EMD-Based Video Clip Retrieval by Many-to-Many Matching

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

[1] Institute of Computer Science and Technology,
Peking University, Beijing 100871, China
pengyuxin@icst.pku.edu.cn
[2] Department of Computer Science,
City University of Hong Kong, Kowloon, HongKong
cwngo@cs.cityu.edu.hk

Abstract. This paper presents a new approach for video clip retrieval based on Earth Mover's Distance (EMD). Instead of imposing one-to-one matching constraint as in [11, 14], our approach allows many-to-many matching methodology and is capable of tolerating errors due to video partitioning and various video editing effects. We formulate clip-based retrieval as a graph matching problem in two stages. In the first stage, to allow the matching between a query and a long video, an online clip segmentation algorithm is employed to rapidly locate candidate clips for similarity measure. In the second stage, a weighted graph is constructed to model the similarity between two clips. EMD is proposed to compute the minimum cost of the weighted graph as the similarity between two clips. Experimental results show that the proposed approach is better than some existing methods in term of ranking capability.

1 Introduction

With the drastic growth of multimedia data in internet, TV stations, enterprises and personal digital archives, an effective yet efficient way of retrieving relevant multimedia information such as video clips is a highly challenging issue. Since the past decade, numerous researches have been conducted for content-based video retrieval. Nevertheless, most works are concentrated on retrieval by single shot, rather than retrieval by multiple shots (video clip). In this paper, we proposed a new approach based on Earth Mover's Distance (EMD) for similarity measure between two 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 shot is only a physical unit, while a clip usually conveys one semantic 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 a semantic event and does not convey full story. From the entropy point of video, video clips are relatively informative and hence the retrieved items should be relatively meaningful. For most casual users, query-by-clip is definitely more concise and convenient than query-by-shot.

W.-K. Leow et al. (Eds.): CIVR 2005, LNCS 3568, pp. 71–81, 2005.

Existing approaches in clip-based retrieval include [1-14]. Some researches focus on the rapid identification of similar clips [1-6], while the others focus on the similarity ranking of video clips [7-14]. In [1, 2, 4, 6], 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, and minor editing in spatial or temporal domain. One successful example is the high accuracy and speed in retrieving commercials clips from large video database [4]. Recently, an index structure based on multi-resolution KD-tree is proposed in [6] to further speed up clip retrieval.

In [7-12, 14], 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 granularity, temporal order and interference among shots. In [8, 9, 13], 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.

Some sophisticated approaches for clip-based retrieval are proposed in [11, 12, 14] where different factors including granularity, temporal order and interference are taken into account. Granularity models the degree of one-to-one shot matching between two clips, while interference models the percentages of unmatched shots. In [11, 12], 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. In [14], a hierarchical video retrieval framework is proposed for similarity measure of video clips. Maximum matching is employed to filter irrelevant video clips, while optimal matching is utilized to rank the similarity of clips according to the visual and granularity factors. Although the approach in [14] are different with the methods in [11, 12], both employ the granularity factor to compute the clip similarity by guaranteeing the one-to-one mapping among video shots. However, one-to-one shot mapping does not always work effectively due to shot composition and video partitioning problems as follows:

- *Video editing effect.* The content of a long shot in a clip may be segmented and appeared as several shots in other editions of the clip. Some segmented shots may be deleted in certain editions. For example, a short commercial clip is displayed in golden broadcast time while its long editions are shown in other time. In addition, the same news event also has short and long editions due to the need of editing effects and the constraint of broadcast periods.
- *Shot boundary detection error.* One shot may be falsely segmented into several short shots. Several shots may also be incorrectly merged as one shot.

The composition or decomposition of shots, either due to editing effects or video partitioning errors, sometime follows the nature of many-to-many scrambling. In the case of one shot being segmented into several shots in another edition, one-to-one shot mapping can only match one shot of the edition, while other shots cannot be matched and measured. In this situation, one-to-many or many-to-many matching techniques among shots are needed to guarantee effective matching. In addition, most approaches [1, 2, 8-13] assume video clips are pre-segmented and always available for matching. As a result, the online segmentation and matching of multiple similar clips in a long video is not supported in [1, 2, 8-13].

In this paper, we propose a new approach for the similarity measure of video clips. The similarity measure is formulated as a graph matching problem in two stages. In the first stage, to allow the matching between a query and a long video, we propose an online clip segmentation algorithm to rapidly locate candidate clips for similarity measure [14]. In the second stage, the detailed similarity ranking is based on many-to-many mapping by EMD. The major contributions of our approach are *similarity ranking*. We model two clips as a weighted graph with two vertex sets: Each vertex represents a shot and is stamped with a signature (or weight) to indicate its significance during matching. The signature symbolizes the duration of a shot. EMD is then employed to compute the minimum cost of the graph, by using the signatures to control the degree of matching under many-to-many shot mapping. The computed cost reflects the similarity of clips.

The remaining of this paper is organized as follows. Section 2 describes the preprocessing steps including shot boundary detection, keyframe representation and shot similarity measure. Section 3 presents the algorithm for online video clip segmentation. Section 4 presents the proposed clip-based similarity measure by EMD. Section 5 shows the experimental results and section 6 concludes this paper.

2 Video Preprocessing

The preprocessing includes shot boundary detection, keyframe representation and shot similarity measure. We adopt the detector in [15] for the partitioning of videos into shots. Motion-based analysis in [16] 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 $\{r_{i1}, r_{i2}, ...\}$, the similarity between two shots is defined as

$$Sim(s_i, s_j) = \frac{1}{2}\left\{\phi(s_i, s_j) + \overset{\wedge}{\phi}(s_i, s_j)\right\} \tag{1}$$

where

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

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

The similarity function $Inter\sec t(r_{ip}, r_{jq})$ 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 Online Video Clip Segmentation

In video databases, 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 *implicitly* approach that exploits the inherent matching relationship between a given query and long videos for online clip segmentation [14].

Given a query clip X and a long video Y (usually $|Y| >> |X|$), an unweighted bipartite graph is constructed by matching the shots in X to the shots in Y by

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

The function *Sim* is based on Eqn (1). A threshold T is set to determine whether there is an edge from shots x_i to y_j ($\omega_{ij} = 1$ represents there is an edge from shots x_i to y_j). Since a clip is composed of a series of shots with same semantic, the color content of shots is usually inter-correlated and similar. Because of this self-similarity property, one shot in X can usually match multiple shots in Y. As a consequence, the mapping of shots in the bipartite graph is usually the many-to-many relationship. Denote $\zeta_j = \{0,1\}$ to indicate whether a shot j in 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 Y with the query clip.

One straightforward way of implicit clip segmentation is to extract the dense clusters directly from the $1D$ ζ space. To do this, a parameter ρ is needed to specify how to extract a cluster. 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.

In the experiment, ρ =2 is set. 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.

4 Clip-Based Similarity Measure

Earth Mover's Distance (EMD) has been successfully employed for image-based retrieval [17]. In this section, we will employ EMD for clip-based similarity measure. A weighted graph is constructed to model the similarity between two clips, and then EMD is employed to compute the minimum cost of the weighted graph as the similarity value between two clips.

EMD is based on the well-known *transportation problem*. Suppose some suppliers, each with a given amount of goods, are required to supply some consumers, each with a given limited capacity to accept goods. For each supplier-consumer pair, the cost of transporting a single unit of goods is given. The transportation problem is: Find a minimum expensive flow of goods from the suppliers to the consumers that satisfies the consumers' demand.

Given two clips X and Y_k, a weighted graph G_k is constructed as follows:

- Let $X = \{(x_1, \omega_{x_1}), (x_2, \omega_{x_2}), ..., (x_m, \omega_{x_m})\}$ as a query clip with m shots, x_i represents a shot in X and ω_{x_i} is the number of frames in shot x_i.
- Let $Y_k = \{(y_1, \omega_{y_1}), (y_2, \omega_{y_2}), ..., (y_n, \omega_{y_n})\}$ as the k^{th} video clip with n shots in a video *Y*, y_j represents a shot in Y_k and ω_{y_j} is the number of frames in shot y_j.
- Let $D = \{d_{ij}\}$ as the distance matrix where d_{ij} is the distance between shots x_i and y_j. In our case, d_{ij} is defined as

$$d_{ij} = 1 - Sim(x_i, y_j) \tag{3}$$

The function Sim is based on Eqn (1).

- Let $G_k = \{X, Y_k, D\}$ as a weighted graph constructed by X, Y_k and D. $V_k = X \cup Y_k$ is the vertex set while $D = \{d_{ij}\}$ is the edge set.

In the weighted graph G_k, we want to find a flow $F = \{f_{ij}\}$ where f_{ij} is the flow between x_i and y_j, that minimizes the overall cost

$$WORK(X, Y_k, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij} \tag{4}$$

subject to the following constraints:

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m, \quad 1 \leq j \leq n \tag{5}$$

$$\sum_{j=1}^{n} f_{ij} \leq \omega_{x_i} \qquad 1 \leq i \leq m \tag{6}$$

$$\sum_{i=1}^{m} f_{ij} \leq \omega_{y_j} \qquad 1 \leq j \leq n \tag{7}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} \omega_{x_i}, \sum_{j=1}^{n} \omega_{y_j}\right) \tag{8}$$

Constraint (5) allows moving frames from X to Y_k and not vice versa. Constraint (6) limits the amount of frames that can be sent by the shots in X to their weights. Constraint (7) limits the shots in Y_k to receive no more frames than their weights, and constraint (8) forces to move the maximum amount of frames. We call this amount the *total flow*. Once the transportation problem is solved, and we have found the optimal flow F, the earth mover's distance is defined as the resulting work normalized by the total flow:

$$EMD(X, Y_k) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \tag{9}$$

The normalization factor is the total weight of the smaller clip as indicated in constraint (8). Finally, the similarity between clips X and Y_k is defined as:

$$Sim_{clip}(X, Y_k) = 1 - EMD(X, Y_k) \tag{10}$$

$Sim_{clip}(X, Y_k)$ is normalized in the range of [0,1]. The higher the value of $Sim_{clip}(X, Y_k)$, the more similar the clips X and Y_k.

5 Experiments

To evaluate the performance of the proposed approach, 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. All the relevant video clips in the database are manually judged and grouped by human subjects.

We compare our approach with optimal matching in [14] and Liu's approach in [11]. The major difference among the three approaches is that our approach utilizes many-to-may shot mapping while the other two approaches employ one-to-one shot mapping. Table 1 summarizes the difference. In [11], 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 [14], the matching of shots and the degree of congregation are measured directly by optimal matching. Dynamic programming is employed to measure the temporal order of two sequences. In [11], this value is measured by calculating the percentage of matching shots that are in reverse order. The interference factor in [14] is same as disturbance in [11]. In our proposed approach, EMD is employed to transport the frames in shots between two clips. Due to the nature of many-to-many mapping among shots, the granularity, temporal order and interference factors are not applicable for clip similarity measure. Only the visual similarity based on Eqn (10) is considered in our approach.

Table 1. Comparison among our approach, optimal matching and Liu's approach

	Our approach	Optimal matching [14]	Liu's approach [11]
Features	Color histogram	Color histogram	Color histogram, Tamura texture
Video clips	Automatically segmented	Automatically segmented	Manually segmented
Similarity factors	EMD for visual similarity	Optimal matching, temporal order, interference factor	Cluster-based matching, temporal order, speed, disturbance, congregate
Shot mapping	Many-to-many	One-to-one	One-to-one
Video Clip ranking	EMD	Linear combination, three weights are set	Five weighting factors are manually optimized

Liu's approach [11] assumes that the video clips are pre-segmented and always available for retrieval. As a result, we manually segmented the 21 hours' videos into clips, and in total, there are 1,288 segmented video clips in our database. In the experiment, while the result of [11] is based on the retrieval of manually segmented video clips, our approach and optimal matching in [14] adopt the online automatic segmentation scheme described in Section 3.

Clip-based retrieval, in general, can be divided into two categories: identical matching and approximate matching. Identical matching includes commercials clips matching, and the approximate matching includes news and sports clips matching. The identical matching is relatively easy while the approximate matching is always difficult. In the experiment, we conduct testing on both kinds of matching. To assess the ranking capability of the three tested approaches, we use AR (Average Recall) and ANMRR (Average Normalized Modified Retrieval Rank) [18] 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 [18].

Table 2 summaries the experimental results for identical matching (commercial clips) while Table 3 shows the details of approximate matching (news and sport clips). In total, 40 queries are used for testing, include 20 commercials clips and 20 news and various sports clips. The commercial retrieval is relatively easy since the visual content of the relevant commercial clips is usually similar and the major differences are in the temporal order and duration due to different ways of shot composition. Overall, three 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 clips, our approach is constantly better than optimal matching and 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 cannot always give satisfactory results. Optimal matching, although better than Liu's approach, the performance is not always satisfactory due to the enforcement of one-to-one mapping among video shots. In contrast, our proposed approach can always achieve better results in term of AR and ANMRR. Furthermore, 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.

Table 2. Experimental results for the retrieval and ranking of commercial clips

Query type	# of queries	Our approach		Optimal matching		Liu's approach	
		AR	ANMRR	AR	ANMRR	AR	ANMRR
Commercial	20	1.000	0.000	1.000	0.000	0.990	0.009

Table 3. Experimental result for the retrieval and ranking of news and sport clips

Query clip	Relevant clip #	Our approach		Optimal matching		Liu's approach	
		AR	ANMRR	AR	ANMRR	AR	ANMRR
1	8	0.625	0.490	0.625	0.300	0.500	0.570
2	6	1.000	0.000	0.833	0.136	0.667	0.284
3	6	0.833	0.272	0.667	0.321	0.833	0.210
4	4	0.750	0.224	0.750	0.259	1.000	0.000
5	4	0.500	0.466	0.500	0.466	0.500	0.466
6	4	1.000	0.000	1.000	0.000	0.750	0.224
7	3	0.667	0.303	0.667	0.364	0.667	0.303
8	3	1.000	0.000	1.000	0.000	0.667	0.303
9	3	0.667	0.303	0.667	0.303	0.333	0.636
10	2	1.000	0.000	1.000	0.200	1.000	0.000
11	8	0.750	0.420	0.500	0.420	0.625	0.530
12	7	0.857	0.176	0.857	0.165	0.714	0.341
13	7	0.571	0.473	0.429	0.505	0.714	0.286
14	7	0.857	0.297	0.714	0.264	0.571	0.363
15	6	0.833	0.247	0.833	0.161	0.333	0.679
16	4	0.750	0.397	0.750	0.500	0.500	0.483
17	4	0.750	0.224	0.750	0.224	0.750	0.224
18	3	0.667	0.303	0.667	0.303	1.000	0.061
19	3	1.000	0.000	1.000	0.000	0.667	0.303
20	3	0.667	0.303	0.667	0.303	0.667	0.515
Average	4.8	0.787	0.245	0.744	0.260	0.673	0.339

Figures 1 and 2 show the retrieval and ranking results of news query #8 and sport query #11 respectively (due to the limitation of space, we do not show all the shots). Compared with commercials 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 1 and 2. Despite the difficulties, the proposed approach is still able to match and rank the relevant video clips with reasonably good results.

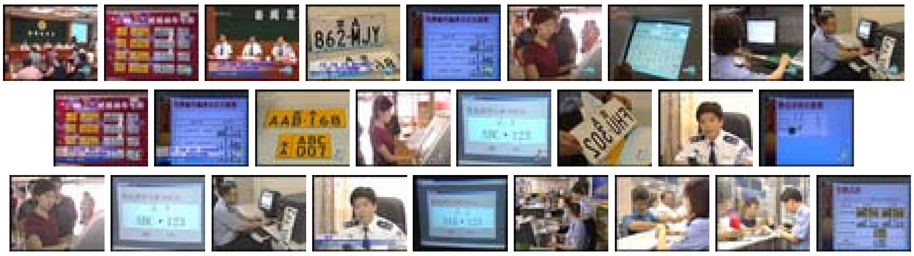

Fig. 1. Retrieval and ranking results of news query #8 (new policies in the ministry of police). Query clip is listed in 1^{st} row. The correct matches are shown one row after another according to the ranked order

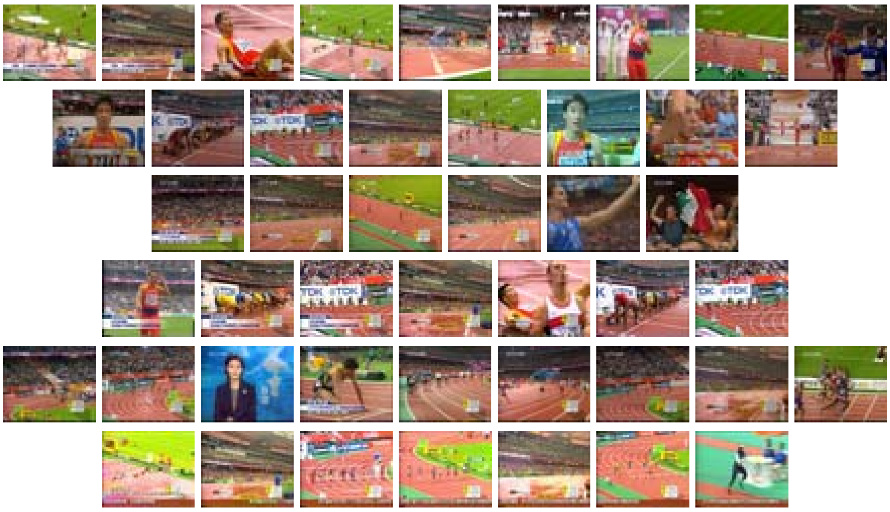

Fig. 2. Retrieval and ranking results of sport query #11 (running). Query clip is listed in 1^{st} row. The correct matches are shown one row after another according to the ranked order

6 Conclusions

We have presented a new EMD-based similarity measure for video clip retrieval. Experimental results on a 21 hours' video database indicate that EMD is capable of effectively retrieving and ranking the relevant video clips. The proposed matching mechanism is suitable not only for identical matching (*e.g.*, commercial clips), but also approximate matching (*e.g.*, news and sport clips).

Currently, we use duration (number of frames) to represent the weight (or signature) of a shot for controlling the degree of many-to-many matching. This scheme, although straightforward and yield encouraging experimental results, can be further improved if other "content indicators" such as motion and audio cues are jointly taken into account to characterize the signature of a shot.

Acknowledgements

The work described in this paper was fully supported by two grants from City University of Hong Kong (Project No. 7001470 and Project No. 7001804). In addition, we thank Yossi Rubner for his EMD code.

References

1. S. C. Cheung and A. Zakhor. Efficient Video Similarity Measurement with Video Signature. IEEE Trans. on Circuits and Systems for Video Technology, Vol. 13, No. 1, Jan, 2003.
2. S. C. Cheung and A. Zakhor. Fast Similarity Search and Clustering of Video Sequences on the World-Wide-Web. IEEE Trans. on Multimedia, 2004.
3. T. C. Hoad and J. Zobel. Fast Video Matching with Signature Alignment. ACM Int. Workshop on Multimedia Information Retrieval, pp. 262-268, 2003.
4. K. Kashino, T. Kurozumi, and H. Murase. A Quick Search Method for Audio and Video Signals based on Histogram Pruning, IEEE Trans. on Multimedia, Vol. 5, No. 3, Sep, 2003.
5. M. R. Naphade, M. M. Yeung and B. L. Yeo. A Novel Scheme for Fast and Efficient Video Sequence Matching Using Compact Signatures, SPIE: Storage and Retrieval for Media Databases, pp. 564-572, 2000.
6. J. Yuan, L.-Y Duan, Q. Tian and C. Xu. Fast and Robust Short Video Clip Search Using an Index Structure, ACM Int. Workshop on Multimedia Information Retrieval, Oct, 2004.
7. L. Chen, and T. S. Chua. A Match and Tiling Approach to Content-based Video Retrieval, Int. Conf. on Multimedia and Expo, pp. 417-420, 2001.
8. N. Dimitrova, and M. Abdel-Mottaled. Content-based Video Retrieval by Example Video Clip, SPIE: Storage and Retrieval of Image and Video Databases VI, Vol. 3022, pp. 184-196, 1998.
9. A. K. Jain, A. Vailaya, and W. Xiong. Query by Video Clip, Multimedia System, Vol. 7, pp. 369-384, 1999.
10. R. Lienhart and W. Effelsberg. A Systematic Method to Compare and Retrieve Video Sequences, Multimedia Tools and applications, Vol. 10, No. 1, Jan, 2000.
11. X. Liu, Y. Zhuang , and Y. Pan. A New Approach to Retrieve Video by Example Video Clip, ACM Multimedia Conf., 1999.
12. Y. Wu, Y. Zhuang, and Y. Pan. Content-based Video Similarity Model, ACM Multimedia Conf., 2000.
13. 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.

14. Y. Peng, C. W. Ngo. Clip-based Similarity Measure for Hierarchical Video Retrieval, ACM Int. Workshop on Multimedia Information Retrieval, Oct, 2004.
15. C. W. Ngo, T. C. Pong, and R. T. Chin. Video Partitioning by Temporal Slices Coherency, IEEE Trans. on Circuits and Systems for Video Technology, Vol. 11, No. 8, pp. 941-953, 2001.
16. 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.
17. Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. Int. Journal of Computer Vision, Vol. 40, No. 2, pp. 99-121, 2000.
18. MPEG video group. Description of Core Experiments for MPEG-7 Color/Texture Descriptors, ISO/MPEGJTC1/SC29/WG11 MPEG98/M2819, July, 1999.