VIDEO CLIP RETRIEVAL BY MAXIMAL MATCHING AND OPTIMAL MATCHING IN GRAPH THEORY

Yu-Xin Peng [1], Chong-Wah Ngo [2], Qing-Jie Dong [1], Zong-Ming Guo [1], Jian-Guo Xiao [1]

[1] Institute of Computer Science and Technology
Peking University
Beijing 100871, China
peng_yuxin@icst.pku.edu.cn

[2] Department of Computer Science
City University of Hong Kong
Tat Chee Avenue, Kowloon, HK
cwngo@cs.cityu.edu.hk

ABSTRACT

In this paper, a novel approach for automatic matching, ranking and retrieval of video clips is proposed. Motivated by the maximal and optimal matching theories in graph analysis, a new similarity measure of video clips is defined based on the representation and modeling of bipartite graph. Four different factors: visual similarity, granularity, interference and temporal order of shots are taken into consideration for similarity ranking. These factors are progressively analyzed in the proposed approach. Maximal matching utilizes the granularity factor to efficiently filter false matches, while optimal matching takes into account the visual, granularity and interference factors for similarity measure. Dynamic programming is also formulated to quantitatively evaluate the temporal order of shots. The final similarity measure is based on the results of optimal matching and dynamic programming. Experimental results indicate that the proposed approach is effective and efficient in retrieving and ranking similar video clips.

1. INTRODUCTION

Due to the drastic advances in internet and multimedia applications such as digital library, video-on-demand and distance learning, an effective yet efficient way of retrieving relevant video data is a highly challenging issue. In broad, we can categorize the video retrieval techniques into (a) shot-based retrieval and (b) clip-based retrieval. A shot is typically defined as a series of frames with continuous camera motion. A video clip, on the other hand, is a series of shots that are coherent from the narrative point of view. To date, most approaches are developed for shot-based retrieval due to its simplicity. Relatively few works have been done for clip-based retrieval [1][2][3][4][5][6]. Compared to shot-based retrieval, clip-based retrieval is relatively robust in term of retrieval accuracy since video clips normally consist of more meaningful and concise information. For most casual users, an input query to a video database should be a video clip rather than just one single video shot.

Clip-based retrieval, in general, is built upon the shot-based retrieval. Besides relying on the visual similarity between shots, clip-based retrieval should consider the inter-relationship such as the granularity, temporal order and interference factors among video shots. In brief, the similarity measure between video clips should include:

1. *Visual similarity* of shots between two video clips.
2. *Granularity*: One shot in a clip may be similar to more than one shot in the other one. The similarity among the video shots of two clips could be many-to-many, many-to-one, one-to-many mapping. Certain objective criteria are required to model different cases. For instance, two clips with many-to-one relationship should be given smaller similarity value.
3. *Temporal order*: Two clips with similar visual information but different temporal order among shots should not be considered as dissimilar. However, two clips with similar visual content and temporal order should be given higher similarity value than two clips with similar visual information but different temporal order.
4. *Interference factor*: Some shots in a clip may not be similar to any shot in the other one. This factor should affect the final similarity value.

To date, most existing approaches on clip-based retrieval are based on dynamic programming [3] and ad-hoc assumption [1][2][3][4]. The disadvantages of these approaches include slow matching time and ill-defined shot similarity measures. In [1][2][3], temporal order is imposed as a hard constraint. In other words, similar clips must obey the same temporal order. As a result, video clips with similar content but different shot order will not be retrieved. In contrast to [1][2][3], the proposed approach in [4] ignored the influence of temporal order and granularity. The clip similarity depends mainly on the number of matching shots. As a consequence, the similarity of two clips with one-to-one relationship may be same as two clips with one-to-many relationship. In [5][6], the proposed approach takes into account the visual, temporal order, temporal duration, interference and granularity factors. The final similarity is based on the weighted linear combination of these factors. The similarity measures due to different factors are heuristically defined without validation. And, it requires manual segmentation of video sequences into clips prior to clip matching.

The proposed approach, as in [5][6], will take different factors into consideration. However, in contrast to [5][6], the similarity measure will be formulated directly based on the representation and modeling of bipartite graph. An obvious advantage is that the effectiveness of the proposed approach can be verified through maximal and optimal matching theories in graph analysis. In addition, instead of adopting linear

combination to integrate similarity measures due to different factors, our similarity measure is carried out in a progressive manner. Because the similar clips due to dynamic programming (DP) are always a subset of similar clips due to optimal matching (OM), while the similar clips due to OM are always a subset of similar clips due to maximal matching (MM). These two properties allow us to effectively prune dissimilar items step by step without missing any potential candidates. In addition, because MM is computationally faster than OM, these two properties also allow us to considerably speed up the computational time of the proposed approach. Besides similarity measure, the proposed approach could also automatically identify the boundaries of similar video clips without prior manual segmentation.

The rest of this paper is organized as follows. In section 2, we focus on the overview of the proposed approach. Shot-based retrieval is described in section 3. Clip segmentation and MM algorithm are presented in section 4. OM and DP for similarity measure are given in section 5. Experimental results are shown in section 6 and conclusion remarks are provided in section 7.

2. OVERVIEW OF THE PROPOSED APPROACH

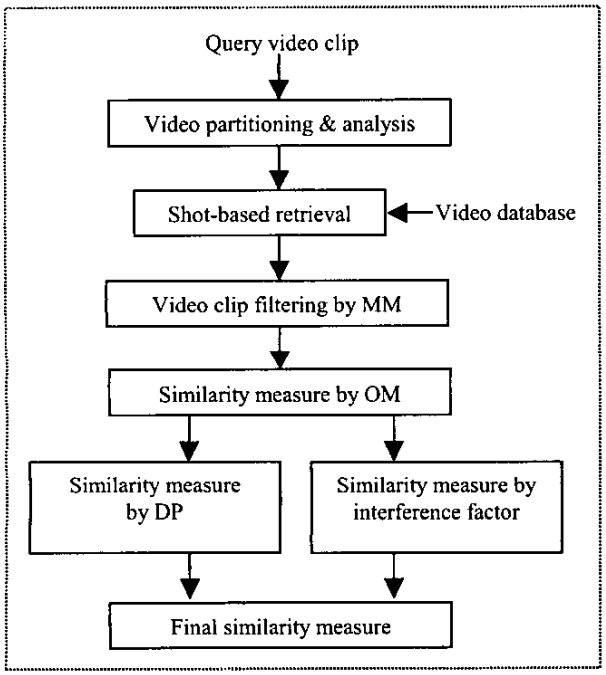

Figure 1: Proposed framework

Figure 1 illustrates the proposed framework. Initially a query video clip is partitioned into shots through spatio-temporal slices analysis [7]. Keyframes are extracted and constructed to represent the content of video shots [8]. The shot similarity measure is based on a result of comparing the keyframes of query clip and video database [8]. By modeling the continuity of the similar shots, candidates of similar video clips are segmented from video database. To improve speed efficiency, MM is then adopted to filter false candidates and retain only the similar video clips for further investigation. OM measures the visual similarity between query clip and its similar clips by guaranteeing the one-to-one mapping among video shots. Based on the output of OM, DP measures the temporal order of shots between two clips, and interference factor is also further measured. The final similarity measure is based on the results of OM and DP.

The novelty of the approach is mainly based on MM and OM [9][10][11] of video clips, on one hand, to speed up matching time, on the other hand, to enforce one-to-one mapping between two clips for effective similarity measure. Both MM and OM are classical problems in graph theory. Let $G = \{X, Y, E\}$ as a bipartite graph, where $V = X \cup Y$ is the vertex set, $E = \{e_{ij}\}$ is the edge set. A matching M of G is a subset of the edges with the property that no two edges of M share the same node. Given the unweighted bipartite graph G, as illustrated in Figure 2, MM is to find a matching M that has as many edges as possible. OM is basically an extension of MM, by assigning weight to every edge in G, as illustrated in Figure 3. OM is to find the matching M that has the largest total weight.

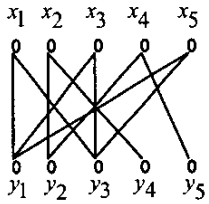

Figure 2: Maximal matching

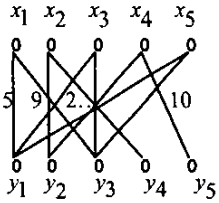

Figure 3: Optimal matching

In the proposed approach, the matching process is represented by two sets of shots X and Y_k in a bipartite graph G_k. Let $X = \{x_1, x_2, \ldots, x_n\}$ as a query clip and $Y_k = \{y_1, y_2, \ldots, y_m\}$ as a segmented video clip in the video database, e_{ij} represents the shot similarity between x_i and y_j. The purpose of MM is to quickly eliminate dissimilar clips Y_k with X by just counting the number of edges of one-to-one matching in G_k. While OM does not use the matching obtained by MM, it assigns weight e_{ij} to every edge in the original graph G_k, and measure the clip similarity between X and Y_k by maximizing the total value of e_{ij}.

3. SHOT-BASED RETRIEVAL

The approach in [8] is employed to measure similarity for all pairs of shots between X and video sequences Y in the database. A shot x_i matches y_j if their similarity value exceeds a threshold T. Since the purpose of T is to prune most of the incorrect matches while ensuring high recall rate, T can be set as low as possible. T will not be sensitive to the final results because dissimilar shots which pass T are very difficult to meet the constraint of the number of edges in one-to-one matching (See section 4). According to the experimental result, T=0.5 is set.

4. VIDEO CLIP FILTERING

After we retrieve the similar shots y_j, we sort y_j in an ascending order. If $|y_{j+1} - y_j| > D$, the video sequences Y can then be segmented into video clips Y_k. In the approach, we set D=2.

Because Y_k includes dissimilar clips and similar clips with X, MM based on Hungarian algorithm [9][10][11] is then adopted to filter dissimilar clips by counting the number of edges in one-to-one matching M . The original Hungarian algorithm is modified as follows for the application:

1. $M \leftarrow \phi$.
2. If all vertices in X have been tested, M is the maximal matching of G_k and the algorithm ends. Otherwise, go to next step.
3. Find a vertex $x_i \in X$ where x_i has not been tested. Let $A \leftarrow \{x_i\}, B \leftarrow \phi$. A and B are different sets.
4. If $N(A) = B$, x_i can not join M , label x_i as tested, go to step 2, otherwise go to next step. $N(A) \subseteq Y_k$ corresponds to the set of vertices that matches the vertices in set A.
5. Find a vertex $y_j \in N(A) - B$.
6. If $(y_j, z) \in M$, let $A \leftarrow A \cup \{z\}, B \leftarrow B \cup \{y_j\}$, then go to step 4. Otherwise go to next step.
7. There exists an augmenting path P from x_i to y_j , let $M \leftarrow M \oplus E(P)$, label x_i as tested, go to step 2.

The computational complexity of Hungarian algorithm is $O(n \cdot |E_k|)$ where $|E_k|$ is the number of edges in G_k . In the approach, a pair of clips is considered as similar if $|M| \geq \lceil n/2 \rceil$. These pairs are retained for OM and DP for similarity measure.

5. SIMILARITY MEASURE OF VIDEO CLIPS

5.1. Optimal matching

After we retrieve the similar clips Y_k , OM based on Kuhn-Munkres algorithm [9][11] is employed to measure similarity between X and Y_k by assigning weights e_{ij} to edges in the original graph G_k , The algorithm is as follows:

1. Start with initial label of $l(x_i) = \max_j e_{ij}, l(y_j) = 0, i, j = 1,2...,t$, $t = \max(n,m)$
2. Find edge set $E_l = \{(x_i, y_j) \mid l(x_i) + l(y_j) = e_{ij}\}$, $G_l = (X, Y_k, E_l)$ and the matching M in G_l
3. If M consists of all vertices in X, M is the optimal matching of G_k and the algorithm ends. Otherwise, go to next step.
4. Find a vertex $x_i \in X$ where x_i is not inside M. Let $A \leftarrow \{x_i\}$ and $B \leftarrow \phi$. A and B are different sets.
5. If $N_{G_l}(A) = B$, then go to step 9, otherwise go to next step. $N_{G_l}(A) \subseteq Y_k$ corresponds to the set of vertices that matches the vertices in set A.
6. Find a vertex $y_j \in N_{G_l}(A) - B$
7. If $(y_j, z) \in M$, let $A \leftarrow A \cup \{z\}, B \leftarrow B \cup \{y_j\}$, then go to step 5. Otherwise go to next step.
8. There exists an augmenting path P from x_i to y_j , let $M \leftarrow M \oplus E(P)$, and go to step 3.
9. Compute the value of

$$a = \min_{\substack{x_i \in A \\ y_j \notin N_{G_l}(A)}} \{l(x_i) + l(y_j) - e_{ij}\}$$

, then construct a new label l' by

$$l'(v) = \begin{cases} l(v) - a & v \in A \\ l(v) + a & v \in B \\ l(v) & otherwise \end{cases}$$

Compute the value of $E_{l'}, G_{l'}$ according to l'

10. Replace $l \leftarrow l', G_l \leftarrow G_{l'}$, go to step 6.

The computational complexity of Kuhn-Munkres algorithm is $O(t^3)$, where $t = \max(n,m)$. After that, we can get the optimal matching M and the total value ω of e_{ij} in M. The visual factor is defined as follows:

$$Vision = \frac{\omega}{n}$$

5.2. Dynamic programming

Based on the output of OM, we use dynamic programming (DP) to further measure the clip similarity based on the temporal order of shots. The algorithm is as follows:

$$c[i,j] = \begin{cases} 0 & i = 0, \text{ or } j = 0 \\ c[i-1, j-1] + 1 & i,j > 0, (x_i, y_j) \in M \\ \max(c[i, j-1], c[i-1, j]) & i,j > 0, (x_i, y_j) \notin M \end{cases}$$

The computational complexity of DP is $O(nm)$. The temporal order factor is calculated as follows:

$$order = \frac{c[i,j]}{n}$$

5.3. Interference factor

At last, some vertices of X and Y_k are not in the optimal matching M. The interference factor is computed as follows:

$$Interference = \frac{2 \cdot |M|}{n+m}$$

5.4. Final similarity measure

Based on the above analysis, the final similarity between X and Y_k can be calculated as follows:

$$Similarity(X, Y_k) = \omega_1 \cdot Vision + \omega_2 \cdot Order + \omega_3 \cdot Interference$$

Where ω_1 , ω_2 , ω_3 are the weights of different factors, in the system, we assign ω_1 =0.4, ω_2 =0.3, ω_3 =0.3.

6. EXPERIMENTAL RESULTS

We conducted the experiment on TV programs of 191 minutes, with totally 4714 shots (286936 frames). The video database is very challenging because it contains a diversity of video programs, including news, commercials, movie, sports, etc. There are many repeated video clips, such as the commercials and the news logo, and it also contains some similar video clips, such as the commercials with different length and order, and different tennis ball games. To evaluate the performance of the

proposed approach, we implement the approach in [4] for comparison purpose. Beside the precision and recall, we also compare the retrieval speed of the two approaches. The test is performed on a computer of dual CPU PIII-1G, 256M RAM.

Figure 4 shows an example of retrieving and ranking similar video clips with query clip. It shows different editions of a commercial. In every row, we use keyframes to represent the video clip. The clip in the first row is the query clip, and the others are the similar clips found by the system. As can be seen, these results are similar to the query clip, and they are ranked in a descensive order of similarity. The ranking clips embody the subjective visual judgement of human. For example, the first result is the query clip itself because its similarity is certainly the highest. In addition, the three video clips at the top are more similar than the two clips at the bottom with the query clip in temporal order.

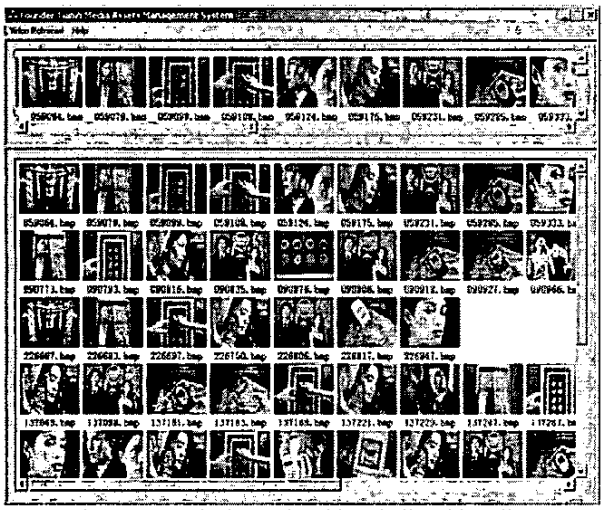

Figure 4: Retrieving and ranking similar video clips

We first test the two approaches on exact video clip retrieval using five different queries. One query uses the news logo, one uses the football games in news, and the other three cases are from three different commercials.

The Proposed Approach			Approach in [4]		
Precision	Recall	Speed	Precision	Recall	Speed
100%	100%	107s	81.7%	100%	148s

Table 1: Performance on exact video clip retrieval

Table 1 gives the experimental results, the results show that the two approaches achieve the same average recall of 100%, but in average precision and retrieval speed, the proposed approach is better than the approach in [4].

Then, we test the two approaches on similar video clip retrieval using five different queries. One query uses the tennis ball game, one uses the scene of doctors salving patient in the hospital, and the other three cases are from three different commercials.

The Proposed Approach			Approach in [4]		
Precision	Recall	Speed	Precision	Recall	Speed
92%	87.1%	100s	87.1%	70%	189s

Table 2: Performance on similar video clip retrieval

In table 2, the experimental results indicate that the proposed approach achieves better performance than the approach in [4] in average precision, recall and retrieval speed. In addition, as shown in Figure 4, the approach is better in ranking similar clips because we take different factors into consideration but the approach in [4] ignored the influence of these factors.

7. CONCLUSIONS

In this paper, we have proposed a novel approach for video clip retrieval and ranking based on maximal matching and optimal matching in graph theory. Four different factors: visual similarity, granularity, interference and temporal order of shots are taken into consideration for similarity ranking. These factors are progressively analyzed in the proposed approach. First, maximal matching utilizes the granularity factor to efficiently filter false matches. Then, optimal matching and dynamic programming measure the clip similarity based on the four factors. Experimental results indicate that the proposed approach is effective and efficient in retrieving and ranking similar video clips.

8. ACKNOWLEDGEMENT

The work described in this paper was partially supported by a grant from City University of Hong Kong (Project No. 7001470).

9. REFERENCES

[1] N. Dimitrova, and M. Abdel-Mottaled, "Content-based Video Retrieval by Example Video Clip," In *SPIE Proceeding: Storage and Retrieval of Image and Video Databases VI*, Vol. 3022, pp. 184-196, 1998.

[2] A. K. Jain, A. Vailaya, and W. Xiong, "Query by Video Clip," *Multimedia System*, 7, pp. 369-384, 1999.

[3] Y. P. Tan, S. R. Kulkarni, and P. J. Ramadge, "A Framework for Measuring Video Similarity and Its application to Video Query by Example," *Int. Conf. On Image Processing*, Vol.2, pp. 106-110, 1999.

[4] L. Chen, and T.-S. Chua, "A Match and Tilting Approach to Content-based Video Retrieval," *Int. Conf. On Multimedia and Expo*, 2001.

[5] X. Liu, Y. Zhuang , and Y. Pan, "A New Approach to Retrieve Video by Example Video Clips,", *ACM Multimedia*, 1999.

[6] Y. Wu, Y. Zhuang, and Y. Pan, "Content-based Video Similarity Model," *ACM Multimedia*, 2000.

[7] C. W. Ngo, T. C. Pong, and R. T. Chin, "Video Partitioning through Temporal Slices Analysis," *IEEE Trans on Circuits and Systems for Video Technology*, Vol. 11, No. 8, pp. 941-953, 2001.

[8] C. W. Ngo, T. C. Pong, and H. J. Zhang, "Motion-based Video Representation for Scene Change Detection," *Int. Journal of Computer Vision*, Vol. 50, No. 2, Nov 2002.

[9] W.-S. Xiao, *Graph Theory and Its algorithms*, Beijing Aviation Industrial Press, 1993.

[10] J. A. McHugh, *Algorithmic Graph Theory*, Prentice Hall, 1990.

[11] L. Lovasz, and M. D. Plummer, *Matching Theory*, Amsterdam: North Holland, 1986.