Hot Event Detection and Summarization by Graph Modeling and 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 proposes a new approach for hot event detection and summarization of news videos. The approach is mainly based on two graph algorithms: optimal matching (OM) and normalized cut (NC). Initially, OM is employed to measure the visual similarity between all pairs of events under the one-to-one mapping constraint among video shots. Then, news events are represented as a complete weighted graph and NC is carried out to globally and optimally partition the graph into event clusters. Finally, based on the cluster size and globality of events, hot events can be automatically detected and selected as the summaries of news videos across TV stations of various channels and languages. Our proposed approach has been tested on news videos of 10 hours and has been found to be effective.

1 Introduction

Due to the rapidly growing amount of video collections, an effective yet efficient way for video browsing and retrieval is a highly challenging issue. Although traditional query-based retrieval is useful for known facts, it is deficient for generic retrieval such as "What happened?" or "What's new?". Suppose one person return from his office and want to know what happened in the world. Watching all the news videos in all channels is a daunting task, and query about unknown facts is unrealistic. In such situation, applications such as broadcasting hot events summarization from all channels today for users are highly demanded. In these applications, the basic problem is the hot events detection and summarization. In general, the repeated broadcast number of the same event by different channels can reflect whether an event is important and hot. To measure the number of relevant events from different channels, two basic techniques need to be developed:

- How to measure the similarity between two events?
- How to cluster the relevant news events?

In the past decade, most approaches in news video retrieval focus on the news events detection [1, 2]. To date, representative news video retrieval systems include

W.-K. Leow et al. (Eds.): CIVR 2005, LNCS 3568, pp. 257 – 266, 2005.

Informedia project [3] and VideoQA [4]. The recent work in Informedia project [3] introduced video collages as an effective interface for browsing and interpreting video collections. The system supports queries by allowing users to retrieve information through map, text and other structured information. In VideoQA system [4], users interact with VideoQA using text-based query, the system returns the relevant news fragments as the answer.

The existing news retrieval systems in [2, 3, 4] are mainly the query-based retrieval, generic retrieval such as "what's hot events today?", however, has not yet been addressed. In this paper, we propose a new approach for hot events detection and summarization. The proposed approach lies on the similarity measure of news events by optimal matching (OM), and clustering of events by normalized cut (NC) [5, 23] based on graph theory. Hot events can be automatically detected and summarized by investigating the properties of event clusters. The major contributions of our approach are as follows:

- *Similarity matching and measure.* We model two clips as a weighted bipartite graph: Every vertex in the bipartite graph represents one shot in a clip, and the weight of every edge represents the visual similarity between two shot. Then optimal matching is employed to measure the similarity between two clips according to the visual and granularity factors.
- *Highlight detection and summarization.* Based on the results of clip similarity measure by OM, all news events are represented as a complete weighted graph. Normalized cut [5, 23] is carried out to globally and optimally partition the graph into event clusters. Based on the cluster size and globality of events, hot events can be automatically detected and selected as the summarization of news videos across TV stations of various channels and languages.

Currently, our approach is based on the visual similarity for matching and clustering of news events. Multi-model features such as speech and caption cues are not considered since the broadcasts from different TV channels can be in different languages. To incorporate speech and caption recognition, multilingual translation problem need to be explicitly handled. In fact, different broadcasts of hot events, although different in term of language, and naming of person and location, partially share some common visual content that can be vividly explored for event similarity measure. In this paper, we adopt two graph-based approaches, namely OM and NC, to measure and cluster the relevant events in different channels by utilizing visual information.

2 Clip-Based Similarity Measure

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 semantic event. Existing approaches in clip-based similarity measure include [7-19]. Some researches focus on the rapid identification of similar clips [7-12], while the others focus on the similarity ranking of video clips [13-19]. In [7, 8, 10, 12], fast algorithms are pro-

posed 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 [10]. Recently, an index structure based on multi-resolution KD-tree is proposed in [12] to further speed up clip retrieval.

In [13-18], 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 [14, 15, 19], 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-based retrieval is proposed in [17, 18] 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 [17, 18], 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 this section, we propose a new approach for the similarity measure of video clips based on optimal matching (OM). Instead of adopting cluster-based algorithm as in [17, 18], we formulate the problem of shot matching as a bipartite graph matching. An obvious advantage is that the effectiveness of our proposed approach can be verified through OM in graph theory. In addition, temporal order and interference factors in [17, 18] are not considered because they will only affect the ranking but not the clustering of clips. OM is able to measure the similarity of clips under the one-to-one shot mapping constraint. Compared with commercials clips, the effective similarity measure of news events is difficult since a same event is usually reported in different profiles, editions and camera shooting. Despite the difficulties, our proposed approach is still able to match and cluster the relevant clips with reasonable results as shown in Section 5.

2.1 Video Preprocessing

The preprocessing includes shot boundary detection, keyframe representation and shot similarity measure. We adopt the detector in [20] for the partitioning of videos into shots. Motion-based analysis in [21] 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

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

where

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

$$\hat{\phi}(s_i, s_j) = \hat{\max}_{p=\{1,2,...\}, q=\{1,2,...\}} Inter\sec t\{r_{ip}, r_{jq}\}$$

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 ($18\times3\times3$) distinct color sets.

2.2 Notation

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

- Let $X = \{x_1, x_2, ..., x_p\}$ as a clip with p shots and x_i represents a shot in X.
- Let $Y = \{y_1, y_2, ..., y_q\}$ as another clip with q shots and y_j is a shot in Y.
- Let $G = \{X, Y, E\}$ as a weighted bipartite graph constructed by X and Y. $V = X \cup Y$ is the vertex set while $E = (\omega_{ij})$ is the edge set. ω_{ij} represents the shot similarity between x_i and y_j based on Eqn (1).

2.3 Optimal Matching (OM)

Given two clips X and Y, a weighted bipartite graph G is formed by applying Eqn (1). 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 and vice versa. Although the shot mapping in G_{OM} may be not unique, the total weight in G_{OM} is unique. The similarity of X and Y is assessed based on the total weight in G_{OM} as follows

$$Sim_{OM}(X, Y) = \frac{\sum \omega_{ij}}{\min(p, q)} \quad (2)$$

where the similarity is normalized by $\min(p, q)$. The implementation of OM is based on Kuhn-Munkres algorithm [6]. The details are given in Figure 1. The running time of OM is $O(n^4)$ where $n = p + q$ is the total number of vertices in G.

1. Start with the initial label of $l(x_i) = \max_j(\omega_{ij})$ and $l(y_j) = 0$, where $i, j = 1,2,...,t$ and $t = \max(p,q)$.
2. Compute $E_l = \{(x_i, y_j) \mid l(x_i) + l(y_j) = \omega_{ij}\}$, $G_l = (X, Y_k, E_l)$ 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 $B \leftarrow \phi$, where A and B are two 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 $a = \min_{\substack{x_i \in A \\ y_j \notin N_{G_l}(A)}} \{l(x_i) + l(y_j) - \omega_{ij}\}$, then construct a new label $l'(v)$ by

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

   Compute $E_{l'}, G_{l'}$ based on l'.
10. Set $l \leftarrow l', G_l \leftarrow G_{l'}$, goto step 6.

Fig. 1. Kuhn-Munkres Algorithm for Optimal Matching

3 Graph-Based Clustering

Given a set of video clips, we model the similarity among clips as a weighted undirected graph $\hat{G} = (V, E)$ where V is a set of video clips, and E is a set of edges that describes the proximity of clips. Our aim is to decompose $\hat{G}$ into sub-graphs (or clusters) so as to minimize the intra-cluster distance while maximizing the inter-cluster distance. We adopt the normalized cut algorithm [5] for the recursive bipartition of $\hat{G}$ into the clusters of clips. Normalized cut aims to globally and optimally partition a graph $\hat{G}$ into two disjoint sets A and B $(A \cup B = V)$ by minimizing

$$Ncut(A,B) = \frac{cut(A,B)}{volume(A)} + \frac{cut(A,B)}{volume(B)} \tag{3}$$

where

$$cut(A,B) = \sum_{i \in A, j \in B} Sim_{OM}(i,j) \tag{4}$$

$$volume(A) = \sum_{i \in A, j \in V} Sim_{OM}(i,j) \tag{5}$$

$cut(A,B)$ is the sum of inter-clip similarity between A and B, $volume(A)$ is the total similarity for all pairs of clips that connect A and V, and $Sim_{OM}(i,j)$ is the similarity between clips i and j based on Eqn (2). Eqn (3) can be transformed to a standard eigen system

$$D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}}z = \lambda z \tag{6}$$

where D and W are $|V| \times |V|$ matrices. D is a diagonal matrix with $D(i,i) = \sum_{j \in V} Sim_{OM}(i,j)$ and W is a symmetrical matrix with $W(i,j) = Sim_{OM}(i,j)$.

In Eqn (6), the eigen vector that corresponds to the second smallest eigen value is used to find the sets A and B. The value 0 is selected as the splitting point to divide the eigen vector into two parts that correspond to A and B respectively. The algorithm will run recursively to further bipartition the resulting sets (or clusters). The procedure terminates when the average similarity for all pairs of video clips in a cluster is below $\mu + \alpha\sigma$, where μ and σ are respectively the mean and standard deviation of all clip similarity in $\hat{G}$ and α is an empirical parameter.

4 Highlight Detection and Summarization

Based on the event clusters obtained in Section 3, highlight can be readily detected by selecting the representative clips from the clusters with large size. Assuming the skimming time S of a summary is given, we use two heuristic criterions to select the highlight from clusters:

- *Cluster size.* Highlighted events are usually repeatedly broadcasted by different TV channels at different periods of time. Therefore, the number of times an event is broadcasted is a vivid hint in deciding the highlight. Based on the skimming time constraint S, we select the clusters for highlight summarization in the descending of their cluster size.
- *Globality* of an event. An event broadcasted by different TV channels is intuitively more important than an event that is broadcasted by one channel only. Similarly, an event that is broadcasted at different periods of time (e.g., morning, afternoon, night) is more important than an event reported in a particular time of a day only. Hence, we use these two hints (the number of channels and the number of periods) that an event is broadcasted to decide the highlight, when the cluster sizes of two events are same.

For each selected cluster C, one representative clip is chosen for highlight summary. We select the clip (medoid) that is most centrally located in a cluster as representative. The medoid clip M_c is the clip whose sum of similarity with all other clips in its cluster is maximum, i.e.,

$$M_c = \max_{i \in c} \left\{ \sum_{j \in c} Sim_{OM}(i, j) \right\} \tag{7}$$

5 Experiments

We use 10 hours of news videos for testing. The videos are recorded continuously in four days from seven different TV channels. There are a total of 40 different news programs with duration ranging from 5 minutes to 30 minutes. As observed from these videos, the same events are repeatedly broadcasted in different editions and profiles by different stations. Even a same event reported in one channel, it appears differently at different time of reporting.

We manually segment the videos into clips. In total, there are 439 news clips. The numbers of events that are reported for more than one time are summarized in Table 1. In total, there are 115 clips involved in reporting 41 events. Our aim is to group news clips that describe a same event under a cluster, and then select the clusters as well as the representatives of clusters for summarization.

Table 1. The number of news events that are broadcasted for more than one time

Broadcast #	Number of events
6	3
4	5
3	11
2	22

5.1 Clustering

We employ F-measure [22] to evaluate the performance of video clip clustering. F-measure evaluates the quality of clusters by comparing the detected and ground-truth clusters. Let Q be the set of ground-truth clusters and D be the set of detected clusters, the F-measure F is given as

$$F = \frac{1}{Z} \sum_{C_i \in Q} |C_i| \max_{C_j \in D} \{ \Re(C_i, C_j) \} \tag{8}$$

$$\Re(C_i, C_j) = \frac{2 \times \mathrm{Re}\, call(C_i, C_j) \times \mathrm{Pr}\, ec(C_i, C_j)}{\mathrm{Re}\, call(C_i, C_j) + \mathrm{Pr}\, ec(C_i, C_j)} \tag{9}$$

where

$$\mathrm{Re}\,call(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i|} \tag{10}$$

$$\mathrm{Pr}\,ec(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_j|} \tag{11}$$

The term $Z = \sum_{C_i \in Q} |C_i|$ is a normalization constant. The value of F ranges [0, 1], and $F = 1$ indicates perfect clustering. By the normalized cut algorithm and clip-based similarity, we detect 291 clusters in the ten hours of videos. The value of F-measure is F =0.8225, where $|Q| = 290$ and $|D| = 291$. Table 2 shows the details of few clustering results. Some clusters such as events #1 and #3 are over-segmented into two clusters respectively. Some false clips are included due to the similarity in background color, but none of the relevant clip is missed. Because we select the medoid of a cluster as representative, false clips are not selected in video summaries. Figure 2 shows the clustering result of event #6 in Table 2. Our approach successfully

Table 2. Clustering Results of Some News Events

	News event	Number of clips in the event	Average number of shots	Final cluster(s)	Falsely included clips
1	Six-way talk about North Korea	6	55	2	2
2	New financial policy	6	22	1	2
3	The death of an Iraq aga in bomb	6	21	2	0
4	A conflict event in Iraq	4	15	1	2
5	Economic development of Beijing	4	8	1	1
6	Conflict between Israel and Palestine	3	11	1	0
7	Report about blaster virus	3	6	1	0

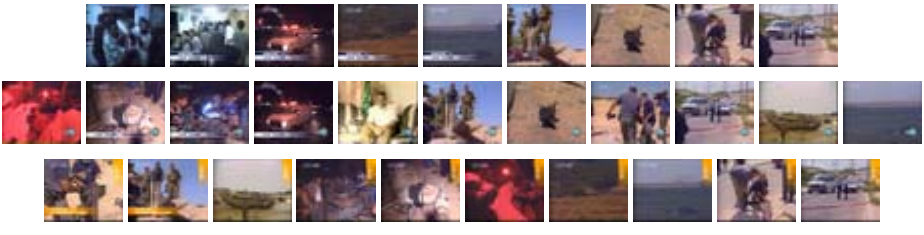

Fig. 2. The clustering results of event #6 in table 2. The three news clips are clustered correctly. The cluster medoid is listed in 2^{nd} row

groups the three video clips in one cluster although they are from three different TV channels and appear differently.

5.2 Summarization

Given a skimming time, our approach selects clusters based on the cluster size and globality of events. The medoids of selected clusters are then included in the summary. The ground-truth summary ifs manually generated in a same way based on the ground-truth clusters. For instance, when the skimming time equals to 10min, the ground-truth summary will include all the three events that are broadcasted for six times and other three events that are reported for four times (see Table 1). Table 3 shows the results of summarization. Experimental results indicate that our approach can include most of the expected events for summarization. Some events are repeated due to the over-segmentation of clusters.

Table 3. Results of summarization from videos of 10 hours

Skimming time (Minute)	Number of Expected events (Ground-truth)	Number of clips included in summary	Detected events	Missed events	Repeated events
10	6	8	4	2	0
20	11	14	8	3	0
30	24	26	21	3	0
40	39	39	31	8	1
45	41	42	34	7	2

6 Conclusions

We have presented a new approach for hot events detection and summarization. Optimal matching is employed to measure the similarity of news events, and normalized cut is employed to cluster news events. Hot events are automatically detected and summarized by investigating the properties of event clusters. The experimental results show the effectiveness of our proposed approach.

Currently, news events are detected manually. In addition, event-based similarity measure considers only color features. In future, automatic news events detection will be developed and incorporated in our system. Besides, other features such as motion and audio classes (e.g., speech, music, environmental sound and silence) can also be incorporated in the proposed approach for more effective clip-based similarity measure.

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. 7001546).

References

1. H. J. Zhang, Y. Gong, S. W. Smoliar, and S. Y. Tan. Automatic Parsing of News Video, Int. Conf. on Multimedia Computing and Systems, pp. 45-54, 1994.
2. L. Chaisorn, T. S. Chua, and C. H. Lee. The Segmentation of News Video into Story Units, Int. Conf. on Multimedia and Expo, 2002.
3. M. G. Christel, A. G. Hauptmann, H. D. Wactlar, and T. D. Ng. Collages as Dynamic Summaries for News Video, ACM Multimedia Conf., 2002.
4. H. Yang, L. Chaisorn, *et. al.*. VideoQA: Question Answering on News Video, ACM Multimedia Conf., 2003.
5. J. Shi, and J. Malik. Normalized Cuts and Image Segmentation, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, Aug, 2000.
6. W. S. Xiao, Graph Theory and Its Algorithms, Beijing Aviation Industrial Press, 1993.
7. 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.
8. S. C. Cheung and A. Zakhor. Fast Similarity Search and Clustering of Video Sequences on the World-Wide-Web. IEEE Trans. on Multimedia, 2004.
9. T. C. Hoad and J. Zobel. Fast Video Matching with Signature Alignment. ACM Int. Workshop on Multimedia Information Retrieval, pp. 262-268, 2003.
10. 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.
11. 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.
12. 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.
13. L. Chen, and T. S. Chua. A Match and Tiling Approach to Content-based Video Retrieval, Int.. Conf.. on Multimedia and Expo, 2001.
14. 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.
15. A. K. Jain, A. Vailaya, and W. Xiong. Query by Video Clip, Multimedia System, Vol. 7, pp. 369-384, 1999.
16. R. Lienhart and W. Effelsberg. A Systematic Method to Compare and Retrieve Video Sequences. Multimedia Tools and Applications, Vol. 10, No. 1, Jan, 2000.
17. X. Liu, Y. Zhuang , and Y. Pan. A New Approach to Retrieve Video by Example Video Clip, ACM Multimedia Conf., 1999.
18. Y. Wu, Y. Zhuang, and Y. Pan. Content-based Video Similarity Model, ACM Multimedia Conf., 2000.
19. 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.
20. 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, Vol. 11, No. 8, pp. 941-953, 2001.
21. 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, 2002.
22. M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering Techniques, KDD Workshop on Text Mining, 2000.
23. C. W. Ngo, Y. F. Ma, and H. J. Zhang. Video Summarization and Scene Detection by Graph Modeling, IEEE Trans. on CSVT, vol. 15, no. 2, pp. 296-305, Feb, 2005.