Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems | School of Computing and Information Systems

10-2004

Clip-based similarity measure for hierarchical video retrieval

Yuxin PENG

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the Databases and Information Systems Commons, and the Graphics and Human Computer Interfaces Commons

Citation

PENG, Yuxin and NGO, Chong-wah. Clip-based similarity measure for hierarchical video retrieval. (2004). *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, New York, October 15-16*. 53-60.
Available at: https://ink.library.smu.edu.sg/sis_research/6507

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Clip-based Similarity Measure for Hierarchical Video Retrieval

Yuxin Peng[1,2] and Chong-Wah Ngo[2]

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

[2] Department of Computer Science
City University of Hong Kong
cwngo@cs.cityu.edu.hk

ABSTRACT

This paper proposes a new approach and algorithm for the similarity measure of video clips. The similarity is mainly based on two bipartite graph matching algorithms: maximum matching (MM) and optimal matching (OM). MM is able to rapidly filter irrelevant video clips, while OM is capable of ranking the similarity of clips according to the visual and granularity factors. Based on MM and OM, a hierarchical video retrieval framework is constructed for the approximate matching of video clips. To allow the matching between a query and a long video, an online clip segmentation algorithm is also proposed to rapidly locate candidate clips for similarity measure. The validity of the retrieval framework is theoretically proved and empirically verified on a video database of 21 hours.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models.
General Terms: Algorithms, Experimentation, Theory.
Keywords: Clip-based similarity, hierarchical video retrieval.

1. INTRODUCTION

Due to the drastic advances in multimedia and Internet applications, the effective techniques for video retrieval and summarization are increasingly demanded. One critical component in these techniques is the similarity measure of visual information. While the issues in shot-based similarity have been intensively addressed for retrieval, clustering and summarization, clip-based similarity remains a difficult problem that has not yet been fully exploited. In this paper, we propose a hierarchical framework based on the bipartite graph matching algorithms for the similarity filtering and ranking of video clips.

A shot is a series of frames with continuous camera motion, while a clip is a series of shots that are coherent from the narrative point of view. A clip usually conveys one meaningful event. Shot-based retrieval is useful for tasks like the detection of known objects and certain kinds of videos like sports. For most general videos, retrieval based on a single shot, may not be practical since a shot itself is only a part of an event and does not convey full story. For most casual users, query-by-clip is definitely more concise and convenient than query-by-shot.

Existing approaches in clip-based retrieval include [1-9, 12, 16, 17]. Some researches focus on the rapid identification of similar clips [2, 3, 5, 7, 12], while the others focus on the similarity ranking of videos clips [1, 4, 6, 8, 9, 16, 17]. In [2, 3, 7], fast algorithms are proposed by deriving signatures to represent the clip contents. The signatures are basically the summaries or global statistics of low-level features in clips. The similarity of clips depends on the distance between signatures. The global signatures are suitable for matching clips with almost identical content but little changes due to compression, formatting, minor editing in spatial or temporal domain. One successful example is the high accuracy and speed in retrieving commercial clips from large video databases [7].

In [1, 4, 6, 8, 9, 17], clip-based retrieval is built upon the shot-based retrieval. Besides relying on shot similarity, clip similarity is also dependent on the inter-relationship such as the temporal order, granularity and interference among shots. In [4, 6], shots in two clips are matched by preserving their temporal order. These approaches may not be appropriate since shots in different clips tend to appear in various orders due to editing effects. Even a commercial video, several editions are normally available with various shot order and duration.

One sophisticated approach for clip retrieval is proposed in [9, 17] where different factors including temporal order, granularity and interference are taken into account. Granularity models the degree of one-to-one shot matching between two clips, while interference models the percentage of unmatched shots. A cluster-based algorithm is employed to match similar shots. The aim of clustering is to find a cut (or threshold) that can maximize the centroid distance of similar and dissimilar shots. The cut value is used to decide whether two shots should be matched. A slightly similar approach to [9, 17] is [8]. A threshold value is predefined to determine the matching of shots. Two measures, re-sequence and correspondence, are used to assess the similarity of clips. The correspondence measure can partially evaluate the degree of granularity.

Most approaches [2, 3, 4, 6, 8, 9, 16, 17] assume video clips are pre-segmented and always available for matching. As a result, the matching and ranking of multiple similar

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MIR'04, October 15–16, 2004, New York, New York, USA.
Copyright 2004 ACM 1-58113-940-3/04/0010 ...$5.00.

instances in a long recorded video is not supported. In addition, most algorithms [1, 2, 3, 4, 5, 6, 8, 9, 16, 17] does not incorporate the capability of filtering irrelevant clips prior to similarity ranking. The retrieval speed could be seriously affected particularly for large video database.

In this paper, we propose a hierarchical retrieval framework focusing mainly on the similarity ranking of video clips. Our proposed similarity measure is in line with [9, 17], but with the capabilities of clip filtering and online segmentation. Instead of adopting cluster-based algorithm as in [9, 17], we formulate the problem of shot matching as a bipartite graph matching in two stages. In the first stage, the candidate clips are located and segmented from videos while the irrelevant clips are rapidly filtered. In the second stage, the detailed similarity ranking is conducted by considering the quality of matching determined jointly by the granularity, temporal order and interference factors. The major contributions of our approach are as follows.

- *Matching and filtering.* We adopt two bipartite graph matching algorithms, namely maximum matching (MM) and optimal matching (OM), for the matching of shots in clips. Both algorithms are constrained under one-to-one mapping. MM, by computing the maximum cardinality of matching, is capable of rapidly filtering irrelevant clips. OM, by optimizing the total weight of matching, is able to rank relevant clips based on the similarity of visual and granularity. MM and OM can thus form a hierarchical framework for filtering and retrieval. By the definitions of MM and OM [10, 18], the validity of the hierarchical framework can be justified by showing that MM will never filter clips that are considered similar by OM.
- *Similarity ranking.* The clip similarity is jointly determined by visual, granularity, order and interference factors. While visual and granularity are measured by OM, temporal order similarity is evaluated effectively by the dynamic programming. The measure of interference is based on the output of OM.
- *Online segmentation.* The segmentation of videos into clips is implicitly tailored to the content of a query clip. Given a query and a video, a bipartite graph is constructed by many-to-many mapping. The mapping usually results in the following properties: Some shots in the video are densely matched along the temporal dimension, while most shots are sparsely matched or unmatched. Our algorithm will automatically locate the dense regions as potential candidate clips.

The remaining of this paper is organized as follows. Section 2 describes the preprocessing steps including the shot boundary detection and shot similarity measure. Section 3 presents the proposed clip-based similarity measure by MM and OM. Section 4 justifies the validity of the hierarchical video retrieval framework formed by MM and OM. The algorithm for online video clip segmentation is also presented. Section 5 presents the experimental results.

2. VIDEO PREPROCESSING

The preprocessing includes shot boundary detection, keyframe representation and shot similarity measure. We adopt the detector in [13] for the partitioning of videos into shots. Motion based analysis in [14] is then employed to select and construct keyframes for each shot. For instance, a sequence with pan is represented by a panoramic keyframe, while a sequence with zoom is represented by two frames before and after the zoom.

Let the keyframes of a shot s_i be $\{r_{i1}, r_{i2}, \ldots\}$, the similarity between two shots is defined as

$$Sim(s_i, s_j) = \frac{1}{2}\{\phi(s_i, s_j) + \hat{\phi}(s_i, s_j)\} \quad (1)$$

where

$$\phi(s_i, s_j) = \max_{p=\{1,2,\ldots\}, q=\{1,2,\ldots\}} \text{Intersect}(r_{ip}, r_{jq})$$

$$\hat{\phi}(s_i, s_j) = \hat{\max}_{p=\{1,2,\ldots\}, q=\{1,2,\ldots\}} \text{Intersect}(r_{ip}, r_{jq})$$

The similarity function $\text{Intersect}(r_{ip}, r_{jq})$ is the color histogram intersection of two keyframes r_{ip} and r_{jq}. The function $\hat{\max}$ returns the second largest value among all pairs of keyframe comparisons. The histogram is in HSV color space. Hue is quantized into 18 bins while saturation and intensity are quantized into 3 bins respectively. The quantization provides 162 ($18 \times 3 \times 3$) distinct color sets.

3. CLIP-BASED SIMILARITY

The similarity is mainly based on maximum matching (MM) and optimal matching (OM). Both MM and OM are classical matching algorithms in graph theory [10, 18]. MM computes the maximum cardinality matching in an unweighted bipartite graph, while OM optimizes the maximum weight matching in a weighted bipartite graph.

3.1 Notation

For the ease of understanding, we use the following notations in the remaining paper:

- Let $X = \{x_1, x_2, \ldots, x_p\}$ as a query clip with p shots and x_i represents a shot in X.
- Let $Y_k = \{y_1, y_2, \ldots, y_q\}$ as the k^{th} video clip with q shots in a video $\mathbf{Y}$ and y_j is a shot in Y_k.
- Let $G_k = \{X, Y_k, E_k\}$ as a bipartite graph constructed by X and Y_k. $V_k = X \cup Y_k$ is the vertex set while $E_k = \{\omega_{ij}\}$ is the edge set. For an unweighted graph, $\omega_{ij} = \{0, 1\}$ and 1 represents there is an edge (or a match) from shot x_i to shot y_j. For a weighted graph, ω_{ij} represents the shot similarity between x_i and y_j.

3.2 Video Clip Filtering by MM

Given X and Y_k, an unweighted bipartite graph G_k is formed by

$$\omega_{ij} = \begin{cases} 1 & Sim(x_i, y_j) > \mathcal{T} \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

The function Sim is based on Eqn(1). A threshold [1] $\mathcal{T}$ is set to determine whether there is an edge from x_i to y_j. Since a clip is composed of a series of shots with same semantic, the color content of shots are usually inter-correlated and similar. Because of this self-similarity property, one shot in

[1]To ensure high recall rate, the value of $\mathcal{T}$ is set as low as possible. $\mathcal{T}$ will not be sensitive to the final matching since the edges of dissimilar shots will not meet the one-to-one constraint in MM and will be filtered ultimately.

X can usually match multiple shots in Y_k. As a consequence, the mapping of shots in G_k is usually the many-to-many relationship. To maximize the matching of shots between X and Y_k under the one-to-one mapping constraint, MM is used due to its effectiveness and efficiency. The output of MM is a bipartite graph G_{MM} with each x_i matches with at most one y_j and vice versa. Based on the number of edges in G_{MM}, we can rapidly filter dissimilar video clips while retain only potentially relevant clips for the detailed similarity ranking. In general, if only few shots in X can match Y_k, Y_k should be considered as dissimilar to the query clip X. In our case, we define two clips as dissimilar if $|M| < \frac{|X|}{2}$, where $|M|$ is the number of edges in G_{MM} and $|X| = p$ is the number of shots in a query clip.

We employ maximum cardinality matching algorithm (Kuhn algorithm) for the implementation of MM [18]. The details are given in Figure 1. The computational complexity of MM is $O(nm)$, where $n = p + q$ is the number of vertices (shots) and m is the number of edges in G_k.

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

Figure 1: Algorithm for Maximum Matching.

3.3 Video Clip Ranking

3.3.1 Optimal Matching (OM)

Based on a weighted bipartite graph G_k formed by applying $\mathcal{T}$ as in Eqn(2), OM is employed to maximize the total weights of matching under the one-to-one mapping constraint. The output of OM is a weighted bipartite graph G_{OM} where one shot in X can match with at most one shot in Y_k and vice versa. The similarity of X and Y_k is assessed based on the total weight in G_{OM} as follows

$$Sim_{OM}(X, Y_k) = \frac{\sum \omega_{ij}}{p} \tag{3}$$

where the similarity is normalized by the number of shots p in the query clip X. The implementation of OM is based on Kuhn-Munkres algorithm [18]. The details are given in Figure 2. The running time of OM is $O(n^4)$ where $n = p + q$ is the total number of vertices in G_k.

3.3.2 Dynamic Programming (DP)

Given a bipartite graph G_{OM} computed by OM, the similarity of two clips based on the temporal order of shot matching can be formulated by DP. Denote $\mathcal{C}$ as a cost matrix indicating the number of shot pairs that are matched along

1. Start with the initial label of $l(x_i) = \max_j\{\omega_{ij}\}$ and $l(y_j) = 0$, where $i, j = 1, 2, \ldots, t$ and $t = \max\{p, q\}$.
2. Compute $E_l = \{(x_i, y_j) | l(x_i) + l(y_j) = \omega_{ij}\}$, $G_l = (X, Y_k, E_l)$ and one matching M in G_l.
3. If M contains all the vertices in X, M is the optimal matching of G_k and the algorithm ends. Otherwise, goto step 4.
4. Find a vertex $x_i \in X$ and x_i is not inside M. Set $A \leftarrow \{x_i\}$ and $B \leftarrow \emptyset$, where A and B are different sets.
5. Let $N_{G_l}(A) \subseteq Y_k$ as the set of vertices that matches the vertices in set A. If $N_{G_l}(A) = B$, then goto step 9. Otherwise, goto step 6.
6. Find a vertex $y_j \in N_{G_l}(A) - B$.
7. If $(z, y_j) \in M$, set $A \leftarrow A \cup \{z\}$, $B \leftarrow B \cup \{y_j\}$ and goto step 5. Otherwise, goto step 8.
8. There exists an augmenting path P from x_i to y_j. Set $M \leftarrow M \oplus E(P)$ and goto step 3.
9. Compute $a = \min_{x_i \in A, y_j \notin N_{G_l}(A)}\{l(x_i) + l(y_j) - \omega_{ij}\}$, then construct a new label $\hat{l}(v)$ by
$$\hat{l}(v) = \begin{cases} l(v) - a & v \in A \\ l(v) + a & v \in B \\ l(v) & \text{otherwise} \end{cases}$$
Compute $E_{\hat{l}}$, $G_{\hat{l}}$ based on $\hat{l}$.
10. Set $l \leftarrow \hat{l}$, $G_l \leftarrow G_{\hat{l}}$, goto step 6.

Figure 2: Algorithm for Optimal Matching.

the temporal order, we have

$$\mathcal{C}[i, j] = \begin{cases} 0 & i = 0 \text{ or } j = 0 \\ \mathcal{C}[i-1, j-1] + 1 & i, j > 0, (x_i, y_j) \in M \\ \max\{\mathcal{C}[i, j-1], \mathcal{C}[i-1, j]\} & i, j > 0, (x_i, y_j) \notin M \end{cases} \tag{4}$$

where M is the optimal matching. The running time of Eqn(4) is $O(pq)$, where p and q are respectively the number of shots in X and Y_k. The similarity between two clips based on the temporal order is defined as

$$Sim_{DP}(X, Y_k) = \frac{\mathcal{C}[p, q]}{p} \tag{5}$$

3.3.3 Interference Factor (IF)

Interference factor counts the number of unmatched shots in G_{OM}, i.e., $p + q - 2 \times |M|$. The similarity between two clips based on IF is

$$Sim_{IF}(X, Y_k) = \frac{2 \times |M|}{p + q} \tag{6}$$

Since the values of $|M|$, p and q are known, $Sim_{IF}(X, Y_k)$ can be computed in $O(1)$ time.

3.3.4 Clip Similarity

Given X and Y_k, the similarity is measured jointly by the degree of granularity (and visual similarity), temporal order of matching and interference factor as follows

$$Sim_{clip}(X, Y_k) = \sum_{i \in \{OM, DP, IF\}} \alpha_i Sim_i(X, Y_k) \tag{7}$$

where $\sum_i \alpha_i = 1$ are the weights of different similarity measures. The value of α_i controls the ranking of similar video clips. In most video retrieval related tasks, the degree of granularity and visual similarity which reflect respectively the number and proximity of matching shots, should carry more weight than temporal order and interference factor.

Thus, we set $\alpha_{OM} > \alpha_{DP} = \alpha_{IF}$ ($\alpha_{OM} = 0.4$, $\alpha_{DP} = \alpha_{IF} = 0.3$) in our experiments. These values can also be set based on user preference.

4. VIDEO RETRIEVAL

The retrieval of video clip can be conducted by the similarity measure based on OM, DP and IF. However, since the total complexity is $O((p+q)^4)+O(pq)+O(1)$ for each comparison, the algorithm is inefficient particularly in a large video database. The properties of MM and OM, nevertheless, allow us to effectively set up a hierarchical framework for efficient retrieval. Since the complexity of MM is $O((p+q)m)$, it can be employed to rapidly filter irrelevant video clips. The combination of OM+DP+IF has higher time complexity but is more effective in similarity measure. They can serve to rank only a few clips retained by MM.

4.1 Hierarchical Framework

To construct the hierarchical framework, we need to show that MM will not filter any video clip that will not be filtered by OM as well. In other words, if $\mathcal{R}_1$ and $\mathcal{R}_2$ are the sets of similar clips retained by MM and OM respectively, then $\mathcal{R}_2 \subseteq \mathcal{R}_1$. If the claim is correct, the hierarchical framework is not only efficient but also as effective as using OM, DP and IF alone. The claim can be proved based on the definition of MM and OM in graph theory as follows.

DEFINITION 1. Denote a bipartite graph as $G = \{X, Y, E\}$, $M \subseteq E$ is a matching if any two edges in M are not adjacent.

DEFINITION *2.* Suppose $\hat{M}$ contains the matched pairs in G and satisfies Definition 1. Then $\hat{M}$ is the maximum cardinality matching if there exists no matching M in G such that $|M| > |\hat{M}|$.

DEFINITION *3.* Let $\hat{M}$ contains the matched pairs in G and satisfies Definition 1. Then $\hat{M}$ is the optimal matching in G if

$$\Omega(\hat{M}) = \max\{\Omega(M)|M \text{ is a matching in } G\} \quad (8)$$

where

$$\Omega(M) = \sum_{\omega_{ij} \in M} \omega_{ij} \quad (9)$$

is the sum of similarity in M and ω_{ij} is the similarity between shot i and shot j.

THEOREM *1.* Let $|MM|$ as the number of edges by maximum matching (MM) and $|OM|$ as the number of edges by optimal matching (OM). Then

$$|MM| \geq |OM| \quad (10)$$

PROOF. By Eqn(8) in Definition 3, OM is a matching with maximum weight. This implies that OM is also a cardinality matching. Hence, based on Definition 2, we have $|MM| \geq |OM|$. □

THEOREM *2.* Let $\mathcal{R}_1$ as the set of video clips retained by MM, and $\mathcal{R}_2$ as the set of video clips retained by OM. Then

$$\mathcal{R}_2 \subseteq \mathcal{R}_1 \quad (11)$$

PROOF. Denote p as the number of shots in a query clip, and $1 \leq \lambda \leq p$ is a parameter such that $\frac{p}{\lambda}$ decides if a clip in the database should be filtered. If $|OM| \geq \frac{p}{\lambda}$, then $|MM| \geq \frac{p}{\lambda}$ since $|MM| \geq |OM|$ by Theorem 1. Thus, $\mathcal{R}_2 \subseteq \mathcal{R}_1$. □

In setting the hierarchical framework, λ is a parameter that controls the number of clips to be retained for OM. If the value of λ is large, the response time of a query will be slow. In our implementation, the λ is set to 2 as mentioned in Section 3.2.

4.2 Video Clip Segmentation

In a video database, clips are not always available for retrieval. While shots boundaries can be readily located and indexed, clips boundaries are relatively harder to be obtained since the detection of boundaries usually involves a certain degree of semantic understanding. The decomposition of videos into semantic clips is, in general, a hard problem. In this paper, instead of *explicitly* locating the boundaries of clips prior to video retrieval, we propose an *implicit* approach that exploits the inherent matching relationship between a given query and videos for online clip segmentation.

Given a query clip X and a video $\mathbf{Y}$ (usually $|\mathbf{Y}| \gg |X|$), a bipartite graph is constructed by matching the shots in X to the shots in $\mathbf{Y}$ by Eqn(2). The mapping is many-to-many relationship, i.e., a shot can map to multiple shots in $\mathbf{Y}$ as long as they are considered similar based on the definition in Eqn(2). Denote $\zeta_j = \{0, 1\}$ to indicate whether a shot j in $\mathbf{Y}$ is matched by a shot in X. The mapping usually forms a number of dense and sparse clusters (with $\zeta_j = 1$ represents a match) along the one dimensional space of ζ. The dense clusters indicate the presence of potentially similar video clips in $\mathbf{Y}$ with the query clip, while the sparse clusters can probably mean the noisy matching.

One straightforward way of implicit clip segmentation is to extract the dense clusters directly from the $1D$ ζ space. To do this, we need two parameters (ρ, ϑ) where ρ specifies how to extract a cluster while ϑ specifies how to filter sparse clusters. The algorithm is formulated as follows. We check the distance d between all adjacent shots with $\zeta_j = 1$. All the adjacent shots with $d \leq \rho$ are grouped in one cluster. In other words, the shot at the boundary of a cluster has at least $\rho+1$ consecutive unmatched shots with other clusters. Once the clusters $Y_{k=\{0,1,\ldots\}}$ are extracted, we filter those clusters whose $|Y_k| < \vartheta$.

In the experiments, we set $\rho = 2$ and $\vartheta = \frac{|X|}{2}$. A large value of ρ can cause under-segmentation, while a small value of ρ can cause over-segmentation of video clips. The value of ρ is not easy to set, however, when $\rho = \{2, 3, 4, 5\}$, the setting mostly yield satisfactory results for our database of approximately 21 hours' videos and 20,000 shots. The value of ϑ is set based on λ described in Theorem 2. Since $\lambda = 2$, any clip with $|Y_k| < \frac{|X|}{2}$ can never satisfy $|MM| \geq \frac{p}{\lambda}$ and thus should not be considered.

A major advantage of our approach is that the segmentation is always tailored to the content of a query clip. Only those clips related to query will be segmented for retrieval. However, an implicitly segmented video clip may not be a precise scene or story since its boundary may contain shots from other clips and, furthermore, the clip itself could probably be composed of more than one clip due to under-

Table 1: Comparison between Liu's approach and ours.

	Liu's Approach	*Ours*
Features	color histogram, Tamura texture	color histogram
Video clips	manually segmented	online automatically segmented based on the content of query
Similarity factors	cluster-based matching, temporal order, speed, disturbance, congregate	optimal matching, temporal order, interference factor
Video clip filtering	linear combination, a manually optimized threshold is set to filter irrelevant clips	based on MM, $\lambda = 2$ as in Theorem 2
Video clip ranking	five weighting factors are manually optimized in the database	linear combination, three weights are set as in Eqn(7)

Table 2: Experimental results for video clip filtering and retrieval

Query type	# of queries	Average # of shot	Average # of relevant clip	Our Approach		Liu's Approach	
				Precision	Recall	Precision	Recall
Commercial	20	12.9	4.0	0.935	1.000	0.628	0.990
News	20	18.5	4.2	0.794	0.735	0.649	0.622
Sport	10	15.0	5.1	0.765	0.684	0.601	0.509
Average	-	15.5	4.4	0.831	0.806	0.626	0.707

segmentation. Some of these deficiencies, auspiciously, can be got rid of during the similarity ranking of optimal matching. OM can be utilized not only to match similar shots, but also to split a clip and refine its boundary. Given a video clip $Y_k = \{y_1, y_2, \ldots, y_q\}$ and a query clip X, suppose only the shots $\{y_\alpha, \ldots, y_\beta\}$ are matched with X, and $1 < \alpha < \beta < q$. The unmatched shots $Y_{k'} = \{y_1, \ldots, y_{\alpha-1}\}$ and $Y_{k''} = \{y_{\beta+1}, \ldots, y_q\}$ can be pruned if $|\alpha - 1| < \frac{|X|}{\lambda}$ and $|q - \beta| < \frac{|X|}{\lambda}$ respectively. Otherwise, $Y_{k'}$ and $Y_{k''}$ are split from Y_k as the new clips for similarity ranking by OM.

5. EXPERIMENTS

To evaluate the performance of the proposed hierarchical framework, we set up a database that consists of approximately $1,272$ minutes (more than 21 hours) of videos. The genres of videos include news, sports, commercials, movies and documentaries collected from different TV stations. In total, there are $19,929$ shots.

We compare our approach with Liu's approach in [9]. The major difference between these two approaches are summarized in Table 1. In [9], a clustering based algorithm is used to decide the matching of shots in two clips. The aim of the algorithm is to cluster the pairwise similarities of shots into two groups which correspond to the matched and unmatched shots. This is achieved by maximizing the centroid distance between two groups. Based on the matched shots, the temporal order, speed (duration difference), disturbance (number of unmatched shots) and congregation (number of one-to-one mapping) are computed for similarity measure. In our approach, the matching of shots and the degree of congregation are measured directly by OM. Dynamic programming is employed to measure the temporal order of two sequences. In [9], this value is measured by calculating the percentage of matching shots that are in reverse order. Our interference factor is same as disturbance, and we do not use speed since duration is not a critical factor in reflecting similarity particularly when the unmatched shots are available.

Liu's approach [9] assumes that the video clips are presegmented and always available for retrieval. As a result, we manually segment the 21 hours' videos into clips, and in total, there are 1288 segmented video clips in our database. In the experiment, while the results of [9] is based on the retrieval of manually segmented video clips, our approach adopts the online automatic segmentation described in Section 4.2 for retrieval.

All the relevant video clips in the database are manually judged and grouped by human subjects. We experiment various types of query for performance evaluation. These queries include clips from commercials, news and sports videos. We compare the performance of both approaches in term of clip filtering and clip ranking capabilities.

5.1 Video Clip Filtering

We use precision and recall to measure the performance. The recall and precision are defined as follows:

$$\text{Precision} = \frac{\text{Number of relevant clips being retained}}{\text{Number of clips being retained}}$$

$$\text{Recall} = \frac{\text{Number of relevant clips being retained}}{\text{Number of relevant clips}}$$

In [9], no mechanism is proposed for the filtering of irrelevant clips. During the implementation, we set an optimized threshold for this purpose. We systematically try different threshold values and select the one which gives the best overall recall and precision in our database as the threshold.

Table 2 shows the experimental results of both approaches. In total, 50 queries are used for testing. The average number of shots in each query is 15.5, while the average number of relevant clips is 4.4. The commercial retrieval is relatively easy since the visual content of the relevant commercial clips is usually similar and the major differences are in temporal order and duration due to different shot composition. Both approaches achieve high recall, but our approach gets better precision. Compared with commercial clips, the effective retrieval of news and sport video clips is harder because different newscasts tend to report a same event with different camera shootings and editions. In addition, more shots will

Table 3: Experimental results for the filtering and retrieval of news clips (MM+OM+DP+IF)

	Query clip	Shot #	Relevant Clip #	Our Approach		Liu's Approach	
				Precision	Recall	Precision	Recall
1	Power cut accident in London	18	7	1.000	0.429	1.000	0.286
2	Bus bomb event in Israel	12	6	1.000	0.667	0.429	0.500
3	Six-way talk about North Korea	45	6	0.833	0.833	1.000	0.167
4	The death of an Iraq aga in bomb	29	6	0.714	0.833	1.000	0.667
5	New finance policy	33	6	0.500	0.833	0.800	0.667
6	UK premier follows investigation	15	5	0.200	0.600	0.167	0.400
7	Taiwan politic issue	22	4	1.000	0.750	0.600	0.750
8	National singing competition	22	4	1.000	0.750	1.000	0.750
9	Iraq Policy	10	4	1.000	0.500	0.235	1.000
10	Resignation of a UK official	11	4	0.500	0.500	0.250	0.500
11	CCTV program promotion	11	4	1.000	1.000	0.667	1.000
12	Chinese vice president meets foreigners	19	4	0.333	1.000	0.750	0.750
13	Iraq war	9	3	1.000	0.667	1.000	0.333
14	A UN official died in Iraq	17	3	1.000	0.667	0.222	0.667
15	New policies in the ministry of police	21	3	1.000	1.000	0.667	0.667
16	Soccer association election	16	3	1.000	0.667	0.400	0.667
17	Match for solar energy bus	23	3	1.000	0.667	0.500	0.333
18	Report about blaster virus	8	3	0.400	0.667	1.000	0.667
19	Conflict between Israel and Palestine	17	3	1.000	0.667	1.000	0.667
20	Intel CEO visits China	11	2	0.400	1.000	0.286	1.000
	Average	18.5	4.2	0.794	0.735	0.649	0.622

generally be included for a clip reported in a news program of longer duration. The details of news and sport queries are listed in tables 3 and 4. Experimental results indicate that our proposed approach is, in overall, superior to Liu's approach in term of recall and precision, particularly in the retrieval of news and sports video clips collected from different TV channels. By manually investigating the retrieval results, we find that the superiority of our approach is mainly due to: 1) effectiveness of online clip segmentation in removing the sparse clusters of clips from graph matching; 2) capability of MM in filtering large amount of irrelevant clips; 3) capability of OM and DP in clip ranking.

Figures 3 and 4 show the retrieval results of news query #2 and sport query #4 respectively (due to the limitation of space, we do not show all the shots). Compared with commercial clips, the effective retrieval of news and sport clips is difficult since a same event is usually reported in different profiles, editions and camera shooting as shown in figures 3 and 4. Despite the difficulties, the proposed retrieval framework is still able to filter, match and then rank the relevant clips with reasonably good accuracy.

5.2 Video Clip Ranking

In this experiment, because our aim is to compare the ranking capability of both approaches, the MM is excluded from testing. We use AR (average recall) and ANMRR (average normalized modified retrieval rank) [11] for performance evaluation. The values of AR and ANMRR range from $[0, 1]$. A *high* value of AR denotes the superior ability in retrieving relevant clips, while a *low* value of ANMRR indicates the high retrieval rate with relevant clips ranked at the top.

Table 5 summarizes the experimental results while tables 6 and 7 shows the details of news and sport retrieval. We use same set of queries (in total 50) as in Table 2 for testing. For the retrieval of commercial clips, both approaches attain almost perfect AR and ANMRR. This implies that all relevant clips are retrieved and ranked at top. For the retrieval of news and sport events, our approach is constantly better than Liu's approach. By tracing the details of experimental results, we found that the cluster-based and temporal order algorithms used in Liu's approach can not always give satisfactory results. In contrast, the proposed clip-based similarity can rank at least half of the relevant clips at the top $C(q)$ of the retrieved clips[2].

Even though the retrieved clips by our approach are online segmented, the boundaries of most clips are precisely located. Only very few over or under-segmentation of clips happen in our test queries. On a Pentium-M 1.5GHz machine with 512M memory, the average retrieval time for a query by using OM+DP+IF is approximately 1.639 seconds. If MM+OM+DP+IF is used, the average retrieval time is 0.971 seconds. Although the MM and OM are not linear time algorithm, they are still very efficient even in large database since the online segmentation (linear time algorithm) has removed large portions of video segments from consideration before MM and OM matching.

Table 5: Experimental results for video clip retrieval

Query type	# of queries	Our Approach		Liu's Approach	
		AR	ANMRR	AR	ANMRR
Commercial	20	1.000	0.000	0.990	0.009
News	20	0.809	0.200	0.711	0.277
Sport	10	0.783	0.230	0.666	0.371
Average	-	0.864	0.143	0.789	0.219

[2]Let $NR(q)$ as the number of relevant clips for a query q, and Q as the set of queries, then

$$C(q) = \min\{4 \times NR(q), 2 \times \max_{k=1}^{Q} NR(k)\}$$

Figure 3: Retrieval results of news query #2. Query clip is listed in 1^{st} row. The correct matches are shown one row after another according to the ranked order.

Table 4: Experimental results for the filtering and retrieval of sport clips (MM+OM+DP+IF)

	Query clip	Shot #	Relevant Clip #	Our Approach		Liu's Approach	
				Precision	Recall	Precision	Recall
1	Running	14	8	0.571	0.500	0.375	0.375
2	Swimming	8	7	0.600	0.857	0.429	0.429
3	Tennis	7	6	1.000	0.500	1.000	0.500
4	Gym	10	6	0.546	1.000	0.250	0.667
5	Judo	24	6	0.600	0.500	1.000	0.167
6	Boating	16	5	0.333	0.400	1.000	0.200
7	Diving	10	4	1.000	0.750	1.000	0.750
8	Basketball	22	3	1.000	1.000	0.500	1.000
9	Volleyball	19	3	1.000	0.667	0.333	0.667
10	Weight lifting	20	3	1.000	0.667	0.125	0.333
Average		15	5.1	0.765	0.684	0.601	0.509

Table 6: Experimental results for the retrieval of news clips (OM+DP+IF)

Query clip	Our Approach		Liu's Approach	
	AR	ANMRR	AR	ANMRR
1	0.714	0.231	0.571	0.528
2	1.000	0.136	0.500	0.457
3	1.000	0.000	0.667	0.284
4	0.667	0.284	0.833	0.161
5	0.667	0.321	0.833	0.198
6	0.800	0.243	0.400	0.557
7	0.750	0.224	0.750	0.224
8	1.000	0.052	0.750	0.224
9	0.750	0.259	1.000	0.000
10	0.500	0.466	0.500	0.466
11	1.000	0.000	1.000	0.000
12	1.000	0.000	0.750	0.224
13	0.667	0.364	0.667	0.303
14	0.667	0.303	0.667	0.303
15	1.000	0.000	0.667	0.303
16	1.000	0.000	0.667	0.303
17	0.667	0.303	0.667	0.394
18	0.667	0.303	0.667	0.303
19	0.667	0.303	0.667	0.303
20	1.000	0.200	1.000	0.000
Average	0.809	0.200	0.711	0.277

6. CONCLUSIONS

We have presented the proposed algorithm for clip-based similarity measure. A hierarchical video retrieval framework have also been described and experimented. Encouraging results have been obtained through the performance evaluation in a databases with 21 hours of video. Experimental results suggest that the proposed MM is effective in filtering irrelevant clips, while OM is capable of effectively retrieving and clustering video clips of same event. The proposed retrieval matching mechanism is not only suitable for identical matching (*e.g.,* commercial clips), but also approximate matching (*e.g.,* news and sports). Although the current clip-based similarity measure considers only color features, other features such as motion and audio can also be incorporated. Currently, the implementation of MM and OM is based on Kuhn and Kuhn-Munkres algorithms which require $O(nm)$ and $O(n^4)$ respectively, where n and m are the number of shots and matching edges. Faster versions of MM and OM algorithms exists [15], for instance, MM can run in $O(\sqrt{n}m)$ and OM can run in $O(n(m + n \log n))$. In future, algorithms will be incorporated in our framework for more efficient retrieval.

Figure 4: Retrieval results of sport query #4. Query clip is listed in 1^{st} row. The correct matches are shown one row after another according to the ranked order.

Table 7: Experimental results for the retrieval of sport clips (OM+DP+IF).

Query clip	Our Approach		Liu's Approach	
	AR	ANMRR	AR	ANMRR
1	0.625	0.300	0.625	0.530
2	0.857	0.165	0.714	0.341
3	0.833	0.136	0.833	0.198
4	1.000	0.124	0.667	0.284
5	0.833	0.161	0.333	0.679
6	0.600	0.586	0.400	0.571
7	0.750	0.224	0.750	0.224
8	0.667	0.303	1.000	0.061
9	1.000	0.000	0.667	0.303
10	0.667	0.303	0.667	0.515
Average	0.783	0.230	0.666	0.371

Acknowledgement

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

7. REFERENCES

[1] L. Chen and T. S. Chua. A match and tiling approach to content-based video retrieval. In *Int. Conf. on Multimedia and Expo*, pages 417–420, 2001.

[2] S. C. Cheung and A. Zakhor. Efficient video similarity measurement with video signature. *IEEE Trans. on Circuits and Systems for Video Techn.*, 13(1), Jan 2003.

[3] S. C. Cheung and A. Zakhor. Fast similarity search and clustering of video sequences on the world-wide-web. *IEEE Trans. on Multimedia*, 2004.

[4] N. Dimitrova and M. Abdel-Mottaled. Content-based video retrieval by example video clip. In *SPIE Proc. Stograge and Retrieval of Image and Video Database VI*, volume 3022, pages 184–196, 1998.

[5] T. C. Hoad and J. Zobel. Fast video matching with signature alignment. In *Int. Workshop on Multimedia Information Retrieval*, pages 262–268, 2003.

[6] A. K. Jain, A. Vailaya, and W. Xiong. Query by video clip. In *ACM Multimedia System*, volume 7, 1999.

[7] K. Kashino, T. Kurozumi, and H. Murase. A quick search method for audio and video signals based on histogram pruning. *IEEE Trans. on Multimedia*, 5(3), Sep 2003.

[8] R. Lienhart and W. Effelsberg. A systematic method to compare and retrieve video sequences. *Multimedia Tools and Applications*, 10(1), Jan 2000.

[9] X. Liu, Y. Zhuang, and Y. Pan. A new approach to retrieve video by example video clip. In *ACM Multimedia*, 1999.

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

[11] mpeg video group. Description of core experiments for mpeg-7 color/texture descriptors. In *ISO/MPEGJTC1/SC29/WG11 MPEG98/M2819*, July 1999.

[12] M. R. Naphade, M. M. Yeung, and B. L. Yeo. A novel scheme for fast and efficient video sequence matching using compact signatures. In *SPIE: Storage and Retrieval for Media Databases*, pages 564–572, 2000.

[13] C. W. Ngo, T. C. Pong, and R. T. Chin. Video partitioning by temporal slice coherency. *IEEE. Trans. on Circuits and Systems for Video Technology*, 11(8), Aug 2001.

[14] C. W. Ngo, T. C. Pong, and H. J. Zhang. Motion-based video representation for scene change detection. *Int. Journal of Computer Vision*, 50(2), Nov 2002.

[15] A. Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*, volume A. Berlin: Springer, 2003.

[16] Y. P. Tan, S. R. Kulkarni, and P. J. Ramadge. A framework for measuring video similarity and its application to video query by example. In *ICIP*, 1999.

[17] Y. Wu, Y. Zhuang, and Y. Pan. Content-based video similarity model. In *ACM Multimedia*, 2000.

[18] W. S. Xiao. *Graph Theory and Its Algorithms*. Beijing: Aviation Industry Press, 1993.