ELSEVIER

Pattern Recognition Letters 22 (2001) 517–532

Pattern Recognition Letters

www.elsevier.nl/locate/patrec

A framework for temporal similarity measures of content-based scene retrieval

Man-Kwan Shan [a,*], Suh-Yin Lee [b]

[a] Department of Computer Science, National Cheng Chi University, Taipei, Taiwan, ROC

[b] Institute of Computer Science and Information Engineering, National Chiao Tung University, HsinChu, Taiwan, ROC

Abstract

The distinguished features of video retrieval lie in the similarity measures and content-based retrieval. Most research on content-based video retrieval represents the content of video as a set of frames, leaving out the temporal ordering of frames in the shot. In this paper, 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. All the algorithms corresponding to the similarity measures are based on the approach of dynamic programming. © 2001 Published by Elsevier Science B.V.

Keywords: Content-based video retrieval; Similarity measure; Sequence mapping; Dynamic programming

1. Introduction

Video access is one of the important design issues in the development of multimedia information systems, video-on-demand and digital libraries. Video can be accessed by attributes of traditional database techniques (Little et al., 1993; Rowe et al., 1994), by semantic descriptions of traditional information retrieval technique (Wu et al., 1995), by visual features and by browsing (Furht et al., 1995).

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 (Zhang et al., 1993). 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.

For each video shot, the video index is constructed based on visual features of the corresponding sequence of key frames. That is, 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, $1 \leqslant j \leqslant N$, is an f-dimensional vector of visual feature value.

Most researchers defined shot similarity as the similarity of the images chosen to represent each shot (Flickner et al., 1995; Zhang et al., 1995). Some approaches defined shot similarity based on the similarities between two key frame sets (Yeung and Liu, 1995; Zhang et al., 1995). The similarity between two shots U and V is defined as $\mathrm{SIM}(U, V) = \mathrm{Max}(\{\mathrm{sim}(u_i, v_j) \mid \forall i,\ 1 \leqslant i \leqslant M,\ \forall j,\ 1 \leqslant j \leqslant N\})$.

[*] Corresponding author.
E-mail addresses: mkshan@cherry.cs.nccu.edu.tw (M.-K. Shan), sylee@csie.nctu.edu.tw (S.-Y. Lee).

0167-8655/01/$ - see front matter © 2001 Published by Elsevier Science B.V.
PII: S0167-8655(00)00120-3

This definition assumes that the similarity between two clips can be determined by the most similar pair of key frames.

Zhang et al. (1997) also defined another similarity according to the sum of the most similar pair of key frames as $\mathrm{SIM}(U, V) = (1/M)\sum_{i=1}^{M} \mathrm{Max}(\{\mathrm{sim}(u_i, v_j) \mid \forall j,\ 1 \leqslant j \leqslant N\})$.

Note that this definition is asymmetric. The complexity of calculating the above similarity measures is $\mathrm{O}(M \times N)$. The above similarity measures leave out all the temporal information of the shots. A more precise measure of shot similarity should incorporate temporal information.

In the approach developed by Zhong et al. (1995), temporal variation, camera operation and statistic motion features were used to represent the temporal information of shots for similarity measure. However, this approach did not take temporal ordering into consideration.

Dimitrova and Abdel-Mottaled (1997) presented an approach for video retrieval from a large archive of MPEG or Motion JPEG compressed video clips. Video clips are characterized by a sequence of representative frame signatures, which are constructed from DC coefficients and motion information. The similarity between two video clips is determined by the average distance of corresponding frames between two videos as the similarity measure.

In the approach proposed by Mohan (1998), videos are matched based on similarity of temporal activity. Video sequences are represented as a sequence of feature vectors computed from compressed MPEG videos. Query video sequence is matched against the video sequences in a database using sequential matching.

We proposed similarity measures based on similarity of frame sequence (Shan and Lee, 1998). A set of similarity measures was developed. The similarity was measured on the basis of the maximum-likelihood criterion.

Adjeroh et al. (1999) presented a distance measure for video sequences. In this approach, a video sequence is represented as a string, called the *vstring*. The problem of video sequence matching is therefore formulated as the pattern matching problem. The distance measure is based on the concept of string-editing distance. Five editing operations, insertion, deletion, substitution, swap, and fusion, are considered.

Standard video hierarchical model classifies video sequences into four levels according to level of temporal resolution. The lowest level is a series of frames. At the next higher level, frames are grouped into shots. A shot is a continuous camera recording. Consecutive shots are aggregated into scenes based on story-telling coherence. For scene-based video retrieval, it is beneficial for users to access by incorporating temporal ordering information.

In this paper, we propose a framework for similarity measures between video clips. We will use clip to stand for any temporal level of video in the rest of this paper. In contrast to the work of Adjeroh et al., where a unique distance is defined, a series of similarity measures based on similarity of frame sequence is proposed. The proposed approach gives users the freedom to specify the temporal constraints. Moreover, our proposed measure is measured on the basis of maximum likelihood of frames, instead of minimum editing distance. In other words, we consider the mapping of similar frames and measure the similarity as the sum of the distance between mapped similar frames.

For each similarity measure, the corresponding algorithm is also described. All these algorithms are based on the dynamic programming approach (Aoe, 1994; Saoke and Chiba, 1990) and take $\mathrm{O}(M \times N)$ of computation time.

In the next section, we present the proposed similarity measures and algorithms. The result of performance analysis is described in Section 3. Section 4 concludes the paper.

2. Similarity of frame sequence

In our framework, we use similarity measures that are consistent to the perception of human beings. People often judge the similarity between videos by common subsequence. In this section, we present several similarity algorithms based on the similarity of frame sequence. Also note that, in this section, examples are given to illustrate the computation of the algorithms. For the sake of clarity, each frame in the examples is represented as an integer, instead of an f-dimensional feature vector. This does not affect the correctness of the algorithms, since the assumption of these algorithms lies in the availability of distance function between frames.

2.1. Symmetric similarity measures

A similarity measure is symmetric if $D(U,V) = D(V,U)$. The symmetric measure is used in the video clustering or query-by-video example. The straightforward measure of similarity is the one-to-one optimal 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 the shorter clip (clip 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 clips $U = (u_1, u_2, \ldots, u_M)$, $V = (v_1, v_2, \ldots, v_N)$ and the distance $d(u_i, v_j)$ $\forall i,\ 1 \leqslant i \leqslant M,\ \forall j,\ 1 \leqslant j \leqslant N$, 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) $|R_\mathrm{M}| = \min\{M, N\}$, where $|R_\mathrm{M}|$ 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 clips $U = (u_1, u_2, \ldots, u_M)$ and $V = (v_1, v_2, \ldots, v_N)$, the distance between U and V for a given mapping R_M, $D_{R_\mathrm{M}}(U,V)$, is defined as

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

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

$$D_\mathrm{OM}(U,V) = \min_{\forall R_\mathrm{M}} \{D_{R_\mathrm{M}}(U,V)\}.$$

The solution of $D_\mathrm{OM}(U,V)$ can be found based on the approach of dynamic programming. Assume that the shorter clip is U and the longer one is V. First, from Definition 1, it is obvious that each frame of U must be matched with a frame of 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 $(v_1, v_2, \ldots, v_j)$. Dynamic programming tries to find a relation between $D[m,n]$ and $D[i,j]$ for some combinations of smaller is and js. It is not hard to see that there are two possibilities:

map: the frame v_n is mapped with the frame u_m, $D[m,n] = D[m-1, n-1] + d(u_m, v_n)$;

ignore: the frame v_n is not selected to map with the frame u_m, $D[m,n] = D[m, n-1]$.

Note that it is not permitted to ignore frames of U. The reason, which has been discussed earlier, comes from Definition 1. Combining these two cases, we get the following recurrence relation for the solution of $D_\mathrm{OM}(U,V)$:

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

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

The setting of initial values can be understood as follows. $D[0, j]$ denotes the distance between ( ) and $(v_1, v_2, \ldots, v_j)$, where ( ) is a null sequence. Therefore, $D[0, j] = 0$, for all j, $1 \leqslant j \leqslant N$. $D[i, 0]$ denotes the distance between $(u_1, u_2, \ldots, u_i)$ and ( ). As discussed earlier, each frame of U must be mapped with a frame of V. It is impossible to map frames of U with those of null sequence ( ). Therefore, $D[i, 0] = \infty$, for all i, $1 \leqslant i \leqslant M$.

The algorithm for the solution of optimal mapping is listed as follows. Moreover, note that the relation R_{M} can be constructed by backtracking the matrix $D[M, N]$.

Algorithm (*Optimal mapping* (*OM*)).

```
if $M \geqslant N$ then{
    for $i = 0$ to $M$ do $D[i, 0] = 0$;
    for $j = 1$ to $N$ do $D[0, j] = \infty$;
    for $i = 1$ to $M$ do
        for $j = 1$ to $N$ do
            $D[i, j] = \min(D[i-1, j-1] + d(u_i, v_j), D[i-1, j])$;}
else{
    for $i = 1$ to $M$ do $D[i, 0] = \infty$;
    for $j = 0$ to $N$ do $D[0, j] = 0$;
    for $i = 1$ to $M$ do
        for $j = 1$ to $N$ do
            $D[i, j] = \min(D[i-1, j-1] + d(u_i, v_j), D[i, j-1])$;}
return $D[M, N]$
```

Example 1. Given two video clips $V = (2, 6, 7, 1, 6, 10, 4)$ and $U = (5, 3, 2, 8, 3)$, the computation process of the distance of optimal mapping is described in Table 1. Fig. 1 shows the mapping of frames between V and U.

Optimal mapping is 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 clips, more number of key frames are extracted for the clip with lower threshold. It is necessary to measure the sequence similarity based on many-to-many frame mapping.

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

Table 1
Computation of optimal mapping of Example 1

		2	6	7	1	6	10	4
	0	**0**	0	0	0	0	0	0
5	∞	3	**1**	1	1	1	1	1
3	∞	∞	6	**5**	3	3	3	2
2	∞	∞	∞	11	**6**	**6**	6	5
8	∞	∞	∞	∞	18	8	**8**	8
3	∞	∞	∞	∞	∞	21	16	**9**

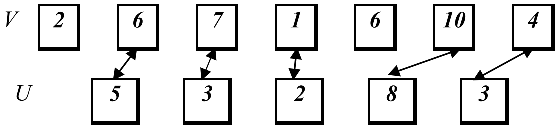

Fig. 1. Optimal mapping of Example 1.

(1) for each $i,\ 1 \leqslant i \leqslant M$, there exists at least one $j,\ 1 \leqslant j \leqslant N$, such that $(i,j) \in R_{\mathrm{MR}}$,
(2) for each $j,\ 1 \leqslant j \leqslant N$, there exists at least one $i,\ 1 \leqslant i \leqslant M$, such that $(i,j) \in R_{\mathrm{MR}}$,
(3) for any two ordered pairs $(i,j),(k,l)$ in $R_{\mathrm{MR}},\ (j \leqslant l)$ if and only if $(i \leqslant k)$.

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

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

Definition 6. Given two video clips $U = (u_1, u_2, \ldots, u_M)$ and $V = (v_1, v_2, \ldots, v_N)$, and $d(u_i, v_j)\ \forall i,\ 1 \leqslant i \leqslant M,\ \forall j,\ 1 \leqslant j \leqslant N$, the distance between U and V for *optimal mapping with replication* (OMR) is defined as

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

Similar to $D_{\mathrm{OM}}(U,V)$, the solution of $D_{\mathrm{OMR}}(U,V)$ can be found based on the approach of dynamic programming. Let $D[i,j]$ be the minimum cost of mapping with replication between $(u_1, u_2, \ldots, u_i)$ and $(v_1, v_2, \ldots, v_j)$. There are three possible relations between $D[m,n]$ and $D[i,j]$ for some combinations of smaller is and js:

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

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

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

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

The initial values are set according to Definition 4. Definition 4 states that each frame of one clip must be mapped with at least one frame of the other clip. It is impossible to map a sequence of frames with a null sequence ( ). Therefore, $D[i,0] = \infty$, for all $i,\ 1 \leqslant i \leqslant M$, and $D[0,j] = \infty$, for all $j,\ 1 \leqslant j \leqslant N$. The algorithm for the solution of optimal mapping with replication is listed as follows.

Algorithm (*Optimal mapping with replication* (*OMR*)).

```
$D[0,0] = 0$;
for $i = 1$ to $M$ do $D[i,0] = \infty$;
for $j = 1$ to $N$ do $D[0,j] = \infty$;
for $i = 1$ to $M$ do
  for $j = 1$ to $N$ do
    $D[i,j] = \min(D[i-1,j-1] + d(u_i,v_j), D[i-1,j] + d(u_i,v_j), D[i,j-1] + d(u_i,v_j))$;
return $D[M,N]$
```

Example 2. For the same example as Example 1, the computation process of the distance of optimal mapping with replication is described in Table 2. Fig. 2 shows the mapping of frames between U and V. Note that the shaded boxes denote the intermediate path to derive the minimum cost.

Definition 7. Let $G = (U, E, V)$ be a bipartite graph, such that U, V are two disjoint sets of vertices and E is a set of edges connecting vertices from U to V. A *redundant mapping* is a set of edges with path length at least 3.

In terms of frame mapping, U and V are respective frame sequences of clips U and V. E is a set of possible mapping between frames of U and V. Fig. 3 illustrates a redundant mapping. The term 'redundant' means that the mapping between frame u_{i-1} and frame v_j (the edge $\overline{u_{i-1}v_j}$) is redundant and can be removed. This is because there has been one mapping between frames u_{i-1} and v_{j-1}, and another mapping between frames u_i, v_j. It is not necessary to map frames u_{i-1}, v_j.

Lemma 1. *Algorithm OMR never produces redundant mapping.*

Proof. As in Fig. 3, the three matching $\overline{u_{i-1}v_{j-1}}$, $\overline{u_{i-1}v_j}$, $\overline{u_iv_{j-1}}$ constitute a redundant mapping. Therefore, in the computation table,

Table 2
Computation of optimal mapping with replication of Example 2

		2	6	7	1	6	10	4
	0	∞	∞	∞	∞	∞	∞	∞
5	∞	**3**	**4**	**6**	10	11	16	17
3	∞	4	6	8	**8**	11	18	17
2	∞	4	8	11	**9**	12	19	19
8	∞	10	6	7	14	**11**	**13**	17
3	∞	11	9	10	9	12	18	**14**

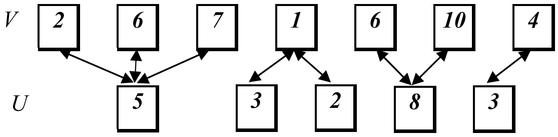

Fig. 2. Optimal mapping with replication of Example 2.

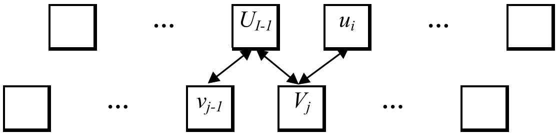

Fig. 3. Redundant mapping.

$$\begin{aligned}
&D[i,j-1] = D[i-1,j-1] + d(u_i, v_{j-1}),\\
&D[i,j] = D[i,j-1] + d(u_i, v_j),\\
&D[i,j] < D[i-1,j-1] + d(u_i, v_j)\\
&\quad \Rightarrow D[i,j-1] + d(u_i, v_j) < D[i-1,j-1] + d(u_i, v_j)\\
&\quad \Rightarrow D[i,j-1] < D[i-1,j-1], \text{ which is a contradiction.} \quad \square
\end{aligned}$$

The effect of video segmentation also affects the performance of sequence mapping. Video segmentation with lower threshold produces clips with much variation. It is possible that clip U is very similar to V except that some frames are very dissimilar. Using OM, these dissimilar frame pairs produce large distance. Therefore, in the next definition, the mapping is constrained by a threshold δ. Two frames with distance larger than δ are not allowed to map. In addition, each unmapped frame is associated with a penalty value ϖ in the computation of clip distance. Otherwise, the result of the distance-constrained optimal mapping with minimum cost would be no mapping at all. No mapping produces the distance of zero.

Definition 8. Given two video clips $U = (u_1, u_2, \ldots, u_M)$ and $V = (v_1, v_2, \ldots, v_N)$, the frame distance tolerance δ, and the distance $d(u_i, v_j)\ \forall i,\ 1 \leqslant i \leqslant M, \forall j,\ 1 \leqslant j \leqslant N$, a distance-constrained mapping between them is a one-to-one relation R_{DM} from $\{1, 2, \ldots, M\}$ to $\{1, 2, \ldots, N\}$, such that

(1) for each ordered pair (i,j) in R_{DM}, $d(u_i, v_j) \leqslant \delta$;

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

Definition 9. Given two video clips $U = (u_1, u_2, \ldots, u_M)$ and $V = (v_1, v_2, \ldots v_N)$, the frame distance constraint δ, and the penalty ϖ for unmapped frames, $\varpi = \delta/2$, the distance between U and V for a given distance-constrained mapping R_{DM}, $D'_{R_{\mathrm{DM}}}(U, V, \delta, \varpi)$, is defined as

$$D'_{R_{\mathrm{DM}}}(U, V, \delta, \varpi) = \varpi \times (M + N - 2 \times |R_{\mathrm{DM}}|) + \sum_{\forall (i,j) \in R_{\mathrm{DM}}} d(u_i, v_j),$$

where $|R_{\mathrm{DM}}|$ denotes the cardinality of the relation R_{DM}.

In the last definition, if the penalty value is less than $\delta/2$, it would be better not to map two frames with distance less than δ. This is why Definition 9 sets the value of penalty ϖ.

Definition 10. Given two video clips $U = (u_1, u_2, \ldots, u_M)$ and $V = (v_1, v_2, \ldots, v_N)$, the distance between U and V for *distance-constrained optimal mapping* is defined as

$$D_{\mathrm{DOM}}(U, V, \delta, \varpi) = \min_{\forall R_{\mathrm{DM}}} \{D_{R_{\mathrm{DM}}}(U, V, \delta, \varpi)\}.$$

The algorithm DOM is adapted from the optimal correspondence of string subsequence (Wang and Pavlidis, 1990). Let $D[i,j]$ be the minimum cost of distance-constrained optimal mapping between $(u_1, u_2, \ldots, u_i)$, and $(v_1, v_2, \ldots, v_j)$. There are three possible relations between $D[m,n]$ and $D[i,j]$ for some combinations of smaller is and js:

map: the key frame v_n is mapped with the key frame u_m, $D[m,n] = D[m-1,n-1] + d(u_m, v_n)$;

ignore v_n: the key frame v_n is ignored and the penalty value is added, $D[m,n] = D[m,n-1] + \delta/2$;

ignore u_m: the key frame u_m is ignored and the penalty value is added, $D[m,n] = D[m-1,n] + \delta/2$.

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

$$D[m,n] = \min \begin{cases} D[m-1,n-1] + d(u_m, v_n) \\ D[m-1,n] + \delta/2 \\ D[m,n-1] + \delta/2 \end{cases},$$

with $D[0,0] = 0$, $D[0,j] = j * \delta/2$, for all j, $1 \leqslant j \leqslant N$, and $D[i,0] = i * \delta/2$, for all i, $1 \leqslant i \leqslant M$.

Algorithm (*Distance-constrained optimal mapping* (*DOM*)).

```
$D[0,0] = 0$;
for $i = 1$ to $M$ do $D[i,0] = i * \delta/2$;
for $j = 1$ to $N$ do $D[0,j] = j * \delta/2$;
for $i = 1$ to $M$ do
    for $j = 1$ to $N$ do
        $D[i,j] = \min(D[i-1,j-1] + d(u_i,v_j), D[i-1,j] + \delta/2, D[i,j-1] + \delta/2)$;
return $D[M,N]$
```

Example 3. For the same example as Example 1, for the frame distance constraint δ equals 3, the process for computation of the distance of distance-constrained optimal mapping is described in the Table 3. One such mapping is shown in Fig. 4. The value of penalty ϖ for unmapped frames is 3/2; then the distance equals $11 - 4 * 1.5 + 4 * 1.5 = 11$. The number of unmapped frames can be derived from the backtracking of Table 3.

2.2. Asymmetric similarity measures

A similarity measure is asymmetric if $D(U,V) \neq D(V,U)$. In general, asymmetric similarity measures are used when users query video by some key 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. Therefore, it is not necessary to compare the length between the video clip U and query video clip Q.

Table 3
Computation of distance-constrained optimal mapping of Example 3

		2	*6*	*7*	*1*	*6*	*10*	*4*
	0	**1.5**	3	4.5	6	7.5	9	10.5
5	**1.5**	3	**2.5**	**4**	5.5	7	8.5	10
3	**3**	**2.5**	4	**5.5**	**6**	7.5	9	9.5
2	4.5	**3**	**4.5**	6	**6.5**	**8**	9.5	11
8	6	4.5	5	**5.5**	**7**	**8.5**	**10**	11.5
3	7.5	7	6.5	7	7.5	9	10.5	**11**

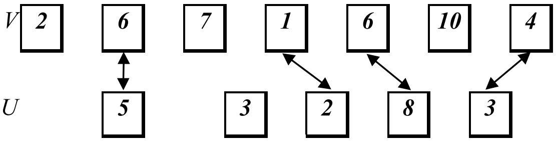

Fig. 4. One of the distance-constrained optimal mappings of Example 3.

Definition 11. Given the query clip $Q = (q_1, q_2, \ldots, q_M)$ and the video clip $V = (v_1, v_2, \ldots, v_N)$, $M \leqslant N$, a subsequence mapping is a one-to-one relation R_{SM} from $\{1, 2, \ldots, M\}$ to $\{1, 2, \ldots, N\}$, such that

(1) for each i, $1 \leqslant i \leqslant M$, there exists one j, $1 \leqslant j \leqslant N$, such that $(i, j) \in R_{\mathrm{SM}}$;
(2) for any two ordered pairs $(i, j), (k, l)$ in R_{SM}, $(j < l)$ if and only if $(i < k)$.

Definition 12. Given the query clip $Q = (q_1, q_2, \ldots, q_M)$, the video clip $V = (v_1, v_2, \ldots, v_N)$, and the distance $d(q_i, v_j)$, $\forall i$, $1 \leqslant i \leqslant M$, $\forall j$, $1 \leqslant j \leqslant N$, the distance between Q and V for a given subsequence mapping R_{SM}, $D_{R_{\mathrm{SM}}}(Q, V)$, is defined as

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

Definition 13. Given the query clip $Q = (q_1, q_2, \ldots, q_M)$ and the video clip $V = (v_1, v_2, \ldots, v_N)$, the distance between Q and V for *optimal subsequence mapping* is defined as

$$D_{\mathrm{OSM}}(Q, V) = \min_{\forall R_{\mathrm{SM}}} \{D_{R_{\mathrm{SM}}}(Q, V)\}.$$

Algorithm (*Optimal subsequence mapping* (*OSM*)).

```
for $i = 1$ to $M$ do $D[i, 0] = \infty$;
for $j = 0$ to $N$ do $D[0, j] = 0$;
for $i = 1$ to $M$ do
  for $j = 1$ to $N$ do
    $D[i, j] = \min(D[i-1, j-1] + d(q_i, v_j), D[i, j-1])$;
return $D[M, N]$
```

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

Definition 14. Given the query clip $Q = (q_1, q_2, \ldots, q_M)$, the video clip $V = (v_1, v_2, \ldots, v_N)$, $M \leqslant N$, 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, $1 \leqslant i \leqslant M$, there exists at least one j, $1 \leqslant j \leqslant N$, such that $(i, j) \in R_{\mathrm{SMR}}$;
(2) for each j, $1 \leqslant j \leqslant N$, there exists one i, $1 \leqslant i \leqslant M$, such that $(i, j) \in R_{\mathrm{SMR}}$;
(3) for any two ordered pairs $(i, j), (k, l)$ in R_{SMR}, $(j < l)$ if and only if $(i \leqslant k)$.

Definition 15. Given the query clip $Q = (q_1, q_2, \ldots, q_M)$, the video clip $V = (v_1, v_2, \ldots, v_N)$, and the distance $d(q_i, v_j)$ $\forall i$, $1 \leqslant i \leqslant M$, $\forall j$, $1 \leqslant j \leqslant N$, the distance between Q and V for a given mapping R_{SMR}, $D_{R_{\mathrm{SMR}}}(Q, V)$, is defined as

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

Definition 16. Given the query clip $Q = (q_1, q_2, \ldots, q_M)$ and the video clip $V = (v_1, v_2, \ldots, v_N)$, the distance between U and V for *optimal subsequence mapping with replication* (OSMR) is defined as

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

Algorithm (*Optimal subsequence mapping with replication* (*OSMR*)).

```
$D[0,0] = 0$;
for $i = 1$ to $M$ do $D[i,0] = \infty$;
for $j = 1$ to $N$ do $D[0,j] = \infty$;
for $i = 1$ to $M$ do
  for $j = 1$ to $N$ do
    $c[i,j] = \min(D[i-1,j-1] + d(q_i, v_j), D[i,j-1] + d[q_i, v_j])$;
return $c[M,N]$
```

Example 4. Given the query clip $Q = (5,3,2,8,3)$ and one video clip $V = (2,6,7,1,6,10,4)$, the computation process of the distance of optimal subsequence mapping with replication is described in Table 4. The mapping is shown in Fig. 5.

Similar to the similarity measure DOM in symmetric measures, we also defined the distance-constrained optimal subsequence mapping (DOSM) in asymmetric measures.

Definition 17. Given the query clip $Q = (q_1, q_2, \ldots, q_M)$, the video clip $V = (v_1, v_2, \ldots, v_N)$, the frame distance tolerance δ and the distance $d(q_i, v_j)$ $\forall i,\ 1 \leqslant i \leqslant M$ $\forall j,\ 1 \leqslant j \leqslant N$, a distance-constrained subsequence mapping between them is a one-to-one relation R_{DSM} from $\{1,2,\ldots,M\}$ to $\{1,2,\ldots,N\}$, such that

(1) for each ordered pair (i,j) in R_{DSM}, $d(u_i, v_j) \leqslant \delta$;
(2) for each i, $1 \leqslant i \leqslant M$, there exists one j, $1 \leqslant j \leqslant N$, such that $(i,j) \in R_{\text{DSM}}$;
(3) for any two ordered pairs $(i,j), (k,l)$ in R_{DSM}, $(j < l)$ if and only if $(i < k)$.

Definition 18. Given the query clip $Q = (q_1, q_2, \ldots, q_M)$, the video clip $V = (v_1, v_2, \ldots, v_N)$, the frame distance constraint δ, and the distance $d(q_i, v_j)$ $\forall i,\ 1 \leqslant i \leqslant M$, $\forall j,\ 1 \leqslant j \leqslant N$, the distance between Q and V for a given mapping R_{DSM}, $D'_{R_{\text{DSM}}}(Q, V, \delta)$, is defined as

Table 4
Computation of optimal subsequence mapping with replication of Example 4

		2	*6*	*7*	*1*	*6*	*10*	*4*
	0	∞	∞	∞	∞	∞	∞	∞
5	∞	**3**	**4**	6	10	11	16	17
3	∞	∞	6	**8**	**8**	11	18	17
2	∞	∞	∞	11	**9**	12	19	20
8	∞	∞	∞	∞	18	**11**	**13**	17
3	∞	∞	∞	∞	∞	21	18	**14**

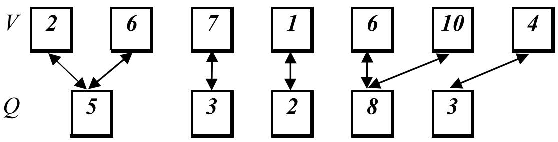

Fig. 5. Optimal subsequence mapping with replication of Example 4.

$$D'_{R_{\mathrm{DSM}}}(Q,V,\delta) = \sum_{\forall(i,j)\in R_{\mathrm{DSM}}} d(q_i,v_j).$$

Definition 19. Given the query clip $Q=(q_1,q_2,\ldots,q_M)$ and the video clip $V=(v_1,v_2,\ldots,v_N)$, the distance between Q and V for *distance-constrained optimal subsequence mapping* (DOSM) is defined as

$$D_{\mathrm{DOSM}}(Q,V,\delta) = \min_{\forall R_{\mathrm{DSM}}}\{D'_{R_{\mathrm{DSM}}}(Q,V,\delta)\}.$$

Note that, in DOSM, it is not necessary to add the penalty to the distance. The reason comes from the fact that, according to Definition 17, each frame of query sequence must be mapped with a frame of video. If there is no such mapping, the distance $D_{\mathrm{DOSM}}(Q,V,\delta)$ is set to ∞.

Algorithm (*Distance-constrained optimal subsequence mapping* (*DOSM*)).

```
$D[0,0] = 0$;
for $i = 1$ to $M$ do $D[i,0] = \infty$;
for $j = 1$ to $N$ do $D[0,j] = 0$;
for $i = 1$ to $M$ do
    for $j = 1$ to $N$ do
        if $d(q_i,v_j) < \delta$ then
            $D[i,j] = \min(D[i-1,j-1] + d(q_i,v_j), D[i,j-1])$;
        else $D[i,j] = D[i,j-1]$;
return $D[M,N]$
```

Example 5. For the same example as Example 4, the frame distance constraint δ being 3, the computation process of the distance of distance-constrained optimal subsequence mapping is described in Table 5. Fig. 6 shows the mapping.

Table 5
Computation of distance-constrained optimal subsequence mapping of Example 5

		2	*6*	*7*	*1*	*6*	*10*	*4*
	0	0	0	0	0	0	0	0
5	∞	**3**	1	1	1	1	1	1
3	∞	∞	**6**	**6**	3	3	3	2
2	∞	∞	∞	∞	**7**	**7**	7	5
8	∞	∞	∞	∞	∞	**9**	**9**	9
3	∞	∞	∞	∞	∞	∞	∞	**10**

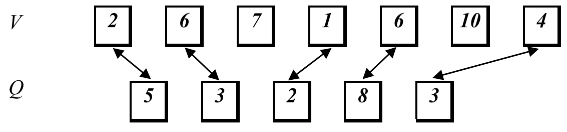

Fig. 6. Distance-constrained optimal subsequence mapping of Example 5.

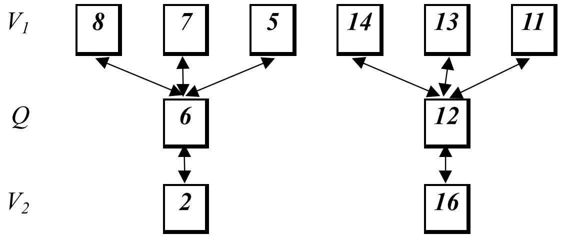

Fig. 7. $R_{\mathrm{OMR}}(Q, V_1)$ and $R_{\mathrm{OMR}}(Q, V_2)$.

2.3. Normalization of similarity measures

In application of the proposed algorithms, it is necessary to normalize two measures. One is the distance measure between key frames for distance-constrained algorithms. The other is the similarity measure between sequence of key frames for algorithms allowing frame replication.

Distance-constrained algorithms are helpful for video retrieval. However, one problem that arises is that users have no idea about the frame distance constraint. People seldom describe color, texture and shape similarities quantitatively. It is difficult for users to specify the constraint of frame distance.

If the frame distance is a normalized measure, it will be helpful for users in specifying the distance constraint. In the previous section, we have described that the color feature is represented as a normalized color histogram. Moreover, the texture and shape distance are also normalized by the inverse variances for each texture, shape component. Therefore, these normalized distances are represented as real numbers which range from 0 to 1.

In addition, sometimes, the threshold values of video segmentation and/or key-frame selection are also useful for users. These threshold values give users hints for the selection of the frame distance constraint. This depends on the availability of the threshold values. For example, threshold values are available in a video information system incorporating functions of uncompressed video segmentation and non-uniform sampling of key frame. The general approach of uncompressed video segmentation is to detect scene changes by detecting significant quantitative color differences between frames. Non-uniform sampling measures the difference between the last selected key frame and the remaining frames in the clip. Frames with considerable variations are selected as key frames.

Concerning the normalization of similarity measures between sequences of frames, it is unfair for video with more number of frames to be measured by the algorithms allowing replications. For example, considering two video sequence V_1, V_2 and the query sequence Q shown in Fig. 7, the distances $D_{\mathrm{OMR}}(Q, V_1)$ and $D_{\mathrm{OMR}}(Q, V_2)$ both equal 8. However, intuitively, V_1 is more similar to Q, though there are more frames in V_1. To deal with this unfairness, we normalize the distance by the cardinality of relation R_{OMR}. In this example, the normalized $D_{\mathrm{OMR}}(Q, V_1)$ becomes 8/6 while the normalized $D_{\mathrm{OMR}}(Q, V_2)$ becomes 8/2.

3. Experiments

To evaluate the performance of the proposed similarity measures and previous work done by Adjeroh et al. (1999), we performed two experiments for the symmetric case. The first experiment measures the performance for video scenes of consecutive shots while the second measures that for video sequence of consecutive scenes. In each experiment, we have a database of 24 video clips. These video clips were extracted from movies “Analyze This”, “Dr. Dolittle”, “Titanic”, “Wild Wild West”, “Saving Private Ryan”, and “Visual Singer”. In the first experiment, the video clips are video scenes between adjacent scene breaks. In the second experiment, the video clips are video sequences of consecutive scenes. For both experiments, each scene is decompressed first and is segmented into consecutive shots. For each shot, a sequence of key frames is extracted by the process of non-uniform key-frame extraction presented in (Yeung and Liu, 1995). Each key frame is represented as a 78-bin color histogram in HSV color space. The number of key frames ranges from 3 to 6 and 5 to 21 for the first and second experiment, respectively.

The video clips in the database are used as the query video clip. That is, there are 24 query clips for each experiment. Each query video clip is represented as a sequence of key frames.

We measure the performance by precision and recall. Precision is defined as the proportion of retrieved video that is relevant, while recall is defined as the relevant video retrieved. The ground truth to determine relevant video is judged by humans. For each query, the ground truth is established by choosing and ranking a list of relevant clips from the video database. Besides the proposed similarity measure, we also implement the measure proposed by Adjeroh et al. with alphabet size 4 and normalization of NORM2 (Adjeroh et al., 1999). For DOM, the distance constraint is set to 0.15. Using each of these similarity measure, each query returns L video clips, $L = 1, \ldots, 20$ and the 20 precision–recall pairs are calculated. Then, the average precision and recall values are derived using the 11-point average which averages precision at 11 recall points. There exist a wide variety of measures for evaluating retrieval performance in visual information retrieval research community (Smith, 1998). We adopt the 11-point average as 11 points average each quantity performance with a single score.

Figs. 8 and 9 show the average precision–recall curves for the first and second experiment, respectively. In Fig. 8, it can be seen that all the four measures perform well. OM performs better when the recall is less than 0.5. However, when the recall increases from 0.7 to 1, the precision of DOM is better while that of OM is the smallest. Despite that, there is no significant difference among these four measures. This is expected since, in the first experiment, frames of consecutive shots are similar where the effect of temporal ordering is not very significant.

On the other hand, in Fig. 9, the precision–recall curves of these measures are more different. DOM performs best among the four different measures. This is because DOM prevents unqualified frame

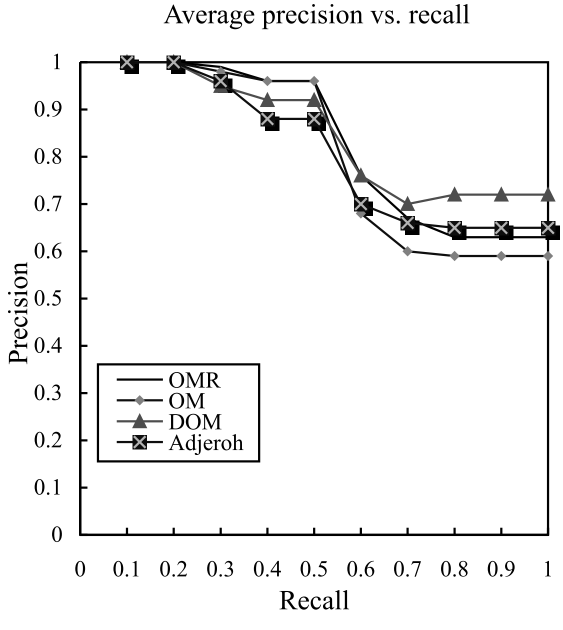

Fig. 8. The average precision–recall curve for the first experiment.

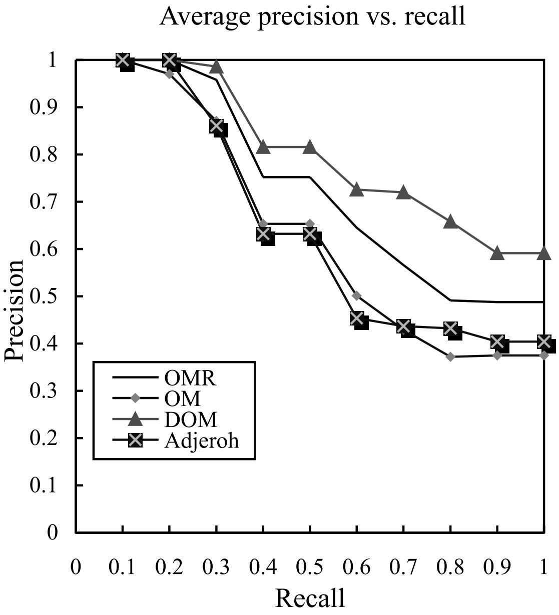

Fig. 9. The average precision–recall curve for the second experiment.

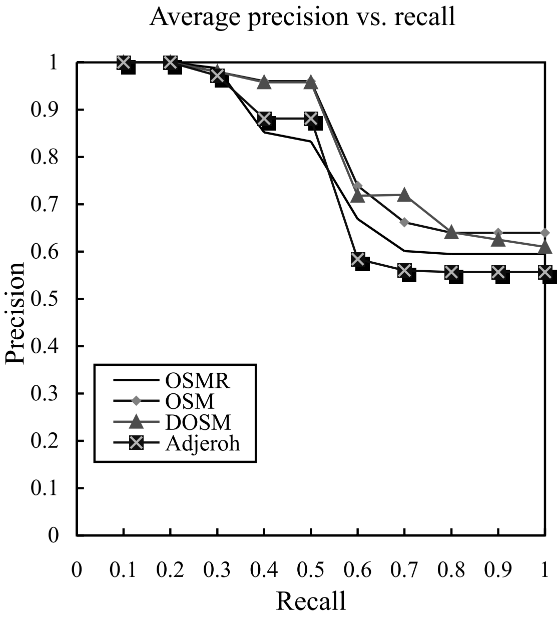

Fig. 10. The average precision–recall curve for the asymmetric case.

mapping. The performance of OMR is next to that of DOM. The performance of the measure proposed by Adjeroh et al. is similar to that of OM. One of the possible reasons that Adjeroh's approach is inferior to OMR comes from the swap operation. In the second experiment with consecutive scenes, users are sensitive to the temporal ordering of frames.

For the asymmetric case, the database and query clips of the second experiment are utilized. However, some of the shots are randomly eliminated from query video. Moreover, each of the remaining shot is represented by only one key frame. Fig. 10 shows the precision–recall curve for the asymmetric case.

For the asymmetric case, it can be seen that distance-constrained-based measure DOSM still performs best. OSM performs better than OSMR. This phenomenon can be realized as follows. OSM is one-to-one mapping and leaves out unmatched frames while OSMR is one-to-many mapping. However, from the human perception's point of view, two key-frame sequences are similar only if there are similar frames between them. Therefore, OSM behaves like human perception.

4. Conclusions

In this paper we have proposed a series of video similarity measures based on similarity of frame sequence. Similarity algorithms based on the approach of dynamic programming are also presented. In fact, the performance is highly dependent on the extraction process of video content. We plan to measure performance by considering the effect of the content extraction process. Furthermore, in this paper, the frame sequence is discussed in the uncompressed domain. Another future work is the extension of the proposed similarity measures to the compressed domain.

References

Adjeroh, D.A., Lee, M.C., King, I., 1999. A distance measure for video sequences. Computer Vision and Image Understanding 75 (1–2), 25–45.

Aoe, J.I. (Ed.), 1994. Computer Algorithms: String Pattern Matching Strategies. IEEE Computer Society Press, Los Alamitos.

Dimitrova, N., Abdel-Mottaled, M., 1997. Content-based video retrieval by example video clip. In: Proc. Storage and Retrieval for Image and Video Databases V, IS&T/SPIE Symp. Electronic Imaging Science and Technology, San Jose, CA. pp. 59–70.

Flickner, M., et al., 1995. Query by image and video content: the QBIC system. IEEE Comput. 28 (9), 23–32.

Furht, B., Smoliar, S.W., Zhang, H.J., 1995. Video and Image Processing in Multimedia Systems. Kluwer Academic Publishers, Boston, MA.

Little, T.D.C., Ahanger, G., Folz, R.J., Gibbon, J.F., 1993. A digital on-demand video service supporting content-based queries. In: Proc. ACM Multimedia'93, Anaheim, CA. pp. 427–436.

Mohan, R., 1998. Video sequence matching. In: Proc. Internat. Conf. Acoustics, Speech and Signal Processing, ICASSP'98, Seattle, WA. pp. 3697–3700.

Rowe, L.A., Boreczky, J.S., Eads, C.A., 1994. Indexes for user access to large video databases. In: Proc. Storage and Retrieval for Image and Video Databases II, IS&T/SPIE Symp. Electronic Imaging Science & Technology, San Jose, CA. pp. 150–161.

Saoke, H., Chiba, S., 1990. Dynamic programming algorithm optimization for spoken word recognition. In: Waibel, A., Lee, K. (Eds.), Readings in Speech Recognition. Morgan Kaufmann, San Mateo, CA.

Shan, M.K., Lee, S.Y., 1998. Content-based similarity measures for video based on similarity of frame sequence. In: Proc. IEEE Internat. Workshop on Multimedia Data Base Management Systems, IW-MMDBMS'98, Dayton, OH.

Smith, J.R., 1998. Image retrieval evaluation. In: Proc. IEEE Internat. Workshop on Content-based Access of Image and Video Libraries, CBAIVL-98.

Wang, Y.P., Pavlidis, T., 1990. Optimal correspondence of string subsequences. IEEE Trans. Pattern Anal. Machine Intell. 12 (11), 1080–1087.

Wu, J.K., Narasimhalu, A.D., Mehtre, B.M., Lam, C.P., Gao, Y.J., 1995. CORE: a content-based retrieval engine for multimedia information systems. Multimedia Syst. 3 (1), 25–41.

Yeung, M.M., Liu, B., 1995. Efficient matching and clustering of video shots. In: Proc. Internat. Conf. Image Processing'95, Washington, DC. pp. 338–341.

Zhang, H.J., Kankanhalli, A., Smoliar, W., 1993. Automatic partitioning of full-motion video. Multimedia Syst. 1 (1), 10–28.

Zhang, H.J., Low, C.Y., Smoliar, S.W., Wu, J.H., 1995. Video parsing, retrieval and browsing: an integrated and content-based solution. In: Proc. ACM Multimedia'95, San Francisco, CA. pp. 15–24.

Zhang, H.J., Tan, S.Y., Smoliar, S.W., Gong, Y., 1995. Automatic parsing and indexing of news video. Multimedia Syst. 2 (6), 256–265.

Zhang, H.J., Zhong, D., Smoliar, S.W., 1997. An integrated system for content-based video retrieval and browsing. Pattern Recognition 30 (4), 643–658.

Zhong, D., Zhang, H.J., Chang, S.F., 1995. Clustering methods for video browsing and annotation. In: Proc. IS&T/SPIE Storage and Retrieval for Image and Video Databases IV, San Jose, CA. pp. 239–246.