting users to search, browse and interact with these collections. In this paper propose and demonstrates a content based video and audio retrieval system. The Content Based Video Retrieval System (CBVRS) is based on DCT and clustering algorithms. The Content Based Audio Retrieval System (CBARS) is based on Mel-Frequency Cepstral Coefficients (MFCCs), the Dynamic Time Warping (DTW) algorithm and the Nearest Neighbor (NN) rule.

Content-Based Video Retrieval Systems (CBVRS) allow similar features to effectively characterise the content of images and then uses such features in the retrieval process. Recent studies have proved that features such as colour, texture, shape and spatial position indeed possess a very high semantic value for image or video retrieval systems [1,2,3]. Such systems represent each image as a set of feature vectors which allows the assessment of similarity between images. IBM QBIC system [1] is an example of using colour, texture and shape as feature vectors.

Huge amounts of multimedia data can now be stored because of the advance of technology and the low cost of hard disk. The major difficulty is the efficient locate images and video frame within huge databases. A searching process would be time-consuming if it needs to search on the whole database which may contain one million of pictures and video. This paper describes the implementations of some clustering algorithms in the proposed CBVRS.

The general information retrieval technology is based on text. Yahoo and Alta Vista are two typical text-based search engines. They both use text-based keywords to search for the information. The engine will locate all the documents which contain the keywords. After that, the degree of match depends on how many keywords in a search option are supplied by a document. The matching degree is higher if a document has more matching keywords. Besides text files, audio and other multimedia files with definitions of text attributes (e.g. id, format, etc.) can be retrieved by this text based searching method. However, most software applications and digital audio frequency are always managed as an opaque stream. There is no defined word or text-based object for comparisons. Thus, a more effective approach to manage the audio retrieval is needed.

Content Based Audio Retrieval System (CBARS) refers to searching for audio data in a database to retrieve a small number of audio files that possess similarity to the user input. The user can input an audio file instead of the traditional query options, keywords of a text. In order to achieve content-based audio retrieval, automated methods are developed to recognize the composition and content of the audio files, and to search the audio data based on the recognized content of the query samples.
Compared to text-based audio retrieval, CBARS is rather general. Text-based audio retrieval relies on how the human operators input the descriptive text for audio files, which is based on the human perception of the audio files. CBARS systems are more flexible in audio retrieval and are more readily adaptable to query variations. In this paper the proposed CBARS is based on Mel-Frequency Cepstral Coefficients (MFCCs), the Dynamic Time Warping (DTW) algorithm and the Nearest Neighbor (NN) rule.

Section 2 describes the proposed Content Based Video Retrieval System (CBVRS). Section 3 explains the proposed Content Based Audio Retrieval System (CBARS). Sections 4 and 5 presents the experimental results and conclusion respectively.

2 Content Based Video Retrieval Systems (CBVRS)

Nowadays people not only use pure images for daily purposes, video is also a popular media for recording TV, diaries etc. As a consequence, effective methods for searching with large database are needed. This suggests the need for using Clustering Based Image Retrieval System for images or videos which allow users to search images or particular image frames according to their preferences.

The proposed Clustering Based Image Retrieval System (CBIRS) comprises three components to obtain desired images or frames efficiently. They are colour histogram, clustering process and searching process. The colour features of each image or video frame are represented by a colour histogram saved in the database. These features define the characteristics of images which are relevant to the database. As a result the database can be relied upon to distinguish and classify images.

The databases can store huge amounts of multimedia data as the advancement of technology and the low cost of hard disk presently. The major difficulty is effectively locating images and video frames within databases. The searching process would be time-consuming if it needs to access the whole database which may contain one million of pictures and video. Hence the clustering approach, and algorithms such as Isodata and K-means have been invented to reduce the searching time. Clustering refers to the process of grouping similar feature vectors together. The groups are termed "clusters". Clustering applied to a multimedia database, groups frames and images into similar grey levels, colours and textures for determining various regions in the images. The “centroid” is an important term within a cluster. It is the average value of features vectors in each cluster.

Searching would be based on a clustered or non-clustered histogram. A non-clustered approach would not be concerned with a “centroid” comparison. However the feature vector of the query feature vector is firstly compared with the centroid of each cluster through a clustered search. The query feature vector would only make a comparison with the most similar group as the distances between the query feature vector and the centroid of all clusters are sorted in order. In this paper three clustering algorithms : Forgy, K-means and Isodata were used to test the proposed CBVRS. Section 2.1 describes in more detail about clustering algorithm.

2.1 Clustering Algorithms

Many retrieval systems calculate the features similarity between the query image and all images in the database and rank the images by sorting their similarities. The problem of this full search approach is very time consuming for large scale databases.

The retrieval time of this full search is the sum of the time to calculate similarity time $T_{sim}$ and the time to sort the images in the database according to their similarity $T_{sort}$. Time for the full search $T_{full}$:

$$T_{full} = T_{sim} + T_{sort}$$

$$T_{full} = nT_{sim(i)} + O(n\log n)$$

Where

- $n$: number of images in the database
- $T_{sim(i)}$: total time to calculate the similarity
- $T_{sort}$: total time to rank $n$ images
- $O(n\log n)$: time to sort $n$ elements

If the images in the database are clustered, the retrieval time is the sum of:

1. The time to calculate the similarity between the query and the cluster centers.
2. The time to calculate the similarity between the query and the images in the nearest clusters; and
3. The time to rank the images in step 2.

Time for cluster search $T_{cluster}$:

$$T_{cluster} = kT_{sim(i)} + lT_{sim(i)} + O(l\log l)$$

Where $k$: number of clusters
number of images in the clusters nearest to the query. Since \( k << n \) and \( l << n \), therefore the clustering search time \( T_{\text{cluster}} \) should be much smaller than the full search time \( T_{\text{full}} \) (i.e. \( T_{\text{cluster}} << T_{\text{full}} \)).

In this work, three different clustering algorithms: Forgy, K-Means and Isodata were tested.

### 2.2 DCT Based Video Retrieval System (DBVRS)

In this work, various sets of DCT coefficients extracted from JPEG compressed images were used. The DCT coefficient values can be regarded as the relative amount of the 2D spatial frequencies contained in the 8x8 block input data. The coefficient with zero frequency is called the “DC coefficient” and the remaining 63 coefficients are called “AC coefficients”. Figure 1 shows the zig zag sequence of 8x8 DCT block with DC and 63 AC coefficients. The forward DCT process is able to concentrate most of the signal in the lower spatial frequency. For the 8x8 block from compressed image, we can only consider the low spatial frequency coefficients to construct the energy histogram. S1, S2, S3 and S4 are four sets of the DCT coefficients that were used for testing, as follows:

- S1 = \([\text{DC}]\)
- S2 = \([c(0,1), c(1,0), c(1,1)]\)
- S3 = \([\text{DC}, c(0,1), c(1,0), c(1,1)]\)
- S4 = \([\text{DC}, c(0,1), c(0,2), c(1,0), c(1,1), c(1,2), c(2,0), c(2,1), c(2,2)]\)

The DCT energy histogram is constructed by counting the number of times of an energy level that occurs in 8 x 8 DCT blocks. After constructing the DCT energy histogram, the histogram intersection method is then be used to perform matching. In the matching process, the reference image DCT energy histogram is compared with different groups in the databases. The matched images will have the higher matching values. Histogram intersection is an efficient way to perform matching. The computation complexity is low and can be implemented by most computers. The matching algorithm is as follows:

\[
\text{match\_value} = \sum_{n=1}^{N} \min(I_n, J_n)
\]

Given a pair of histograms which consist of a reference image (I) and a database image (J), each contains N DCT energy levels.

In order to normalize the matching result between 0 and 1, the measure is finally divided by the total number of coefficients used in the reference image. The histogram intersection technique proposed by [2] is defined as follows:

\[
H(I, J) = \frac{\sum_{n=1}^{N} \min(I_n, J_n)}{\sum_{n=1}^{N} I_n}
\]

### 3 Content Based Audio Retrieval System (CBARS)

The proposed Content-Based Audio Retrieval System (CBARS) includes three steps: (1) feature extraction, (2) classification, and (3) retrieval. Feature extraction refers to transforming the original audio data into the feature vector in which the specific desired characteristics are contained. It is the first step in audio processing, and it has a large correlation with the final accuracy. The classification and retrieval function together, by applying the effective formulation of a distance measure and the classification rules.

The audio features can be divided into two categories: time domain and frequency domain. In time domain, the audio representation is expressed as the basic audio amplitude change with time. Some features can be derived from the statistics of amplitude, such as silence rate (SR) and volume. In the frequency domain, features are obtained by applying the Fourier transform. Such features includes pitch, bandwidth, brightness and harmony.

Fig. 2 shows the block diagram of the proposed CBARS. In the proposed CBARS, Mel-Frequency Cepstral Coefficients (MFCCs) [7][8] were used as the feature extraction algorithm for audio signal. As the length of the input query audio data and the stored audio sequence data are usually different, Dynamic Time Warping (DTW) [9] algorithm was
used to generate the same length of feature vectors. The Nearest Neighbor (NN) [10] rule was used for the classification and retrieval of audio data. Sections 3.1, 3.2 and 3.3 explain in more detail the MFCCs, audio classification and DTW respectively.

The transform formula for the Mel-frequency scale and linear frequency scale is as follows.

\[
mel = \ln \left( 1 + \frac{f}{700} \right) \frac{1000}{\ln(1+1000/700)}
\]

### 3.2 Audio Classification

The classification rules [7][8] are needed for the CBARS. The most commonly used algorithms are: Nearest Neighbor (NN) rule, K-NN rule, HMM rule, Neural Network classifier, NFL rule and SVM rule. In this paper the NN rule was used as the classifier for the proposed CBARS.

#### 3.2.1 Nearest Neighbor (NN)

In the proposed CBARS the Nearest Neighbor (NN)[10] rule is used to classify the observation into the category which only depends on a collection of correctly classified samples. It can be defined as following:

\[X_k \in \theta_n, \text{ where } X_k \text{ is the nearest neighbor of } X \text{ and } X_k \text{ belongs to } \theta_n.\]

#### 3.3 Dynamic Time Warping (DTW)

When the feature vectors are generated, there is one problem arises: the lengths of the input and the stored sequences are unlikely to be the same. The Dynamic Time Warping (DTW) [9] algorithm can be used to solve this problem.

DTW algorithm is defined as follows:

1. Let \( g(1,1) = 2d(1,1), j = 1. \)
2. Let \( i_1 = \max(1, j-r), i_2 = \min(j + r, I), \)
   \[ \text{Compute } g(i, j), (i= i_1, \ldots, i_2). \]
3. \( j < J? \)
   \[ \text{YES} \rightarrow (2) \]
   \[ \text{NO} \rightarrow \text{Compute } D(T, R) = g(I, J) / (I+J) \]

The generalized algorithm is as follows:

\[
g(ck) = \min \left\{ \begin{array}{ll}
g(i-1, j) + d(i, j) \\
g(i-1, j-1) + 2d(i, j) \\
g(i, j-1) + d(i, j) \\
\end{array} \right. 
\]

In this paper the DTW algorithm was used to measure the distance between the MFCCs from different sound clips.
4 Testing Results

4.1 DCT based CBIRS
The purpose of this test is used to determine the efficiency of the clustering algorithms in clustering 910 frames of a video in 60 seconds by using 5, 10, 20, 50, 80, 100 clusters. Three clustering algorithms: Forgy, K-Means and Isodata were used. This test was performed on both RGB and DCT coefficients in 24-dimensions.

<table>
<thead>
<tr>
<th>Algorithm Cluster</th>
<th>Forgy (in seconds)</th>
<th>K-means (in seconds)</th>
<th>Isodata (in seconds)</th>
<th>Average Time (in seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2.6</td>
<td>2.2</td>
<td>2.6</td>
<td>2.47</td>
</tr>
<tr>
<td>10</td>
<td>4.4</td>
<td>2.8</td>
<td>8</td>
<td>5.07</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>9.6</td>
<td>19.6</td>
<td>13.07</td>
</tr>
<tr>
<td>50</td>
<td>27.4</td>
<td>23</td>
<td>60.6</td>
<td>37</td>
</tr>
<tr>
<td>80</td>
<td>39.2</td>
<td>47.8</td>
<td>89</td>
<td>58.7</td>
</tr>
<tr>
<td>100</td>
<td>165.2</td>
<td>40.2</td>
<td>143.2</td>
<td>116.2</td>
</tr>
<tr>
<td>Average time (in seconds)</td>
<td>49.76</td>
<td>25.12</td>
<td>64.6</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3 Average time taken for 3 different clustering algorithms using different number of clusters

Fig. 3 shows the average time taken for three different clustering algorithms using different number of clusters. Fig. 3 shows that the K-means has the best performance in video frames retrieval. The Isodata gave the worst performance. This may due to the nature of video which contains 30 frames per second. Five clusters gave the best performance and 100 clusters required the longest clustering time.

4.1.1 Testing on Retrieval Accuracy Based on 5 Clusters
The purpose of this test is to determine the clustering algorithm and number within five clusters used that will yield the highest accuracy results. The results that in five clusters with and without clustering were compared, Fig.56 shows 10 query images used in the test. Each image has a group of top ten images that are similar to the query image. This test was performed in RGB and DCT coefficients with 24-dimensions on a total of 910 video frames. The results are determined by the Percentage of Retrieval Accuracy Method (PRAM) which is used to determine the percentage of accurate results. The function of PRAM is calculated adding the accuracy results of all query video frames performed and then calculated by multiplying by 100.

<table>
<thead>
<tr>
<th>Algorithm Type</th>
<th>RGB</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>Average results in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>9.1</td>
<td>100</td>
</tr>
<tr>
<td>Isodata</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>8.8</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Forgy</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>9.1</td>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>

Fig. 4 Accuracy results produced by RGB and DCT coefficients based on clustering algorithms in five clusters

Fig. 4 shows the average time taken for three different clustering algorithms using different number of clusters. Fig. 3 shows that the K-means has the best performance in video frames retrieval. The Isodata gave the worst performance. This may due to the nature of video which contains 30 frames per second. Five clusters gave the best performance and 100 clusters required the longest clustering time.

4.2 CBARS
This paper proposes and demonstrates a Content Based Audio Retrieval Systems (CBARS). The proposed CBARS is based on the Mel-Frequency Cepstral Coefficients (MFCCs), Dynamic Time Warping (DTW) algorithm and the Nearest Neighbor (NN) classification rule. For distance measurement the Euclidian distance was computed. To test the classification accuracy two experiments were designed and tested using the Leave-one-out rule (each audio clip is taken as query once).
4.2.1 Test 1: classification of pure music

The first test was to classify the pure music into one of three types (piano, guitar, jazz). The MFCCs in different dimensions (order = 8, 12, 24) were used to classify each set of music. In each class of audio, there were 20 clips, thus a total sum of 80 clips of music were sampled in this experiment. The results are shown in Fig. 6.

<table>
<thead>
<tr>
<th>MFCC class \ music</th>
<th>Piano</th>
<th>Guitar</th>
<th>Jazz</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCCs8</td>
<td>100%</td>
<td>100%</td>
<td>75%</td>
</tr>
<tr>
<td>MFCCs12</td>
<td>100%</td>
<td>100%</td>
<td>85%</td>
</tr>
<tr>
<td>MFCCs24</td>
<td>100%</td>
<td>100%</td>
<td>85%</td>
</tr>
</tbody>
</table>

Fig. 6 The accuracy rate of music classification by MFCCs with DTW measurement.

4.2.2 Test 2: classification of music and reggae

In the second test, reggae music data were used. This test was to discriminate the performance of reggae (singing with music) from pure music, and also, to ascertain the extent to which the music classification is influenced by the involvement of reggae. Varied dimensions of MFCCs (8, 12, 24) were used to test the proposed CBARS. Fig. 10 shows the testing results.

<table>
<thead>
<tr>
<th>MFCC sample class \ music</th>
<th>Piano</th>
<th>Guitar</th>
<th>Jazz</th>
<th>Reggae</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCCs8</td>
<td>100%</td>
<td>100%</td>
<td>65%</td>
<td>60%</td>
</tr>
<tr>
<td>MFCCs12</td>
<td>100%</td>
<td>100%</td>
<td>80%</td>
<td>70%</td>
</tr>
<tr>
<td>MFCCs24</td>
<td>100%</td>
<td>100%</td>
<td>70%</td>
<td>70%</td>
</tr>
</tbody>
</table>

Fig. 7 The accuracy rate of music and reggae classification by MFCCs with DTW measurement.

According to Fig. 6 and Fig. 7, piano and guitar can be easily distinguished by the proposed CBARS, while jazz sometimes is misclassified into piano. Actually, it is understandable because upon checking it was found that the misclassified jazz data were close to brisk piano data acoustically. It is because the reggae is brisk and the tempo data is similar to the jazz data. The experimental results also show that the MFCCs equal to 8 had lower performance than the other two MFCCs. From the testing results the MFCCs equal to 12 had the best performance.

5 Conclusion

Multimedia Retrieval is an important research area. In this paper we have proposed and implemented a Content Based Multimedia Retrieval System (CBMRS) for the effective searching of video and audio data in the large multimedia database. The results in Section 4 show that clustering algorithms in CBVRS can speed up the searching time and give a high accuracy rate in the retrieval system. The proposed CBARS based on Mel-Frequency Cepstral Coefficients (MFCCs), the Dynamic Time Warping (DTW) algorithm and the Nearest Neighbor (NN) classification rule can be used to classify pure music and Reggae. The MFCCs order equal to 12 showed the best performance. Our future work will focus on how to use both content based features from CBIRS and CBARS in order to give even more effective search results.

References:


