1-2011

Mining event structures from web videos

Xiao WU

Yi-Jie LU

Qiang PENG

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

Mining Event Structures from Web Videos

Xiao Wu, Yi-Jie Lu, and Qiang Peng
Southwest Jiaotong University

Chong-Wah Ngo
City University of Hong Kong

This article explores the issues of mining event structures from Web video search results using text analysis, burst detection, clustering, and other techniques.

With the proliferation of social media, the volume of Web videos is growing exponentially. There were 65,000 new videos uploaded to YouTube each day in July 2006. This number increased to 35 hours of videos being uploaded per minute in November 2010. When searching ''911'' on YouTube, there were 203,000 results returned in July 2007, and the number reached 430,000 in September 2008, and 593,000 in June 2009. Facing the overwhelming volume of Web videos, it becomes extremely difficult for users to find the right information. The search results are often mixed with diverse and noisy sets of video thumbnails. In addition, the dramatic growth of social media has made the effective browsing and searching of videos a challenging task.

When searching event-related topics, most users are interested in knowing the major events and their development. However, tracing the event evolution from a long list of videos is by no means easy. With the topic ''Michael Jackson death'' as an example, the search results are merely ranked according to text relevance, as shown in Figure 1. A thumbnail image together with a sparse set of tags is not adequate for giving a clear representation of the video content. Users need to painstakingly explore the search list to understand the each video's relationship to the search query. Manually organizing an event's evolution into a chronological structure like ''Jackson is dead'' → ''last rehearsal'' → ''memorial and funeral'' → ''Michael Jackson tribute'' can be time-consuming. However, automatically providing a concise structure showing the flow of events, together with the representative text keywords and visual shots, could be helpful to users by enabling efficient summarizing of major event highlights.

A news topic is composed of a connected series of events with a common focus or purpose happening in specific places during a given period. An event can be described concisely by a few discriminative and representative terms, for example, ''rehearsal,'' ''staples,'' ''center,'' and ''concert,'' which could be the representative features of the ''Michael Jackson's last rehearsal for London concert'' event. Likewise, a set of strongly correlated features could be used to reconstruct an event description. However, on the social Web, the number of user-supplied text (title and tags) is limited, and they are usually noisy, ambiguous, incomplete, and even misleading. A common difficulty of mining events from search results is that videos usually contain similar title and tags (for example, terms like ''Michael,'' ''Jackson,'' and ''death''). These terms are beneficial for retrieving relevant videos, but not for event mining. Important terms might be absent, leading to poor event coverage represented in the search results.

In contrast to conventional news videos where there are abundant text-from-speech transcripts and closed captions for exploring event relationships, the text feature in Web videos isn't discriminative enough for mining. While text is limited, Web videos carry rich visual content. Near-duplicate shots (keyframes) are often found and embedded as part of videos to signify the milestones of a topic at different timelines. One example is the event ''Michael Jackson's last rehearsal.'' Near-duplicate keyframes (NDKs) frequently appear in events such as ''Jackson's rehearsal on stage'' and ''Jackson's news press for London concert.'' These core shots are popularly manipulated and inserted into videos, either as a reminder or support of viewpoints. Therefore visual near-duplicates can be exploited for grouping videos of similar theme into events.

When a breaking news event occurs, similar Web videos of the same event from different viewpoints can always be found. These videos exhibit patterns in both textual and visual features. Associations and linkages between text

1070-986X/11/$26.00 © 2011 IEEE

Published by the IEEE Computer Society

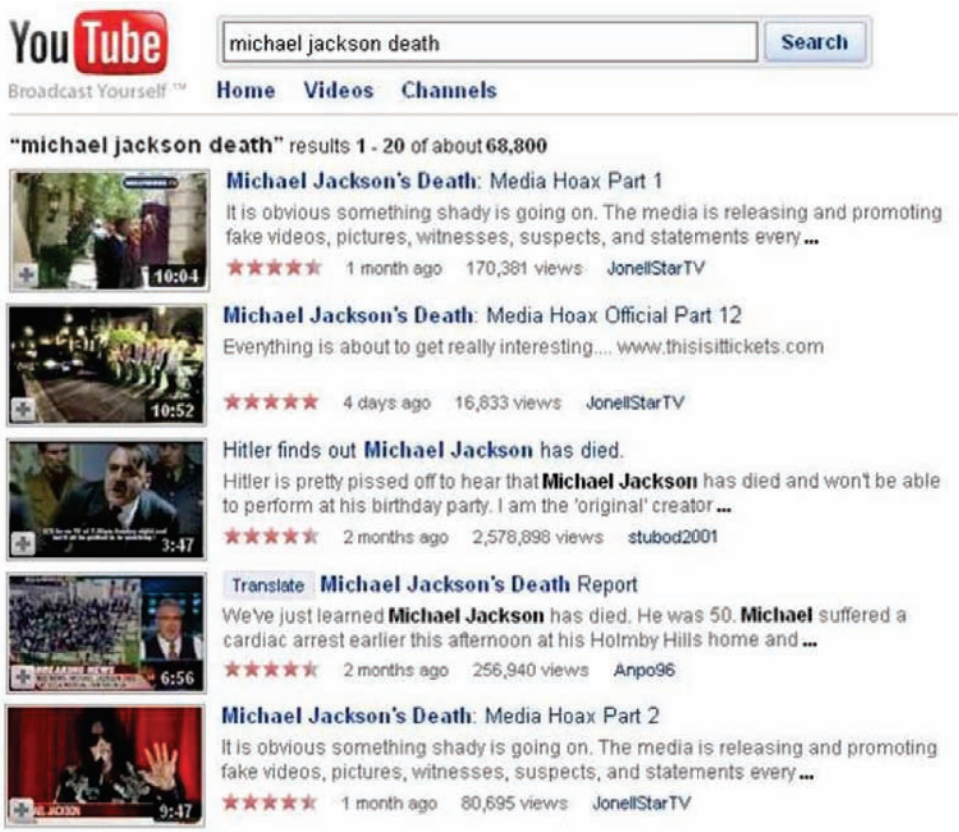


Figure 1. Search results for "Michael Jackson death" on YouTube.

words and visual shots can help discover important events from the search results. However, joint consideration of both features for mining video relationship is not always straightforward.

In this article, we explore event discovery and structure construction from Web video search results. A topic's breaking news period is first detected according to the number of videos uploaded in a certain time. Text patterns are mined from co-occurrence analysis of textual features. Event clusters are further formed by transitive closure, which gives accurate event descriptions from the text perspective. Visual NDKs provide direct linkage for videos. With NDKs, we perform burst detection in time and frequency domains for discovery of the major events. The evolution of NDKs is modeled as a bunch of feature trajectories in a 2D space of time and frequency for event clustering. Finally, events showing similar visual and textual properties are merged. Event structure is constructed by linking and aligning events along the timeline on the basis of event similarity.

Framework

Figure 2 (next page) shows our proposed framework, which is composed of event mining and event structure generation. The input to our framework is the Web videos returned from a search engine, while the output is the constructed event structure. Bursty period detection is first performed to detect the time region. Videos out of this region will be filtered out because most of them are irrelevant. From the perspective of semantics, the coexisting terms of text features associated with videos are detected by using a co-occurrence measure. These terms give clues to the presence of candidate events.

After shot boundary detection and keyframe extraction, each video is represented as a sequence of keyframes. Efficient NDK detection is adopted to detect the NDK among videos. Because of the unique property of visual NDKs for identifying similar events, the visual near-duplicate feature trajectory is constructed to track the appearance patterns of NDK. A clustering algorithm is then performed to group the visual NDK having similar distribution of feature trajectories. The corresponding videos are then

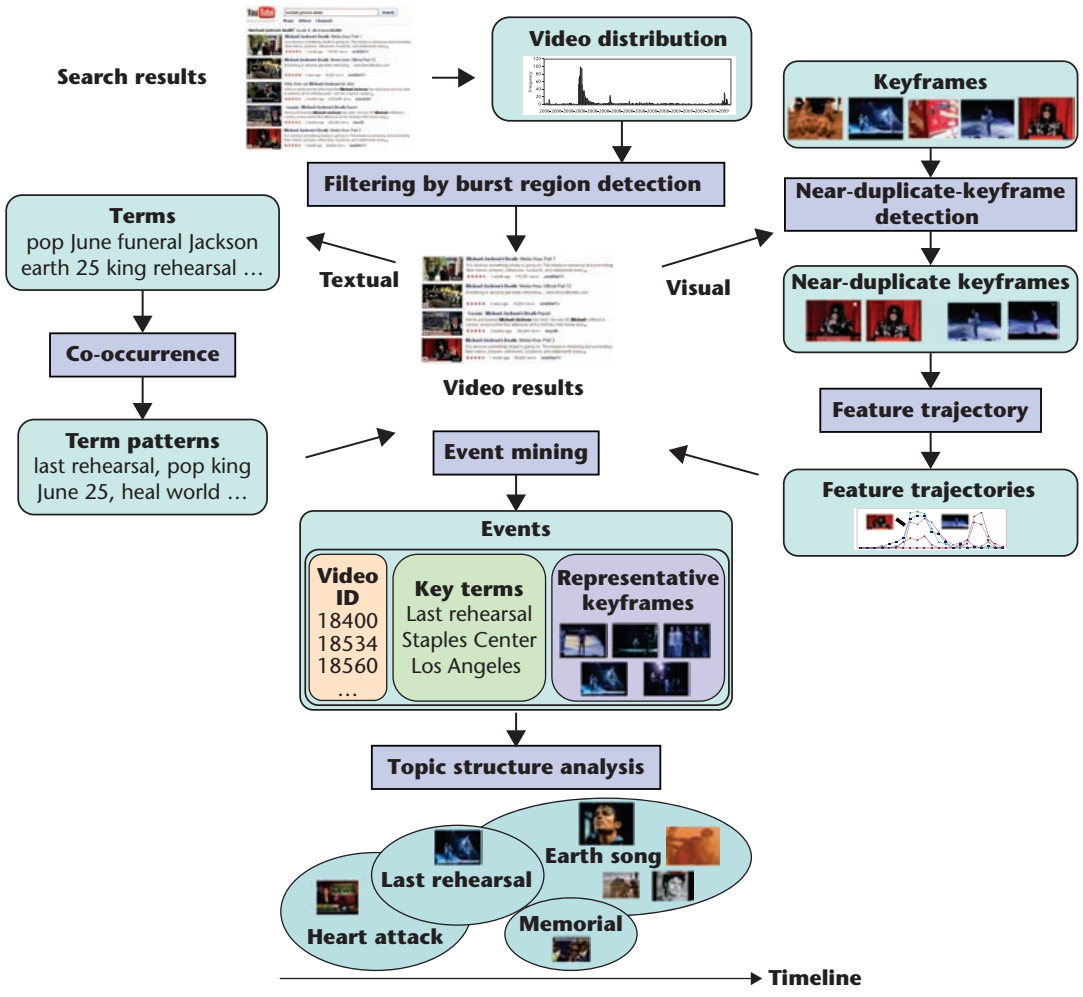


Figure 2. Framework.

clustered to form potential events. Accordingly, the representative key terms and keyframes are extracted to summarize the event, so that the main theme of each event can be easily identified. Due to several factors, such as diverse video content, noisy text features, and false or missed detection of NDKs, the textual and visual features usually demonstrate inconsistent characteristics. Therefore, the events discovered through textual and visual features separately could generate different sets of events.

Overlapped events are further merged. For each potential event, four properties—video set, start and end time, key terms, and representative NDKs—are extracted to represent the event. Finally, the detected events are mapped to the timeline according to the start and end time. The other three properties are exploited to measure the association among events and construct the event structure. A graphical view is used to visualize the structure so that users can easily grasp the flow of events and have a clear picture of how events evolve.

Web video event mining

Web users usually input few general words in search engines to search for new events happening around the world. Meanwhile, search engines return a long list of Web videos according to text relevance. Discovering meaningful videos, especially of new events, from the search results is not a trivial task. Observing that the bursty text features don't necessarily have exact overlap with the bursty features from visual duplicate shots, we study event discovery from two aspects: a relatively coarse granularity event mining from text semantics using text co-occurrence detection, and a relatively fine granularity event mining induced from the visual near-duplicate feature trajectory discovery.

Filtering by bursty period detection

Current Web video search engines return a list of videos determined according to text relevance. Therefore, there are many irrelevant items mixed in the search results. For example,

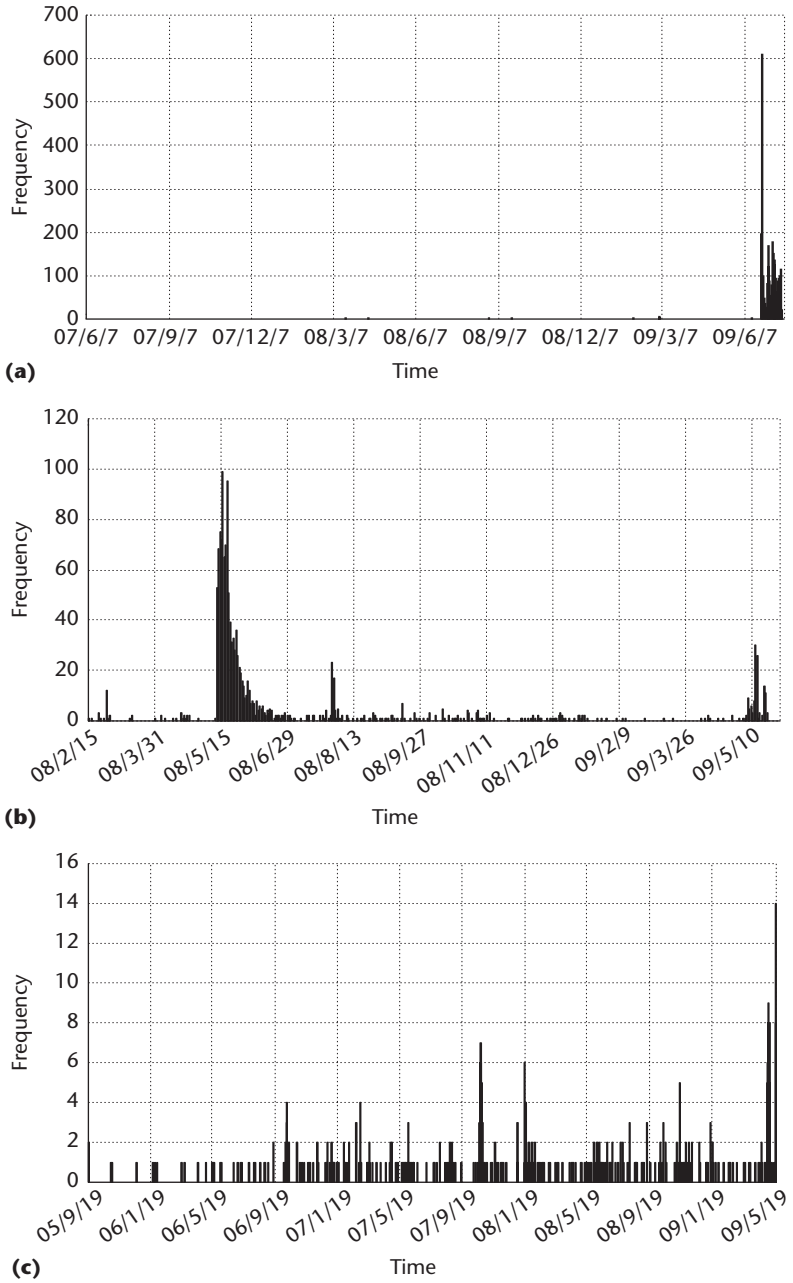


Figure 3. Video distributions according to upload time for queries: (a) ''Michael Jackson death.'' (b) ''Sichuan earthquake,'' and (c) ''California wildfires.''

when searching ''911 terrorist attacks,'' although the users are expected to watch videos on the terrorist attacks that happened in US on September 11, 2001, a large number of videos about 911 emergency calls are returned simply because the title or tags of videos include the keyword ''911.'' In fact, for videos on the 9/11 terrorist attacks, video uploads during that time period occurred intensively.

To discover the desired events from the search results, the first step is to locate the bursty period according to the upload time of videos. The number of videos associated with a particular upload time is a good indication of events, and potentially the periods that events happened. Videos whose upload time is not in that period can be treated as noise and not directly related to the bursty events. Figures 3a–c show the distributions of the

Table 1. Data set Information.

ID	Topic	Search queries	Main time period	Video number (total is 19,972)
1	Economic collapse	Economic collapse/financial crisis/economic crisis	09/14/2008 ~ 05/11/2009	1,025
2	US presidential election	US president election/president election 2008	08/14/2008 ~ 11/24/2008	737
3	Beijing Olympics	Beijing Olympics/Beijing 2008/Beijing Olympic Games 2008/Olympics 2008	03/11/2008 ~ 02/09/2009	1,098
4	Mumbai terror attack	Mumbai terror attack	11/26/2008 ~ 12/23/2008	423
5	Russia Georgia war	Russia Georgia war/Russia Georgia conflict	08/08/2008 ~ 09/28/2008	749
6	Somali pirates	Somali pirates	09/29/2008 ~ 05/15/2009	410
7	Virginia tech massacre	Virginia tech massacre/Virginia tech shooting	04/16/2007 ~ 05/24/2007	683
8	Israel attacks Gaza	Israel attacks Gaza/Gaze conflict	12/27/2008 ~ 02/16/2009	802
9	Beijing Olympic torch relay	Beijing Olympic torch relay/torch relay 2008	03/22/2008 ~ 05/30/2008	652
10	Melamine	Melamine/milk scandal/tainted milk	09/13/2008 ~ 12/29/2008	783
11	Sichuan earthquake	Sichuan earthquake/China earthquake 2008/ Wenchuan earthquake	05/12/2008 ~ 08/02/2008	1,458
12	California wildfires	California wildfires	10/08/2006 ~ 05/22/2009	426
13	London terrorist attack	London terrorist attack/London bombing	05/16/2007 ~ 05/23/2009	784
14	Oil price	Oil price/gas price	03/01/2008 ~ 11/23/2009	759
15	Myanmar cyclone	Myanmar cyclone/Burma cyclone	05/04/2008 ~ 08/07/2008	613
16	Kosovo independence	Kosovo independence	02/13/2008 ~ 03/17/2008	524
17	Russian presidential election	Russian president election/Russian president/ Russian president Medvedev	12/01/2007 ~ 05/24/2009	1,335
18	Iran nuclear program	Iran nuclear program/Iran nuclear weapons	10/01/2006 ~ 05/25/2009	1,056
19	Israeli Palestine peace	Israeli Palestine peace	07/19/2006 ~ 05/27/2009	586
20	Korea nuclear	Korea nuclear/North Korea test	02/01/2009 ~ 05/29/2009	1,060
21	Swine flu	Swine flu/H1N1	04/24/2009 ~ 05/30/2009	1,153
22	Michael Jackson dead	Michael Jackson dead/Michael Jackson dies/ Michael Jackson death	06/25/2009 ~ 07/25/2009	2,850

number of videos according to upload time, from which we can see that there are dominant peaks except for "California wildfires." Interestingly, there are three apparent peaks for the query "Sichuan earthquake," which correspond to the events Sichuan earthquake (12 May 2008), Los Angeles earthquake (29 July 2008), and one year anniversary of Sichuan earthquake (12 May 2009).

In this article, our focus is to discover the events in the dominant period. That is, in this example, we are more interested in finding events that might be associated with "rescue," "mourn," "donation," and so on. To this end, the algorithm based on Gaussian model with Expectation Maximization (EM) can be exploited to model and estimate the bursty peaks. For simplicity, in this article, we use the following heuristic strategy to locate the bursty region:

$$R_j = [t_j - W, t_j + W] \,||\, |V_j| \geq \alpha \left(\sum_{k=1}^{n} |V_k| \right) / n$$

where $|V_j|$ is the number of videos uploaded at time t_j, n is the number of days. The parameters α and W are used to control the peak level and window size respectively. In the experiments, we set α as 1.5 and W as 3. The bursty peaks and their neighbors are first detected and consecutive peaks are merged to form the bursty period. The main time periods of events are listed in Table 1.

Due to limited number of videos or the focus of the event itself, certain queries have no dominant bursty periods, for example, the "California wildfires" in Figure 3c. As a result of bursty detection, no peaks will be detected in the first round detection, indicating that the search results are quite diverse. In this case, a smaller value of α is set to avoid filtering out most of the videos.

Term co-occurrence

From titles and tags, there are numerous frequently accompanied terms that convey useful

information. The Jaccard coefficient is adopted to measure the co-occurrence of terms:

$$d(w_i, w_j) = \frac{|M_i \cap M_j|}{|M_i \cup M_j|}$$

where M_i is the set of videos containing word w_i. Given two words w_i and w_j, the overlapping video set containing both words is $M_i \cap M_j$. Intuitively, a higher score denotes that these two words are more likely correlated. If the score is above a certain threshold, they can be regarded as co-occurrence, denoted as $w_i \rightarrow w_j$. Co-occurrence is a symmetric measure, it also implies that $w_j \rightarrow w_i$.

The transitive closure is used to group co-occurrence terms together. If $w_i \rightarrow w_j$, and $w_j \rightarrow w_k$, then we assume that there exists the association $w_i \rightarrow w_k$. After transitive closure, w_i, w_j and w_k are clustered together. The formed clusters are potential events. Table 2 shows some examples for the queries ''Beijing Olympics'' and ''Michael Jackson death,'' from which we can find many interesting patterns. These patterns potentially correspond to certain events or concepts hidden in the videos. For example, the word pattern ''opening ceremony'' corresponds to events of ''Beijing Olympic Opening Ceremony in Aug 8, 2008,'' while ''Usain Bolt 200 m 100 m Jamaica'' is associated with the event ''Jamaican Usain Bolt broke the 100 m and 200 m world records.'' Another interesting cross-lingual event is ''Huan Liu Sarah Brightman theme song,'' which refers to the theme song of the Beijing Olympics 2008, that is, *You and Me*. The singers' names are also grouped together. Table 3 (next page) lists the top 10 events for ''Beijing Olympics 2008'' detected by text co-occurrence.

Table 2. Sample text patterns mined through co-occurrence.

ID	Text patterns
3	Opening ceremony
	Theme song
	Gold medal
	Michael Phelps swimming
	Torch relay
	National stadium bird nest
	Water cube
	Usain Bolt 200m 100m, Jamaica
	Soccer football
	London 2012
	Fireworks rehearsal
	Finals event
	Huan Liu, Sarah Brightman
22	June 25
	Heart attack cardiac arrest hospital
	Last rehearsal
	Los Angeles
	Black, white
	Staples Center
	Heal, world
	Neverland ranch
	Moon walk
	Pepsi commercial fire hair accident
	911 call
	Prescription drug

Visual feature trajectory

A news video event is usually accompanied with a burst of representative visual near-duplicate shots or keyframes. Some representative visual features suddenly appear frequently when the event emerges, and their frequencies drop when the event fades away. For example, the keyframes of ''Jackson is rehearsal dancing'' and ''Jackson news press'' frequently appear in videos on ''Jackson's last rehearsal video'' and ''Jackson's London concert'' events. An event can also be represented in the form of representative visual features. Representative visual near-duplicate features from the same event share similar distributions over time and are highly correlated. An important event has a set of largely reported representative visual features, whereas a minor event may have no such visual features. In this article, we will explore the visual feature trajectory to discover the events.

After NDK detection and transitive closure, a set of NDK clusters are formed in which each cluster represents one visual scene. For example, the keyframes of the ''bird nest'' stadium are grouped together as a cluster. We track the visual NDKs along the timeline and form the visual feature trajectory. The visual near-duplicate feature trajectory is modeled as the feature distribution along the timeline in a 2D space with one dimension as time and the other as feature weight. The trajectory Y_m of visual feature v_m is defined as follows:

$$Y_m = [y_m(t_i), y_m(t_{i+1}), \ldots, y_m(t_j)]$$

Table 3. Top 10 events detected by text co-occurrence for "Beijing Olympics 2008."

ID	Number of Videos	Key terms	Representative keyframes
1	96	Opening ceremony	
2	83	Gold medal	
3	81	Michael Phelps swimming	
4	76	Theme song	
5	58	Liukin Nastia, Shawn Johnson	
6	52	Field track	
7	50	National stadium, bird nest	
8	42	Finals event	
9	40	Usain Bolt, 200m, 100m, Jamaica	
10	38	Torch relay	

where $y_m(t_i)$ is a measure of visual feature v_m at time t_i. In our case, the visual feature is NDKs, defined according to the *df-idf* representation:

$$y_m(t_i) = \frac{df_m(t_i)}{N(t_i)} \times \log \frac{N}{df_m}$$

where $df_m(t_i)$ is the number of videos containing visual feature v_m at day t_i, df_m is the total number of videos containing visual feature v_m over all times, $N(t_i)$ is the number of videos for day t_i, and N is the total number of videos over all time. In this article, the time unit is set as one day.

Keyframes with similar visual feature trajectories, measured by Euclidean distance between visual feature trajectories, are clustered together. The corresponding videos having the keyframes are then grouped to form events. Figure 4 demonstrates the visual near-duplicate feature trajectories for four NDKs. In this figure, we can see that the first two visual feature trajectories of "Jackson last rehearsal" have similar peaks and trajectory patterns, but different

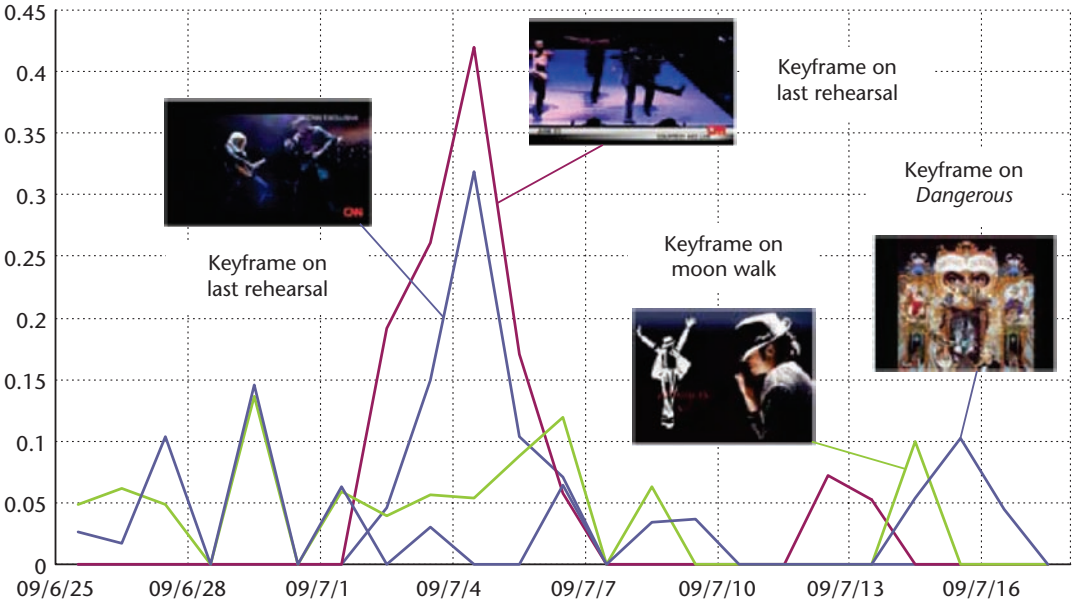


Figure 4. Visual feature trajectory.

from the patterns of the keyframes of ''moon walk'' and ''Dangerous.'' So the first two keyframes should be grouped together, which also indicates that these two keyframes should be associated with a certain event. Table 4 (next page) lists the top 10 events for ''Beijing Olympics 2008'' detected by visual feature trajectory.

Event structure construction

With event mining, two sets of events are discovered separately according to text co-occurrence mining and visual near-duplicate feature-trajectory discovery. These events might overlap and complement each other. The ultimate goal is to build an event structure so that the event relationship can be clearly presented in a graphical view.

Event association measure

Each event is represented by four properties: video set, time period, key terms, and representative keyframes. The association $Sim(e_i, e_j)$ between two events e_i, e_j is a linear combination of three components: video set overlap S_v, term similarity S_t, and keyframe similarity S_k.

$$Sim(e_i, e_j) = S_v(e_i, e_j) + S_t(e_i, e_j) + S_k(e_i, e_j)$$

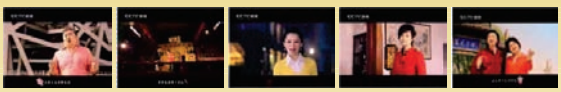
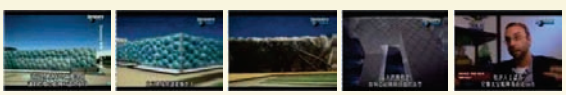
To measure the overlap for individual components, the vector-space model, language models, or other measures can be adopted to evaluate the similarity, but the similarity measure and fusion scheme are not the focus of this article. For simplicity, in this article, we use Jaccard coefficient to evaluate the similarity. After fusion, a similarity score can be obtained to determine the event association, which will be exploited for event structure construction.

Structure construction

The following algorithm describes the steps to automatically construct the event structure. The input of the algorithm is the events discovered by text co-occurrence mining and visual near-duplicate feature trajectory discovery. The output of the algorithm is the event structure:

- Extract four components for each event;
- Map each events to the timeline, and associate corresponding key terms and representative keyframes;
- Set the size of circle to visualize the number of videos in each event;
- Merge the events if their similarity is high;
- According to the order of timeline, compare each event with previous events on the basis of an event-association measure. The closest association determines the arrow between two events.
- Construct linkage according to event similarity. The color of the arrows between

Table 4. Top 10 events detected by visual feature trajectory for "Beijing Olympics 2008."

ID	Number of videos	Key terms	Representative keyframes
1	90	Sarah Brightman official theme song Huan Liu mv [music video]	
2	73	Usain, 100m Jamaica paralympics day bolt 69 seconds field track Waddell	
3	46	58kg men taekwondo women competition show team interview mex [Mexico], Andy night podium Spain training table	
4	42	Beijing welcomes you song mv Chan Jackie Spanish subtitle espa[nol] Jolin	
5	35	Prelims event finals gymnastics vault floor part bars qualification men XXIX women	
6	29	Sarah Brightman theme song Huan Liu mv opening ceremony Liuhuan	
7	25	Mascots piano Fuwa magbaza official WWW cassina 2 horizontal fanfare monster	
8	20	Stadiums game makeover marionette listening	
9	18	Rhythmic clubs ribbon hoop ropes Bessonova gymnastics Anna Kanaeva final hoops Italy rg [rhythmic gymnastique] gymnastic	
10	12	Song theme live China countdown Olimpicos Juegos world dream	

events denotes the degree of event association: red > blue > black.

With this kind of event structure, Web users could have a better understanding of the major event highlights together with a concise structure showing the evolution of events associated with the representative text keywords and visual shots. The event structure conveys useful properties, which enable effective large-scale video visualization and browsing. The timeline indicates the event sequence, the size of the circle denotes the degree of the event's relative significance, while line width refers to the tightness among events. In addition, the association of key terms and representative shots gives direct semantic and visual summarization for each event.

Figure 5 gives an example of the generated event structures for "Beijing Olympics" and "Michael Jackson death," presenting a clear roadmap of the events in the search results. From Figure 5a, we can see that the major events are "opening ceremony," "theme song,

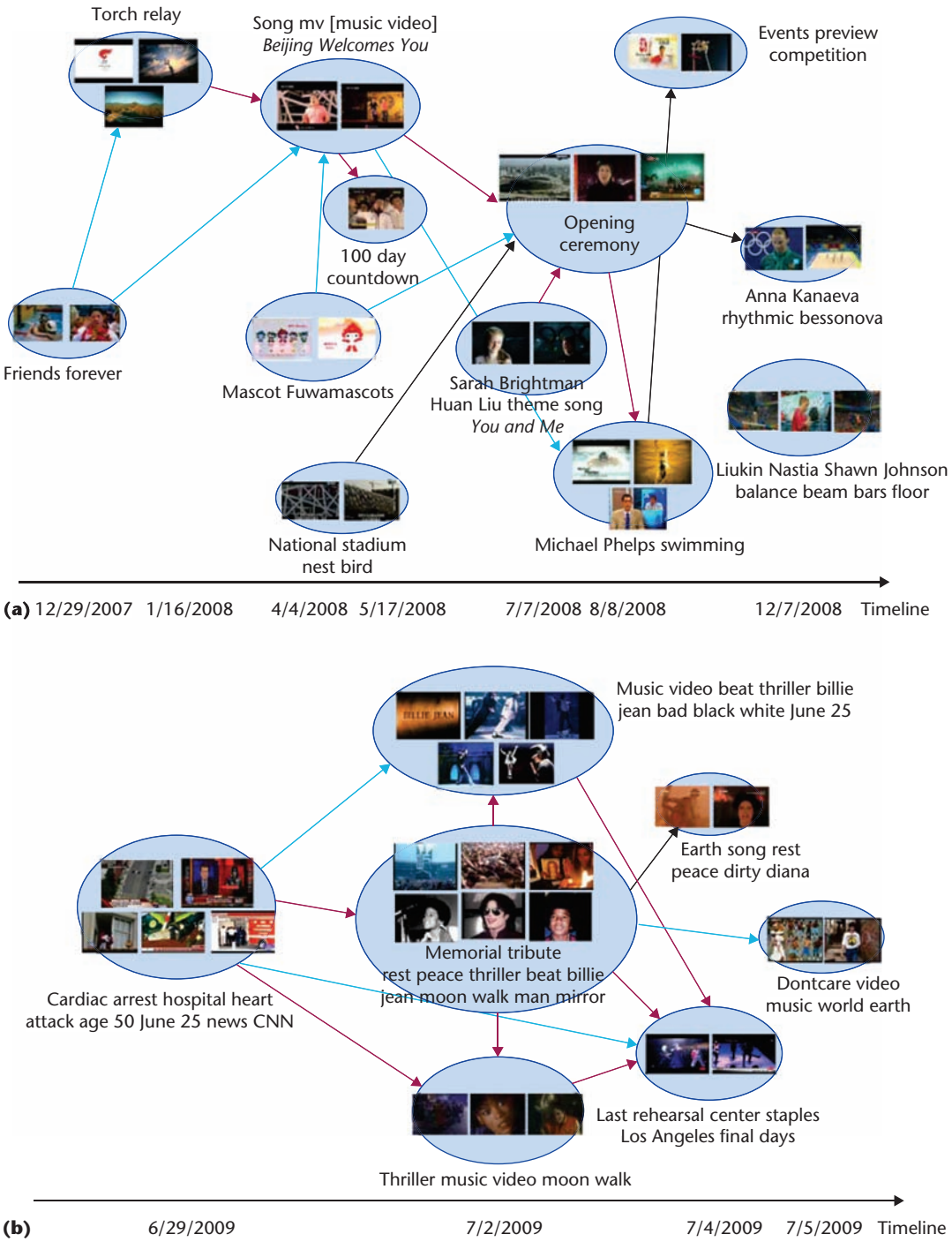


Figure 5. Event structures for queries (a) "Beijing Olympics 2008" and (b) "Michael Jackson death." The color of the arrows between events denotes the degree of event association: red > blue > black.

You and Me, sung by Sarah Brightman and Huan Liu," "MV" [music video], "Beijing welcomes you," and so on. Through the timeline, the sequence of events is clearly aligned. While in Figure 5b, the major events include "Michael Jackson is dead due to heart attack," "memorial," "last rehearsal," and some classic music videos. The mined structure is meaningful.

Experiments

After careful evaluation of the top 10 news events during 2006 and 2009 recommended by CNN, *Time*, and Xinhua, we selected 22 significant events as queries for experiments. We issued these queries to YouTube and crawled the related videos. Usually YouTube only returns at most 1,000 results for each query. To guarantee the coverage of events, multiple

Related Work

Our work is related to the research of topic detection and tracking (TDT),[1] which focuses on detecting new events and tracking known events in text news streams. TDT has been extensively studied, and the previous works mainly focus on text streams.[2–8]

Recently, TDT research has been extended to multimedia. Topics are tracked with visual duplicates and semantic concepts in Hsu and Chang.[9] In Duygulu, Pan, and Forsyth, repeated sequences of news shots are detected and tracked.[10] Similarly, in Zhai and Shah, news stories from different TV channels are linked by textual correlation and keyframe matching.[11] News stories are clustered into topics with the assistance of near-duplicate keyframe (NDK) constraints in Wu, Ngo, and Hauptmann.[12] However, these works don't address the issue of topic structure generation for efficient large-scale browsing. The work of Ide, Mo, and Katayama segments news videos into stories and then constructs dependencies among stories as a graph structure.[13] The thread structure is formed with a chronologically ordered directed graph. In Neo et al., with the help of external resources, such as news articles and blogs, news videos are clustered into a hierarchical structure.[14]

In our previous work, we thread and autodocument news stories according to topic themes.[15] Story clustering is first deployed by exploiting the duality between stories and textual-visual concepts through a coclustering algorithm. The dependency among stories of a topic is tracked by exploring the textual-visual novelty and redundancy detection. A topic structure that chains the dependencies of stories facilitates the fast navigation of a news topic.

The aforementioned approaches target news videos, and yet, few works contribute to Web video event mining and tracking. For news videos, closed captions or text transcripts are usually available. However, this is not the case for Web videos where the speech is often low quality. In Liu et al., Web video topic detection and tracking is modeled by a bipartite graph constructed by Web videos and text keywords.[16] Topic discovery is achieved through coarse topic filtering and fine topic reranking. The topic is tracked by propagating relevant scores and keywords of videos of interests to other relevant videos through the bipartite graph.

Recently, there has been significant interest in modeling an event in text streams as a burst of activities by incorporating temporal information. In He, Chang, and Lim, news stories and feature trends are analyzed to identify important and periodic words, from the perspective of a time-series word signal.[4] Fung et al. propose an approach to construct an event hierarchy in a text corpus.[3] In this approach, bursty features are first identified according to the timestamps and text content. Documents highly related to the bursty features are extracted and organized in a hierarchical structure. In Wang et al., a general probabilistic algorithm is proposed to discover correlated bursty topic patterns and bursty periods in coordinated English and Chinese text streams.[8] All these works focus on the text streams in which the text articles are more informative and have less noise. None of these approaches could be directly applied to Web videos.

While generating trajectories from text features has been a commonly adopted approach, there is no study yet about

text queries with similar meaning were used for each topic to retrieve the Web videos from YouTube. For example, we issued ''Sichuan earthquake,'' ''China earthquake 2008,'' and ''Wenchuan earthquake'' for topic 11. We collected all retrieved videos after removing repeated videos from multiple queries as our data set. All videos were collected in May 2009 except for topic 22. The final data set consists of 19,972 Web videos. The query information and the number of videos are listed in Table 1.

To analyze the performance of our event-mining system, three nonexpert assessors were asked to watch Web videos ordered according to the upload time, and group videos with similar theme into events. To ensure the fairness, the events having less than six Web videos were regarded as noise and pruned out.

Shot boundaries were detected and each shot was represented by a keyframe. In total there were 803,346 keyframes in the set. Local keypoints were detected from the keyframes by Harris-Laplace and described by scale-invariant feature transform.[1] Clustering was adopted to quantize the keypoints into a visual dictionary (20,000 clusters). Each keyframe was encoded as a bag of words. We adopted the approach in Zhao, Wu, and Ngo[2] to detect the NDKs. The detected NDKs were further grouped to form clusters by transitive closure, which represent the same visual scene. There are 134,797 NDKs, which form 43,549 NDK groups. Due to noisy user-supplied tag information, special characters (for example, ?, !, :, #, >, and |) were first removed. After data preprocessing (such as word stemming, special character removal,

the visual trajectories generated from NDKs. Compared to noisy text terms in the social Web, which might cross a long temporal range, NDKs in events appear in a relatively small range and are less noisy. This motivates the study of visual feature trajectory in this article.

References

1. J. Allan, ed., *Topic Detection and Tracking: Event-based Information Organization,* Kluwer Academic Publishers, 2002.
2. C.-C. Chen and M.-C. Chen, ''TSCAN: A Novel Method for Topic Summarization and Content Anatomy,'' *Proc. Int'l ACM Sigir Conf. Research and Development in Information Retrieval,* ACM Press, 2008, pp. 579-586.
3. G. P.-C. Fung et al., ''Time-Dependent Event Hierarchy Construction,'' *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* ACM Press, 2007, pp. 300-309.
4. Q. He, K. Chang, and E.-P. Lim, ''Analyzing Feature Trajectories for Event Detection,'' *Proc. Int'l ACM Sigir Conf. Research and Development in Information Retrieval,* ACM Press, 2007, pp. 207-214.
5. Q. Mei and C. Zhai, ''Discovering Evolutionary Theme Patterns from Text—An Exploration of Temporal Text Mining,'' *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* ACM Press, 2005, pp. 198-207.
6. R. Nallapati et al., ''Event Threading within News Topics,'' *Proc. ACM Int'l Conf. Information and Knowledge Management,* ACM Press, 2004, pp. 446-453.
7. C. Wang et al., ''Automatic Online News Issue Construction in Web Environment,'' *Proc. Int'l World Wide Web Conf., ACM Press,* 2008, pp. 457-466.
8. X.-H. Wang et al., ''Mining Correlated Bursty Topic Patterns from Coordinated Text Streams,'' *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* ACM Press, 2007, pp. 784-793.
9. W.H. Hsu and S-F. Chang, ''Topic Tracking across Broadcast News Videos with Visual Duplicates and Semantic Concepts,'' *Proc. Int'l Conf. Image Processing,* IEEE Press, 2006, pp. 141-144.
10. P. Duygulu, J-Y. Pan, and D.A. Forsyth, ''Towards Auto-Documentary: Tracking the Evolution of News Stories,'' *Proc. ACM Multimedia Conf.,* ACM Press, 2004, pp. 820-827.
11. Y. Zhai and M. Shah, ''Tracking News Stories across Different Sources,'' *Proc. ACM Multimedia Conf.,* ACM Press, 2005, pp. 2-10.
12. X. Wu, C.-W. Ngo, and A.G. Hauptmann, ''Multimodal News Story Clustering with Pairwise Visual Near-Duplicate Constraint,'' *IEEE Trans. Multimedia,* vol. 10, no. 2, 2008, pp. 188-199.
13. I. Ide, H. Mo, and N. Katayama, ''Threading News Video Topics,'' *Proc. ACM Multimedia Information Retrieval,* ACM Press, 2003, pp. 239-246.
14. S.-Y. Neo et al., ''The Use of Topic Evaluation to Help Users Browse and Find Answers in News Video Corpus,'' *Proc. ACM Multimedia Conf.,* ACM Press, 2007, pp. 198-207.
15. X. Wu, C-W. Ngo, and Q. Li, ''Threading and Autodocumenting News Videos,'' *IEEE Signal Processing Magazine,* vol. 23, no. 2, 2006, pp. 59-68.
16. L. Liu et al., ''Web Video Topic Discovery and Tracking via Bipartite Graph Reinforcement Model,'' *Proc. Int'l World Wide Web Conf.,* ACM Press, 2008, pp. 1009-1018.

Chinese word segmentation, and so on), there were 35,136 unique text words.

We used precision and recall to evaluate the performance of event mining, which is defined as:

$$Precision_i = |P_i^+|/|C_i| Recall_i = |P_i^+|/|P_i|$$

where P_i^+ is the number of correctly grouped positive videos for cluster C_i, and P_i is the number of positive samples in ground truth.

Performance evaluation

The performance evaluation is listed in Table 5 (next page). A simplified version of the proposed method in He, Chang, and Lim[3] is treated as the baseline. Word feature trajectories are first extracted. Each feature is defined as a normalized *df-idf* score. Highly correlated word features are grouped to construct events by mapping word sets to video sets. Overall, we can see that it's a challenging task to detect events from the set of diverse and noisy videos returned from search engines. Event mining using visual near-duplicates has inconsistent performance compared to using text features. The event discovery through text co-occurrence forms more general events and the visually dissimilar videos can be grouped together by semantics. In contrast, by NDKs, videos can be densely clustered into events. However, the videos without overlapped NDKs can't be included into any event.

The top 10 events detected by text co-occurrence mining and visual feature trajectory discovery are demonstrated in Tables 3 and 4, respectively. In Table 3, we can see that the key terms are semantically meaningful. The

Table 5. Performance evaluation.

		Baseline		Text		Visual		Text + visual	
ID	Topic	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
1	Economic collapse	0.24	0.17	0.20	0.18	0.27	0.21	0.59	0.16
2	US presidential election	0.26	0.39	0.14	0.47	0.26	0.41	0.57	0.35
3	Beijing Olympics	0.54	0.11	0.51	0.16	0.50	0.06	0.64	0.18
4	Mumbai terror attack	0.31	0.14	0.24	0.24	0.37	0.12	0.49	0.19
5	Russia Georgia war	0.58	0.11	0.43	0.16	0.52	0.07	0.72	0.15
6	Somali pirates	0.49	0.21	0.33	0.18	0.38	0.19	0.48	0.25
7	Virginia tech massacre	0.76	0.05	0.47	0.06	0.51	0.14	0.73	0.33
8	Israel attacks Gaza	0.45	0.12	0.26	0.16	0.52	0.07	0.54	0.16
9	Beijing Olympic torch relay	0.52	0.41	0.44	0.59	0.43	0.18	0.52	0.20
10	Melamine	0.18	0.21	0.11	0.29	0.19	0.33	0.42	0.28
11	Sichuan earthquake	0.52	0.05	0.34	0.09	0.55	0.18	0.76	0.47
12	California wildfires	0.46	0.12	0.42	0.15	0.49	0.05	0.68	0.18
13	London terrorist attack	0.04	0.19	0.09	0.44	0.18	0.36	0.49	0.25
14	Oil price	0.22	0.10	0.14	0.08	0.31	0.13	0.58	0.13
15	Myanmar cyclone	0.39	0.05	0.28	0.08	0.37	0.08	0.68	0.34
16	Kosovo independence	0.66	0.07	0.58	0.09	0.54	0.01	0.78	0.09
17	Russian president election	0.27	0.14	0.15	0.27	0.41	0.13	0.61	0.14
18	Iran nuclear program	0.60	0.07	0.61	0.16	0.67	0.04	0.83	0.10
19	Israeli Palestine peace	0.35	0.17	0.26	0.22	0.53	0.12	0.51	0.16
20	Korea nuclear	0.34	0.16	0.12	0.19	0.26	0.23	0.46	0.24
21	Swine flu	0.15	0.26	0.14	0.35	0.17	0.20	0.25	0.22
22	Michael Jackson death	0.64	0.08	0.56	0.11	0.70	0.07	0.83	0.11
–	**Average**	**0.41**	**0.15**	**0.31**	**0.21**	**0.41**	**0.15**	**0.59**	**0.21**

representative keyframes are extracted from the videos having the key terms. Generally, the videos grouped according to text co-occurrence are diverse. In Table 4, the videos having visual relationship are grouped to form events. These videos are densely coupled. Key terms are then induced from the title and tags of the videos in the events. Compared to Table 3, the extracted key terms in Table 4 are diverse. The social tags and visual content convey inconsistent information. The results combining textual and visual features are listed in Table 5.

Generally, the integration improves the performance. However, this is not always true. A typical failure case is when the noisy videos form clusters by both text and visual features. In this case, combining the clusters will bring more noisy videos to the event, ending up with a larger cluster with messy results. It's worth noting that setting different numbers of clusters will affect the performance. As expected, choosing the appropriate number of clusters is challenging, and it's a subjective issue on how to define the granularity of clusters and events. In addition, an event is sometimes too diverse to be clustered into one cluster. When this happens, videos of the same event can be split into multiple clusters. However, as we adopt timeline visualization, these clusters will still stay close to each other in the final structure.

Conclusion

We performed the event structure mining on the basis of text co-occurrence and visual feature trajectory. We plan to revisit the event discovery by constructing a bipartite graph and integrating visual near-duplicates with text features to improve the clustering performance. We also plan to further study the effect of feature trajectory and co-occurrence for event mining, and ways to fuse text and visual near-duplicates more effectively. **MM**

Acknowledgments

The work described in this article was supported by the National Natural Science Foundation of China (NSFC grant no. 61071184,

60972111, 6103600); the Fundamental Research Funds for the Central Universities (project no. SWJTU09CX032, SWJTU09ZT14); the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), and a grant from the Research Grants Council of Hong Kong Special Administrative Regions, China, (CityU 119610).

References

1. D. Lowe, ''Distinctive Image Features from Scale-Invariant Key Points,'' *Int'l J. Computer Vision,* vol. 60, no. 2, 2004, pp. 91-110.
2. W.-L. Zhao, X. Wu, and C.-W. Ngo, ''On the Annotation of Web Videos by Efficient Near-Duplicate Search,'' *IEEE Trans. Multimedia,* vol. 12, no. 5, 2010, pp. 448-461.
3. Q. He, K. Chang, and E.-P. Lim, ''Analyzing Feature Trajectories for Event Detection,'' *Proc. Int'l ACM Sigir Conf. Research and Development in Information Retrieval,* ACM Press, 2007, pp. 207-214.

Xiao Wu is an associate professor at Southwest Jiaotong University, China. His research interests include multimedia information retrieval, video computing, and data mining. Wu has a PhD in computer science from City University of Hong Kong. Contact him at wuxiaohk@home.swjtu.edu.cn.

Yi-Jie Lu is currently pursuing an MS at Southwest Jiaotong University, China. His research interests include multimedia information retrieval, and video processing. Lu has a BS in computer science from Southwest Jiaotong University. Contact him at iiedii@gmail.com.

Qiang Peng is a professor at Southwest Jiaotong University, China. His research interests include image processing and video coding. Peng has a PhD in computer science from Southwest Jiaotong University. Contact him at qpeng@home.swjtu.edu.cn.

Chong-Wah Ngo is an associate professor at City University of Hong Kong. His research interests include video computing and multimedia information retrieval. Ngo has a PhD in computer science from Hong Kong University of Science and Technology. Contact him at cwngo@cs.cityu.edu.hk.