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 candidates 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

In the proposed approach, the matching process is represented by two sets of shots X and Yk in a bipartite graph Gk. Let X = {x1, x2, …, xn} as a query clip and Yk = {y1, y2, …, ym} as a segmented video clip in the video database, ej represents the shot similarity between xj and yj.

The purpose of MM is to quickly eliminate dissimilar clips Yk with X by just counting the number of edges of one-to-one matching in Gk. While OM does not use the matching obtained by MM, it assigns weight ej to every edge in the original graph Gk, and measure the clip similarity between X and Yk by maximizing the total value of ej.

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 xj matches yj 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 yj, we sort yj in an ascending order. If \( \left| y_{j+1} - y_j \right| > D \), the video sequences Y can then be segmented into video clips Yk. 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 \( O(n^2) \), where \( t = \max(n, m) \). After that, we can get the optimal matching \( M \) and the total value \( \omega \) 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:

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

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:

\[ Interference = \frac{2}{n + m} \]

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 \( \omega_1, \omega_2, \omega_3 \) are the weights of different factors, in the system, we assign \( \omega_1 = 0.4, \omega_2 = 0.3, \omega_3 = 0.3 \).

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 descending 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.

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