行政院國家科學委員會專題研究計畫成果報告

視訊資訊系統中視訊內容的相似擷取
Similarity Retrieval of Video Content in Video Information Systems

計畫編號：NSC 88-2213-E-004-006
執行期限：87 年 10 月 1 日至 88 年 7 月 31 日

主持人：沈錳坤　　政治大學資訊科學系

中文摘要

視訊資訊擷取的主要特性在於相似擷取與內容擷取。本計畫研究視訊內容的相似度衡量。我們提出一系列基於視訊畫面序列相似度的衡量方法及其對應的演算法。視訊畫面序列考量畫面的時間先後順序。本計畫並且評估所提出之相似度衡量的效果。

關鍵詞：視訊資訊系統、視訊內容擷取、相似擷取

1. Abstract

The distinguished features of video retrieval lie in the similarity measures and content-based retrieval.

In this project, the similarity measures of video content are investigated. We propose a series of similarity measures based on the similarity of frame sequence which take temporal ordering into consideration. The corresponding algorithms are also presented. The effectiveness of the developed similarity measures measured by precision and recall is also described.

Keywords: Video Information Systems, Content-Based Video Retrieval, Similarity Retrieval

2、Motivations

Video access is one of the important design issues in the development of multimedia information system, video-on-demand and digital library. Video can be accessed by attributes of traditional database techniques, by semantic descriptions of traditional information retrieval technique, by visual features and by browsing.

To support video access by visual features and browsing, structural and content analysis of video must be performed so that video can be indexed and accessed. Having performed the process of video parsing, a sequence of key frames is extracted from each segmented video shot. A sequence of key frames is a representative set of images for each shot.

Each video shot V is associated with a sequence of visual features, $(v_1, v_2,\ldots, v_N)$, where N is the number of key frames, and v_j, $1 \leqq j \leqq N$, is a f-dimensional vector of visual feature value. Given two video shots U and V, assuming that the distance $d(u_i, v_j)$ is available, $\forall\ i$, $1\leqq i\leqq M$, $\forall\ j$, $1\leqq j\leqq N$, the goal is to define the similarity (or dissimilarity) between these video shots in consideration of temporal features.

In this project, we propose a series of shot similarity algorithms based on similarity of frame or key frame sequence. Since the algorithms apply well to both frame sequence and key frame sequence, we will use frame sequence to stand for frames or key frames in proper context in the rest of the report.

3. Similarity of Frame Sequences

Given two sequences of frames, how to define the (dis)similarity between them? People often judge the similarity between videos by common subsequence. We present several similarity algorithms based on the similarity of frame sequence.

A similarity measure is symmetric if $D(U, V) = D(V, U)$. The straightforward measure of similarity is the one-to-one optimal mapping.

In the one to one mapping, we try to map as many pairs of frames as possible under the constraint that each frame u_i in U corresponds to only one frame v_j in V. Obviously, the maximal number of mapping pairs is equal to the number of frames of shorter shot (shot with less number of frames). The mapping with minimum sum of frame distance is selected as the optimal mapping. The formal definitions are given as follows.

Definition 1 Given two video shots $U = (u_1, u_2,\ldots, u_M)$, $V= (v_1, v_2,\ldots, v_N)$ and the distance $d(u_i, v_j)$, $\forall\ i$, $1 \leqq i \leqq M$, $\forall\ j$, $1 \leqq j \leqq N$, a mapping between them is a one-to-one relation R_M from $\{1, 2,\ldots, M\}$ to $\{1,2, \ldots, N\}$, such that

(1) $|R_M| = \min\{M, N\}$, where $|R_M|$ denotes the cardinality of R_M,

(2) for any two ordered pairs (i, j), (k, l) in R_M, $(j < l)$ if and only if $(i < k)$.

Definition 2 Given two video shots $U = (u_1, u_2,\ldots, u_M)$ and $V= (v_1, v_2,\ldots, v_N)$, the distance between U and V for a given mapping R_M, $D'_{R_M}(U, V)$, is defined as

$$D_{R_M}(U,V) = \sum_{\forall (i,j)\in R_M} d(u_i, v_j)\ .$$

Definition 3 Given two video shots $U = (u_1, u_2,\ldots, u_M)$ and $V= (v_1, v_2,\ldots, v_N)$, the distance between U and V for *Optimal Mapping* (*OM*) is defined as

$$D_{OM}(U,V) = \min_{\forall R_M} \{D_{R_M}(U,V)\}\ .$$

The solution of $D_{OM}(U, V)$ can be found based on the approach of dynamic programming. Assume that the shorter shot is U and the longer one is V. Our goal is to find the subsequence of V which is the most similar to U. Let $D[i, j]$ be the minimum cost of mapping between $(u_1, u_2,\ldots, u_i)$ and $(v_1, v_2,\ldots, v_j)$. It is not hard to see that there are two possibilities:

map: the frame v_n is mapped with the frame u_m, $D[m, n] = D[m\text{-}1, n\text{-}1] + d(u_m, v_n)$.

ignore: the frame v_n is not selected to be mapped with the frame u_m, $D[m, n] = D[m, n\text{-}1]$.

Combining these two cases, we get the following recurrence relation for the solution of $D_{OM}(U, V)$.

$$D[m,n] = \min\begin{cases} D[m-1, n-1] + d(u_m, v_n) \\ D[m, n-1] \end{cases},$$

with $D[0, j] = 0$, for all j, $1 \leqq j \leqq N$, and $D[i, 0] = \infty$, for all i, $1 \leqq i \leqq M$. Note that the relation R_M can be constructed by backtracking the matrix $D[M, N]$.

Optimal mapping is the one-to-one frame mapping. However, sometimes the key frames are extracted by uniform sampling. It is likely that, in the extracted sequence of key frames, two consecutive key frames are similar. In addition, sometimes, two sequences of key frames are extracted by non-uniform sampling but with different thresholds. Therefore, given two similar shots, more number of key frames are extracted for the shot with lower threshold. It is necessary to measure the sequence similarity based on many-to-many frame mapping.

Definition 4 Given two video shots $U = (u_1, u_2,\ldots, u_M)$, $V= (v_1, v_2,\ldots, v_N)$ a mapping with replication is a many-to-many relation R_{MR} from $\{1, 2, \ldots, M\}$ to $\{1,2, \ldots, N\}$, such that

(1) for each i, $1\le i \le M$, there exists at least one j, $1\le j \le N$, such that $(i , j) \in R_{MR}$,

(2) for each j, $1\le j \le N$, there exists at least one i, $1\le i \le M$, such that $(i , j) \in R_{MR}$,

(3) for any two ordered pairs (i , j), (k, l) in R_{MR}, $(j \le l)$ if and only if $(i \le k)$.

Definition 5 Given two video shots $U = (u_1, u_2,\ldots, u_M)$ and $V= (v_1, v_2,\ldots, v_N)$, the distance between U and V for a given mapping R_{MR}, $D'_{R_{MR}}(U, V)$, is defined as

$$D_{R_{MR}}(U,V) = \sum_{\forall (i,j)\in R_{MR}} d(u_i, v_j)\ .$$

Definition 6 Given two video shots $U = (u_1, u_2,\ldots, u_M)$ and $V= (v_1, v_2,\ldots, v_N)$, and $d(u_i, v_j)$, $\forall i$, $1\le i \le M$, $\forall j$, $1\le j \le N$, the distance between U and V for *Optimal Mapping with Replication(OMR)* is defined as

$$D_{OMR}(U,V) = \min_{\forall R_{MR}} \{D_{R_{MR}}(U,V)\}$$

Similar to $D_{OM}(U, V)$, the solution of $D_{OMR}(U, V)$ can be found based on the approach of dynamic programming. There are three possible relations between $D[m, n]$

and $D[i, j]$ for some combinations of smaller i's and j's.:

map: the frame v_n is mapped with the frame u_m, $D[m, n] = D[m\text{-}1, n\text{-}1] + d(u_m, v_n)$.

replicate v_n: the frame v_n is replicated to mapped with the frame u_m, $D[m, n] = D[m\text{-}1, n] + d(u_m, v_n)$.

replicate u_m: the frame u_m is replicated to be mapped with the frame v_n, $D[m, n] = D[m, n\text{-}1] + d(u_m, v_n)$.

Combining these three cases, we get the following recurrence relation for the solution of $D_{OMR}(U, V)$.

$$D[m,n] = \min\begin{cases} D[m-1,n-1] + d(u_m, v_n) \\ D[m-1,n] + d(u_m, v_n) \\ D[m,n-1] + d(u_m, v_n) \end{cases},$$

with $D[0, 0] = 0$, $D[0, j] = \infty$, for all j, $1 \leq j \leq N$, and $D[i, 0] = \infty$, for all i, $1 \leq i \leq M$.

A similarity measure is asymmetric if $D(U, V) \neq D(V, U)$. In general, asymmetric similarity measure is used to map between the query sequence of frames and video sequence of frames. The simplest proposed asymmetric similarity measure is the Optimal Subsequence Mapping (OSM). The algorithm of OSM is similar to that of OM except that the query video sequence must be the shorter sequence.

Similar to the similarity measure OMR in symmetric measures, we defined the OSMR in asymmetric measures as follows.

Definition 7 Given the query shot $Q = (q_1, q_2,\ldots, q_M)$, the video shot $V = (v_1, v_2,\ldots, v_N)$, $M \leqq N$, a subsequence mapping with replication is a one-to-many relation R_{SMR} from $\{1, 2, \ldots, M\}$ to $\{1,2, \ldots, N\}$, such that

(1) for each i, $1 \leqq i \leqq M$, there exists at least one j, $1 \leqq j \leqq N$, such that $(i, j) \in R_{SMR}$,

(2) for each j, $1 \leqq j \leqq N$, there exists one i, $1 \leqq i \leqq M$, such that $(i , j) \in R_{SMR}$,

(3) for any two ordered pairs (i, j), (k, l) in R_{SMR}, $(j < l)$ if and only if $(i \leq k)$.

Definition 8 Given the query shot $Q = (q_1, q_2,\ldots, q_M)$, the video shot $V = (v_1, v_2,\ldots, v_N)$, and the distance $d(q_i, v_j)$, $\forall$ i, $1 \leqq i \leqq M$, $\forall$ j, $1 \leqq j \leqq N$, the distance between Q and V for a given mapping R_{SMR}, $D'_{R_{SMR}}(Q, V)$, is defined as $D_{R_{SMR}}(Q,V) = \sum_{\forall (i,j)\in R_{SMR}} d(q_i, v_j)$.

Definition 9 Given the query shot $Q = (q_1, q_2,\ldots, q_M)$ and the video shot $V = (v_1, v_2,\ldots, v_N)$, the distance between U and V for *Optimal Subsequence Mapping with Replication (OSMR)* is defined as

$$D_{OSMR}(Q,V) = \min_{\forall R_{SMR}} \{D_{R_{SMR}}(Q,V)\}$$

Sometimes, it is required that the mapped frames are consecutive. Therefore, the constraint of mapping relation is more strict.

Definition 10 Given the query shot $Q = (q_1, q_2,\ldots, q_M)$, the video shot $V = (v_1, v_2,\ldots, v_N)$, $M \leqq N$, a consecutive mapping is a one-to-one relation R_{CM} from $\{1, 2,\ldots, M\}$ to $\{1, 2,\ldots, N\}$, such that

(1) for each i, $1 \leqq i \leqq M$, there exists one j, $1 \leqq j \leqq N$, such that $(i , j) \in R_{CM}$,

(2) for any two ordered pairs (i , j), (k, l) in R_{CM}, $[(j - l) = 1]$ if and only if $[(i - k) = 1]$.

Definition 11 Given the query shot $Q = (q_1, q_2,\ldots, q_M)$, the video shot $V = (v_1, v_2,\ldots, v_N)$, and $d(q_i, v_j)$, $\forall$ i, $1 \leqq i \leqq M$, $\forall$ j, $1 \leqq j \leqq N$, the distance between Q and V for a given consecutive mapping R_{CM}, $D'_{R_{CM}}(Q, V)$, is defined as $D_{R_{CM}}(Q,V) = \sum_{\forall (i,j)\in R_{CM}} d(q_i, v_j)$.

Definition 12 Given the query shot $Q = (q_1, q_2,\ldots, q_M)$ and the video shot $V = (v_1, v_2,\ldots, v_N)$, the distance between Q and V for *Optimal Consecutive Mapping* is defined as

$$D_{OCM}(Q,V) = \min_{\forall R_{CM}} \{D_{R_{CM}}(Q,V)\}.$$

Algorithm Optimal Consecutive Mapping (OCM)

```
for $j = 0$ to $M\text{-}N$ do $D[0, j] = 0$;
for $j = 0$ to $M\text{-}N$ do
    for $i = 1$ to $N\text{-}1$ do
        $D[i, i+j] = D[i\text{-}1, i+j\text{-}1]+d(q_i, v_{i+j})$;
$D[M, M] = D[M\text{-}1, M\text{-}1]+d(q_M, v_M)$;
for $j = M+1$ to $N$ do
    $D[M, j] = \min(D[M\text{-}1, j\text{-}1]+d(q_M, v_j), D[M, j\text{-}1])$;
return $D[M, N]$
```

Next, we extend the definition of Optimal Consecutive Mapping to Optimal Consecutive Mapping with Replication (OCMR). In OCMR, each frame of query sequence is permitted to map with more than one frame of video sequence. In addition, the same as OCM, the mapped frames of video sequence must be consecutive.

Definition 13 Given the query video shot Q= (q_1, q_2,…, q_M), the video shot V= (v_1, v_2,…, v_N), M ≦N, a consecutive matching with replication is a one-to-many relation R_{CMR} from {1, 2, …, M} to {1,2, …, N}, such that

(1) for each i, 1 ≦i ≦M, there exists at least one j, 1 ≦j ≦N, such that (i , j) ∈ $R_{CMR,}$

(2) let $p_{max}(i)$ be $\max\{j|(i, j) \in R_{CMR}\}$, $p_{min}(i)$ be $\min\{j|(i, j) \in R_{CMR}\}$, for each j, p_{min}(1) ≦j ≦$p_{max}(M)$, there exists one i, 1 ≦j ≦M, such that (i , j) ∈ $R_{CMR,}$

(3) for any two ordered pairs (i , j), (k, l) in R_{CMR}, if (i < k) then (j < l),

(4) any two ordered pairs (i , j), (k, l) in R_{CMR}, if [(j - l) = 1] then either [(i - k) = 1] or [(i - k) = 0].

Definition 14 Given the query video shot Q= (q_1, q_2,…, q_M), the video shot V= (v_1, v_2,…, v_N) and $d(q_i, v_j)$, ∀ i, 1 ≦i ≦M, ∀ j, 1 ≦j ≦N, the distance between Q and V for a given consecutive mapping R_{CMR}, $D'_{R_{CMR}}(Q, V)$, is defined as

$$D_{R_{CMR}}(Q,V) = \sum_{\forall(i,j)\in R_{CMR}} d(q_i, v_j)\ .$$

Definition 15 Given the query shot Q= (q_1, q_2,…, q_M) and the video shot V= (v_1, v_2,…, v_N), the distance between Q and V for *Optimal Consecutive Mapping with Replication* is defined as

$$D_{OCMR}(Q,V) = \min_{\forall R_{CMR}} \{D_{R_{CMR}}(Q,V)\}\ .$$

The solution of OCMR can be obtained as follows. First, if a mapping R is an OCMR, then the first frame q_1 must be mapped with only one frame of video sequence, so does the last query frame q_M. Otherwise, suppose that q_1 is mapped with frames v_a,… $v_{i\text{-}1}$, v_i, q_M is mapped with v_j, $v_{j\text{-}1}$ …, v_b. We can derive a mapping with less distance by removing the mapping pairs between q_1 and v_a,… $v_{i\text{-}1}$, and those between q_M and $v_{j\text{-}1}$ …, v_b. Therefore the behaviors of q_1 and q_M are the same as those of frames in OCM and the behaviors of q_2 , q_3 , …, $q_{M\text{-}1}$, q_M are the same as those of frames in OSMR.

4. Experiments

To evaluate the performance of the proposed similarity measures, we have a database of 100 video shots. These video shots were MPEG-II video which digitized from CTS news. The length of these video ranges from 32 to 205 seconds. Five sample query video are selected from the database. For each video, a sequence of key frames were extracted. Each video clip is decompressed first and the key frames are extracted by the process of non-uniform key frames extraction. Each key frame is represented as a 64-bin color histogram in HSV color space. We measure the performance by precision and recall. The ground truth to determine relevant video is judged by humans. Using the proposed similarity measures OM and OMR, the sample query returns a list of candidate video and the precision-recall values are calculated.

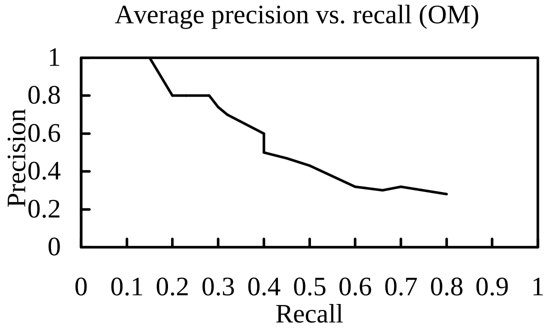

Figure 1. The average precision-recall curve for similarity measure OM.

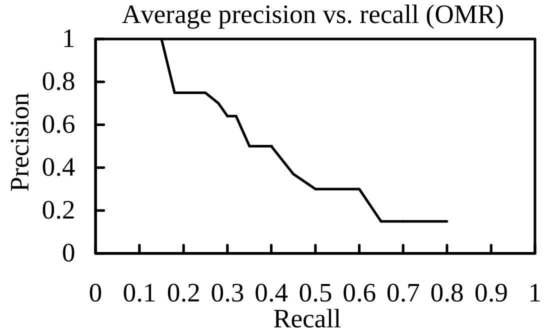

Figure 2. The average precision-recall curve for similarity measure OMR.

Figure 1 and 2 show the average precision-recall curve for similarity measures OM and OMR respectively. The result shows that both measures performs well, especially when only one video is returned for query video. In this case, the precision is one. That is, both measures return the most similar video in the first rank. Moreover, it can be seen that OM performs better than OMR. This phenomenon can be realized as follows. OM is the one-to-one mapping and leaves out unmatched frames while OMR is the many-to-many mapping in which all frames are considered. However, from the human perception human point of view, two video shots are similar only if there are similar frames between them. Therefore, OM behaves like human perception.

5. Conclusions

In this project, we have proposed a series of video similarity measures based on similarity of frame sequence. The similarity algorithms based on the approach of dynamic programming are also presented. The experiment results show that similarity measure OM perform well than OMR.

In fact, the performance highly depends on the extraction process of video content. We plan to measure the performance by considering the effect of content extraction process. In addition, in the proposed similarity measures, two dissimilar frames are permitted to be mapped. The similarity measures in which the dissimilarity constraint is imposed need to be investigated.

6. Contribution

[1] M. K. Shan and S. Y. Lee, Content-Based Similarity Measures for Video Based on Similarity of Frame Sequence, IEEE IW-MMDBMS'98 International Workshop on Multimedia Data Base Management Systems, Dayton, Ohio, 1998.

[2] M. K. Shan and S. Y. Lee, A Generic Framework for Similarity Measures of Content-Based Video Retrieval, submitted to Pattern Recognition Letters, 1999.

7. References

[1] N. Dimitrova and F. Golshani, Rx for Semantic Video Database Retrieval, In *Proceedings of ACM Multimedia'94*, San Francisco, CA, pp. 219-226, 1994.

[2] M. Flickner et al., Query by Image and Video Content: The QBIC System, *IEEE Computer*, Vol. 28, No. 9, pp. 23-32, 1995.

[3] A. Gupta and R. Jain, Visual Information Retrieval, *Communications of ACM*, Vol. 40, No. 5, pp. 71-79, 1997.

[4] L. A. Rowe , J. S. Boreczky and C. A. Eads, Indexes for User Access to Large Video Databases, In *Proceedings. of Storage and Retrieval for Image and Video Databases II, IS&T/SPIE Symposium on Electronic Imaging Science & Technology*, San Jose, CA, pp. 150-161, 1994.

[5] J. K. Wu, A. D. Narasimhalu, B. M. Mehtre, C. P. Lam and Y. J. Gao, CORE: A Content-Based Retrieval Engine for Multimedia Information Systems, *Multimedia Systems*, Vol. 3, No. 1, pp. 25-41, 1995.

[6] M. M. Yeung and B. Liu, Efficient Matching and Clustering of Video Shots, In *Proceedings of International Conference on Image Processing'95*, Washington, DC, pp. 338-341, 1995.

[7] H. J. Zhang, A. Kankanhalli and W. Smoliar, Automatic Partitioning of Full-Motion Video, *Multimedia Systems*, Vol. 1, No. 1, pp. 10-28, 1993.

[8] H. J. Zhang, D. Zhong and S. W. Smoliar, An Integrated System for Content-Based Video Retrieval and Browsing, *Pattern Recognition*, Vol. 30, No, 4, pp. 643-658, 1997.