Topical Summarization of Web Videos by Visual-Text Time-Dependent Alignment

Song Tan, Hung-Khoon Tan and Chong-Wah Ngo
Deptartment of Computer Science
City University of Hong Kong
Hong Kong, China
songtan@student.cityu.edu.hk, {hktan, cwngo}@cs.cityu.edu.hk

ABSTRACT

Search engines are used to return a long list of hundreds or even thousands of videos in response to a query topic. Efficient navigation of videos becomes difficult and users often need to painstakingly explore the search list for a gist of the search result. This paper addresses the challenge of topical summarization by providing a timeline-based visualization of videos through matching of heterogeneous sources. To overcome the so called sparse-text problem of web videos, auxiliary information from Google context is exploited. *Google Trends* is used to predict the milestone events of a topic. Meanwhile, the typical scenes of web videos are extracted by visual near-duplicate threading. Visual-text alignment is then conducted to align scenes from videos and articles from *Google News*. The outcome is a set of scene-news pairs, each representing an event mapped to the milestone timeline of a topic. The timeline-based visualization provides a glimpse of major events about a topic. We conduct both the quantitative and subjective studies to evaluate the practicality of the application.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [Retrieval models]

General Terms

Theory, Algorithms, Performance, Experimentation

Keywords

Timeline-based Summarization, Google Context

1. INTRODUCTION

With the advent of Web 2.0, there is a wealth of information that pervades the Internet. One main concern is how to make sense of these data and organize them in a manner that improves users' understanding of a particular topic. For example, a typical retrieval system only produces a search list without any discernable structure which leave users with only fragmented and incomplete understanding of the topic. Topical summarization aims to detect only the important information and organize them in an easy-to-understand manner. In this paper, we explore the timeline summarization of web videos using multi-modalities processing and external sources. Given a set of videos, topical summarization detects the hot events for the given topic, identifies the typical videos associated with each event and finally finds the dependencies between them to construct a structure which facilitates efficient browsing.

Detection of hot events has been intensively investigated for the textual domain, e.g., by analyzing keyword trajectories [6] across timeline or extracting hot key-terms through sentence modeling [4]. More informative topic unit includes story clusters [8] where a news transcript is cut based on the keyword distributions in a moving windows. The work in [5] proposes a clustering-free algorithm which detects a hot event as a set of bursty features densely focused within a particular time window. In the news video domain, the story units are grouped with keywords and near-duplicates by performing co-clustering using a bi-partite graph while [12] further incorporates facial cues. Then, to thread the detected events, the popular approach is to represent the events as a directed tree [11, 8, 9] where the events are linked in a chronological fashion. For summary presentation, two popular approaches are static storyboard and video skimming [3, 7]. Static storyboard presents a topic as a mosaic of keyframes supplemented with a set of keyword descriptions. Video skimming generates a video synopsis which combines the critical scenes of the topic-of-interest.

However, most of these methods assume the availability of news transcripts as the basis for hot event detection. The over-reliance on transcript renders it unsuitable for web video domain where such transcript is not readily available. Furthermore, the tags or descriptions surrounding each video are always sparse and not discriminative, which makes story segmentation and finding the dependencies between videos challenging. In short, the lack of a strong cue makes summarizing web videos more difficult than news videos. In addition, web videos are diverse where the content and quality vary significantly for different web videos. For example, the videos can be musical videos, video blogs by users, self-made slide shows and furthermore some videos are speech-free or audio-less where automatic speech recognition is not possible. To address these problems, [7] uses keyshots as the basic topic unit where a keyshot is a set of

MM'10, October 25–29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-60558-933-6/10/10 ...$10.00.

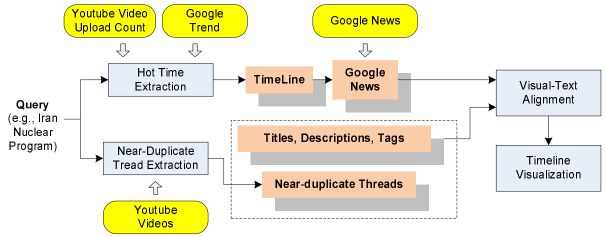


Figure 1: Proposed Google-context video summarization system.

near-duplicate shots. However, keyshots are short in general and as a result when concatenated to form a video skim, the summary video is often fragmented and sketchy, leading to only fragmented understanding of the topic. Furthermore, the chronological order of the keyshots can be severely distorted by homemade and slide-show videos which often cut and paste important scenes from the original video at arbitrary positions of the video.

In this paper, we explore Google-context information as a basis to perform web video summarization. Google-context provides useful cues for web video summarization. First, the search trend from *Google Trends* [1] can accurately pinpoint the time when the most important and defining events (*hot time*) of the topic-of-interest occur. Second, news articles crawled from *Google News* [2] can be paired with specific scenes from the video set to provide a more comprehensive summary of the important events. Furthermore, using a scene-news pairs representation effectively densifies the textual annotation of web videos compared to the textual features of the web video alone. Figure 1 shows the flow of the proposed framework. Given a query, the search trend from [1] and the upload count of the web videos over time is employed to detect the hot times for the topic. The news articles from [2] at the detected 'hot time' are then matched with tags from *near-duplicate threads* mined from web videos, subjected to time consistency constraint. The scene-news pairs from time-dependent alignment are finally presented to the users using a timeline representation.

2. HOT EVENT DETECTION

Hot time refers to the time when certain defining events for the topic occur. Identifying the important events is a rather subjective decision. For example, for the topic 'Somali Pirates', there is an unbroken flow of news reports on piracy activities throughout the years. To identify the set of important incidents that captures the attention of most users, we resort to Google Trends which provides two trend lines for reference: (a) the *search volume index* graph which reflects the number of searches posted by users on Google over time and (b) the *news reference volume* graph which reflects the number of times the topic appears on the Google News stories. For hot event detection, we use mainly the *search volume index* to infer the rise and fall of user interest across time since the search volume on the topic will naturally increase and decrease as the interest build up around some interesting events. To verify the rise in user interest, we further track the upload count of videos on Youtube. The two trends are normalized using min-max normalization and then merged to get the final user interest score. Denote the normalized Google Trends score as H_G and normalized video upload count as H_Y, the user interest score H can be retrieved by merging the scores as follows

$$H(t) = \max(H_G(t), H_Y(t)) \tag{1}$$

The hot times are defined as the time slots where the user interest score is larger than a particular threshold or $H \geq \gamma$. This also allows an option to control the level of details that users prefer to be exposed to when viewing the summarization result.

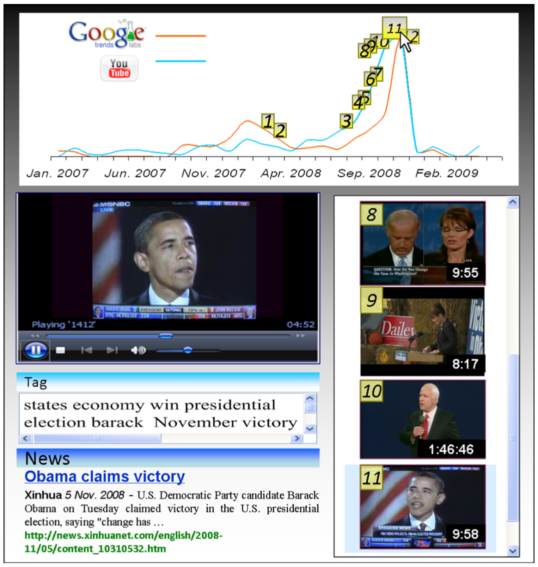
3. SCENE-NEWS ALIGNMENT

Given the set of hot times detected from the previous step, a fundamental task is to align web videos with news articles to retrieve the correspondence set $e_i = (v_i, n_i), i = 1 \ldots M$ such that a scene in web video v_i and news article n_i narrate the same event e_i, and M is the total number of events detected within the hot time slots. To extract important scenes from videos, visual near-duplicate segments are extracted using [10] and overlapping segments are grouped to form near-duplicate threads. Each thread consists of partial segments across different videos which feature the same scene. The scenes found by near-duplicate detection are typically important for the topic especially when they are repeatedly found in videos uploaded by different users. For each thread, the earliest video is chosen to represent the scene. This is based on the assumption that the first video where a scene first appears is the original video. Thus, only the set of earliest videos associated with the scenes is used to align with news articles.

Another problem when aligning web video to news article is that the textual context (tags, titles and descriptions) of web videos are often not sufficiently discriminative. To improve precision, the matching is further subjected to additional constraint which preserves the time consistency between the news articles publication date and video upload time. The similarity measure between a video v_i and n_j is based on the accumulated tf-idf score of the overlapping words in the metadata of v_i and the terms in n_j as follows

$$sim(v_i, n_j) = \begin{cases} \sum_w \text{tf-idf}(w) & \text{if } T(v_i) = T(n_j) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where w is a keyword that appears both in the tags, description or title of the web video v_i and the news article n_j while $T(\cdot)$ refers to the upload time of a web video or the publication date of the news article. We set the granularity of each time unit to one hot month. A greedy method is used to generate a one-to-one mapping between web videos and Google News where during each step, the corresponding scene-news with the highest score is selected and removed from the candidate pool. The process is repeated until no more correspondences can be extracted. The alignment scheme fulfills two important objectives. First it mines the set of corresponding scene-news which represents a single event for the topic. Second, the alignment process also automatically places the detected event to the best time slot which is agreed upon by both the news article and video. This is different from [7] which relies on an expensive quadratic optimization to determine the placement of keyshots based on the orders of near-duplicate frame-pairs. Thus, the proposed time-dependent matching process is generally faster and more reliable.



Tag

states economy win presidential election barack November victory

News

Obama claims victory

Xinhua *5 Nov. 2008* - U.S. Democratic Party candidate Barack Obama on Tuesday claimed victory in the U.S. presidential election, saying "change has ...

http://news.xinhuanet.com/english/2008-11/05/content_10310532.htm

Figure 2: Timeline summarization of Topic 2. The detected events are marked on the trend curves. When an event is selected, the corresponding scene, tags and news snippet are presented to users.

4. TIMELINE-BASED SUMMARIZATION

Timeline summarization is a popular practice by major news station like CNN to summarize major stories and is normally prepared by experts who are familiar with the topic. We adopt the same paradigm to visualize our summarization result. The interface of our system is depicted in Figure 2. The visualization consists of three parts: a) trend chart, b) event selection panel, and c) summarization section. The statistics of Google Trends and video upload count is displayed on the top and the detected events are marked on these trend curves. Through this chart, the *relative importance* of the events, *when* they occur and *the relationship* between them can be intuitively known. Below the trendline, the selection panel on the lower right depicts the list of keyframes, one for each event. Upon selection, the summarization section on the lower left panel will be updated to describe the selected event. In the panel, the tags and descriptions of the events give a brief textual overview of the event and the details are available by viewing the summary video or clicking the online news article. Note that only the extracted scene in the video is played. Compared to video skimming which generates a single synopsis video, the scenes are mapped specifically to each event. Furthermore, different from storyboard summarization, instead of providing a set of independent keywords, our system uses news snippet in the form of a proper sentence to provide a static summarization of an event. Therefore, users do not have to infer the events from keywords which tend to be vague and imprecise.

5. EXPERIMENTS

For evaluation, we evaluate our summarization system on five topics shown in Table 1. These are internationally hot topics which exhibit different characteristics. Topic such as 'Iran Nuclear Program' represents a complex and persistent topic that keeps recurring from time to time. 'Somali Pirates' is a simpler and homogeneous topic but is equally persistent. 'Sichuan Earthquake' and 'US Presidential Election 2008', on the other hand, reflect one-time topics where the latter sustains users' interest over a longer time span. 'Economic Collapse' represents topics that are broad and diverse compared to previous topics. For each topic, the videos are crawled from YouTube from April to May 2009. The important scenes of videos in a topic are extracted by performing near-duplicate thread extraction using [10]. The number of videos and near-duplicate threads are shown in Table 1. For hot time detection, the hot event parameter (Equation 1) is set empirically to $\gamma = 0.2$ for all topics. The time span of the videos is between Jan. 2007 and Apr. 2009 (28 months in total). Both quantitative and qualitative evaluation is presented. The quantitative evaluation gauges the performance of the system in terms of (a) hot spot detection and (b) scene-news alignment. Considering the preference of users is very subjective, we further conduct a user study to gather their feedbacks on the proposed system.

Table 1: Number of videos, detected near-duplicate threads and downloaded news articles.

	Topic	#Videos	#Threads	#News
1	Economic Collapse	1025	95	68
2	US President Election 2008	738	91	44
3	Somali Pirates	410	60	17
4	Sichuan Earthquake	1458	254	19
5	Iran Nuclear Program	1056	145	185

Table 2: Performance of event detection.

Topic	#GT	Hot Time Detection			#GT	Hot Event Detection		
	(HT)	#Det	Prec	Rec	(HE)	#Det	SR	R
1	5	7	0.71	1.00	17	16	0.00	0.31
2	7	7	0.86	0.86	7	12	0.00	0.33
3	2	2	1.00	1.00	2	9	0.11	0.56
4	2	2	1.00	1.00	6	11	0.37	0.27
5	16	15	0.73	0.69	22	21	0.10	0.24
Avg	6	6	0.86	0.91	11	14	0.11	0.34

#GT: Number of groundtruths, HT: Hot Time, HE: Hot Events
#Det (Hot Time): Number of months detected as hot time
#Det (Hot Event): Number of detected events
Rec: Recall, Prec: Precision, SR: Somewhat Related, R: Related

5.1 Event Detection

We evaluate the performance of hot time detection and hot event detection. For hot time detection, we first manually construct the groundtruth by listing the most important events for each topic. The list is produced by going through (a) Wikipedia and (b) expert summaries produced by CNN and MSNBC about the topic. Using one month as a time unit, the detected hot times are then compared against the groundtruth events and the result is shown in Table 2. Both precision and recall is around 0.9 which shows that Google Trends and video upload count are very reliable cues to detect the critical time for the topics. The number of detected hot times generally conforms to our expectation that the lasting events generate more hot times except for 'Somali Pirates'. Although the lifespan for this topic is long, only several dramatic incidents capture the attention of the public.

For hot event detection, we evaluate the performance of the scene-news alignment. The evaluation essentially evaluates the efficacy of the corresponding scene-news pairs for representing events. We define two scales for assessment as

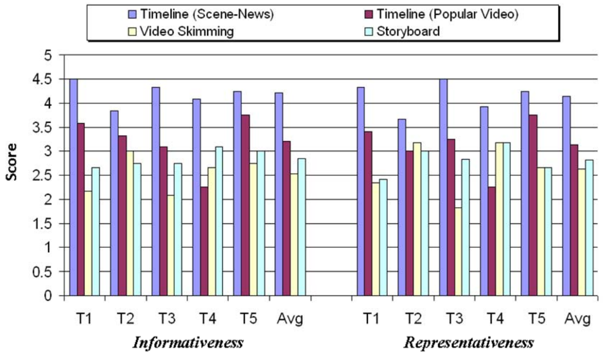


Figure 3: User-based evaluation for four different summarization techniques.

follows: (1) relevant: the scenes are correctly aligned to the news article and placed in the right location, and (2) somewhat relevant: the scene is partly or only vaguely related to the news article. Table 2 shows the result of event detection. In average, 34% of the detected scene-news pairs correspond to some specific events of the topic whereas 11% of the scene-news pairs are related to the topic but does not describe any specific events. Most of the detected scenes are from news web videos. We achieve good alignment performance for Topic 3 (Somali Pirates) and Topic 2 (US Presidential Election 2008) because these topics have very distinct events and therefore are easily distinguished from other events. Surprisingly, the alignment is also reasonably well for Topic 1 (Economic Collapse) although the topic is broad and contains many periphery sub-events. The alignment accuracy is lower for Topic 4 (Sichuan earthquake) since most of the videos for this topic are dedication videos which we do not consider as informative or useful for summarization. For Topic 5 (Iran nuclear program), the events in this topics are somehow more similar to each other since the different triggering events always invoke very similar remarks (and therefore tags) from different parties involved in the crisis.

5.2 User-based Evaluation

We compare timeline summarization as developed in Figure 2 with video skimming and storyboard summarization in [7]. Video skimming concatenates key shots in chronological order to form a synopsis video while storyboard is a static summarization which presents the result in terms of a series of keyframes with tag description embedded. In addition, we also compare to a baseline which selects videos with the highest view count from each month for timeline visualization. We invite twelve assessors to evaluate the systems where they are requested to assign a score between 1 (lowest score) to 5 (highest score) based on their experience with the four systems. We define two criteria for evaluation:

1. Informativeness: What is the coverage of the summarization result?
2. Representativeness: Do you think the summary is helpful to understand the topic well?

Figure 3 shows the user survey result for the systems. Compared to the baseline, using the scenes from scene-news alignment offers better result. This shows that the video clips from aligning with news articles are more accurate than using the most popular video of the month. Most assessors vote favorably for both the two timeline systems over video skimming and storyboard. The preference is due to the wealth of information encrypted within the simple interface. The video clips in timeline system are mapped to specific events, clearly positioned across time and the relative importance of each event can be deduced from the timeline. In contrast, the relationship between different keyshots is lost when viewing the synopsis video for video skimming. Furthermore, [7] uses shots as the basic story unit which are too abbreviated to handle more complex topics and therefore its effectiveness is limited to only one-time and short-duration topics, such as Topic 4 (Sichuan Earthquake). Compared to storyboard, the extracted scenes from web videos in the timeline system carry more information than static keyframes. Furthermore, users are able to extract more intelligible information from snippet-based description since they are expressed in proper sentences as opposed to keyword sets produced by storyboard.

6. CONCLUSIONS

We have presented the topical summarization of web videos for timeline-based multi-modal visualization. Google Trends and video upload count are used to determine the hot times while time-dependent alignment between news articles from Google News and Youtube videos produces the set of scene-news pairs to summarize a query topic. Quantitative evaluation shows that the proposed work is able to perform hot time and event detection reasonably well. This is also confirmed through the user survey which shows that users prefer timeline visualization over video skimming and static storyboard summarization.

7. ACKNOWLEDGMENTS

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119709).

8. REFERENCES

[1] *http://www.google.com/trends.*

[2] *http://news.google.com/.*

[3] H. Benoit and M. Bernard. Automatic video summarization. *Chapter in Interactive Video, Algorithms and Technologies*, pages 27–41, 2006.

[4] K. Y. Chen, L. Luesukprasert, and S. T. Chou. Hot topic extraction based on timeline analysis and multi-dimensional sentence modeling. *TKDE*, 19(8):1016–1025, 2007.

[5] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. *VLDB*, 2005.

[6] Q. He, K. Chang, and E. P. Lim. Analyzing feature trajectories for event detection. *SIGIR*, 2007.

[7] R. Hong, J. Tang, H.-K. Tan, C.-W. Ngo, S. Yan, and T.-S. Chua. Beyond search: Event driven summarization for web videos. *ACM TOMCCAP*, 2010.

[8] I. Ide, H. Mo, N. Katayama, and S. Satoh. Topic threading for structuring a large-scale news video archieve. *CIVR*, 2004.

[9] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. *CIKM*, 2004.

[10] H.-K. Tan, C.-W. Ngo, R. Hong, and T.-S. Chua. Scalable detection of partial near-duplicate videos by visual temporal consistency. *ACM Multimedia*, 2009.

[11] X. Wu, C.-W. Ngo, and Q. Li. Threading and autodocumenting news videos. *IEEE Signal Processing Magazine*, 23(2):59–69, March 206.

[12] Y. Zhai and M. Shah. Tracking news stories across different sources. *ACM Multimedia*, 2005.