12-2011

Cross media hyperlinking for search topic browsing

Song TAN

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

Hung-Khoon TAN

Lei PANG

Cross Media Hyperlinking for Search Topic Browsing *

Song Tan§ Chong-Wah Ngo§ Hung-Khoon Tan† Lei Pang§
songtan@student.cityu.edu.hk cwngo@cs.cityu.edu.hk
thkhoon@utar.edu.my leipang3@student.cityu.edu.hk

Department of Computer Science§
City University of Hong Kong
Kowloon, Hong Kong

Faculty of Information & Communication Technology†
University Tunku Abdul Rahman
Perak, Malaysia

ABSTRACT

With the rapid growth of social media, there are plenty of information sources freely available online for use. Nevertheless, how to synchronize and leverage these diverse forms of information for multimedia applications remains a problem yet to be seriously studied. This paper investigates the synchronization of multiple media content in the physical form of hyperlinking them. The ultimate goal is to develop browsing systems that author search results with rich media information mined from various knowledge sources. The authoring enables the vivid visualization and exploration of different information landscapes inherent in search results. Several key techniques are studied in this paper for developing these browsing features. These techniques include content mining and selection from web videos, space-time alignment of multiple media, and augmenting of search result with *when* and *what* information. We conduct both quantitative and user studies on a large video dataset for performance evaluation. Comparison with traditional techniques including storyboard summarization and video skimming are also presented.

Categories and Subject Descriptors

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

General Terms

Algorithms Performance Experimentation

Keywords

Video browsing, media content synchronization, event extraction, visual summarization

*Area chair: Dick Bulterman

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

1. INTRODUCTION

Browsing search results while digesting the key information from massive pool of returned data is by no means an easy task. The existing web retrieval techniques mostly concentrate on targeted search and present the results by relevancy ranking. Visualizing different facets of information offered by search results is often not supported. In the domain of multimedia particularly, a mechanism that could facilitate the navigation of unexplored results and switch of new search focus is highly demanded. One example of text-based exploratory search is the Google "Wonder Wheel" [1]. In this system, a wheel shows search terms related to a given query. By rolling different wheels, the users are exposed to diverse sub-topics related to the original query.

Developing tools such as Google Wonder Wheel for multimedia search exploration, nevertheless, faces two practical obstacles. First, audio-visual content is difficult to be analyzed if compared to text-based articles due to the semantic gap problem. The meta-data provided by users could be subjective and incomplete. There is no solution for the semantic organization of search results, for examples, to support visualization of thousands of web videos returned for a query topic. Second, web videos of hot topics are massively uploaded to the Internet. It is not surprise to see that these videos are inter-related with partially overlapped visual content [13, 20]. In general, there is no consensus of how these somewhat related but diverse and continue evolving clips should be managed for search results presentation. The information offered by a clip itself is normally short and concise. Providing an overview to visualize different facets of a search topic from thousands of short clips intuitively is a task different from the traditional video condensation [22].

This paper addresses the problems of search topic browsing for web videos through the hyperliking of visual and textual snippets mined from multiple online resources. We consider two issues that enable search result authoring for information exploration: uncover the semantic themes (or events) of a topic; and organize content of short clips for matching potential events. We refer the former issue as *event extraction*, and the later as *visual summarization*. Given hundreds or thousands of returned web videos, video segments which are near-duplicate of each others are first extracted. These segments could be implicitly grouped as visual snippets for signifying different events or themes of a search topic. The problem of *visual summarization* is tackled by uncovering the essential information which is frequently found in video clips. On the other hand, various knowledge sources including wikipedia pages and news articles are collected to

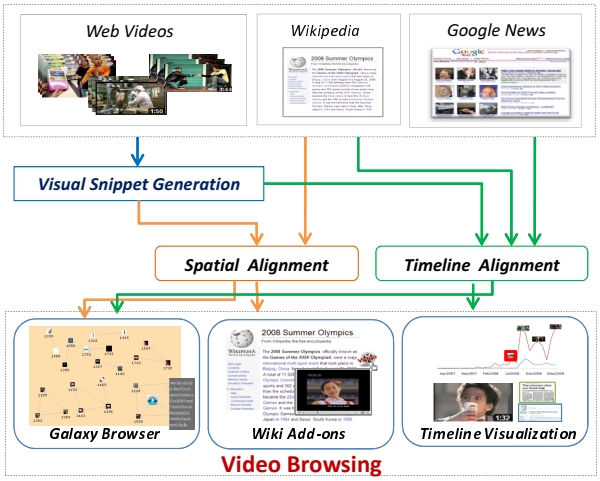


Figure 1: Overview of framework for cross media hyperlinking.

form the pool of textual cues potentially useful for describing visual snippets. Content alignment is then performed *temporally* and *spatially* to synchronize the visual and textual snippets from the video clips and knowledge sources respectively. The outcomes are "media hyperlinks" which link visual snippets with semantic description of events including the when and what information. Meanwhile, hyperlinks are also created for wikipedia pages whenever suitable videos describing the textual snippets are found. In brief, the problem of *event extraction* is tackled by exploiting external sources and aligning multiple information cues.

Figure 1 gives an overview of the proposed framework. First, web videos of a search topic are threaded to form islands of visual snippets based on partial near-duplicate processing (Section 3). Temporal alignment synchronizes the snippets with timeline extracted from wikipedia and news articles by network flow algorithm (Section 4). Meanwhile, spatial alignment matches the snippets with wikipedia page at the sentence level by tag cloud modeling (Section 5). Based on this space-time alignment, three browsing systems: timeline visualization (TL), wiki add-ons (WA) and galaxy browser (GA) are built (Section 6). TL allows timeline based visualization of search topics, and WA enables cross-reference interaction between different media. GA treats each visual snippet as a constellation and integrates TL and WA for exploratory search.

The main contributions of this paper are twofold. First, a computational framework is proposed for mining visual snippets, and more importantly, synchronizing the snippets with multiple online knowledge sources. Second, the framework seamlessly activates variants of browsing mechanisms for more efficient search result visualization and exploration.

2. RELATED WORK

Linear browsing, which clicks-and-plays clips sequentially, has been a traditional way of navigating search results. Nevertheless, different from text-based browsing, watching a long list of web videos of each lasts for a few minutes could be a tedious task. Grasping the overall picture of a search topic from watching hundreds of returned short clips also sounds infeasible. There have been numerous efforts indeed for improving the browsing efficiency from different viewpoints, including relationship mining, search re-ranking and multi-video summarization.

In commercial search engines such as YouTube, video cowatch statistics is a popular technique for video *relationship mining* [4]. The statistics, gathered from millions of web users, are derived by measuring the videos which are cowatched in the same viewing session. The suggested videos by cowatch are likely to be similar, and therefore, have a higher chance of falling into the same event of a topic. Nevertheless, while this feature (known as "related videos") is useful in general, the cowatch statistics does not explicitly tell the connection among videos. A more appealing mechanism is by depicting the video relationship through storyline. Early studies along this line include [7, 11, 18, 26, 28]. These works however are mostly constrained in the domain of news videos, where there are abundant professional-made text transcripts to be explored for constructing the story structure of a topic. For example, the mediaWalker developed by [10] provides a novel browsing interface for users to trace the development of news topics. The interface is built upon a directed graph structure that chronologically chains evolving events for browsing [11]. Similar work has also been demonstrated by [26] which represents evolving and peripheral events of news videos as a topic structure for browsing and summarization. Extending these works to web video domain, nevertheless, faces serious difficulties mainly due to the problem of "sparse text". Specifically the user-supplied text description is less reliable for building of topic structure to describe diverse video content. A recent effort by [15] is to densify the sparse text of web videos by a bipartite graph reinforcement model. The textual information are propagated and enriched by simply connecting videos which share common keywords. The propagated information, however, is limited and imprecise as text cue is considered as the only channel for spreading texts, while visual content similarity is ignored. As evidenced by [15] in the experiments, the amount of text after densification remains limited in describing topics. The work in [24] recently investigated the feasibility of constructing event structure for web videos by clustering user meta-data and visual feature trajectories. As shown in [24], constructing such structure is difficult as the features are rather noisy compared to news domain.

Another viewpoint for search results presentation is by *re-ranking* the retrieval list according to sub-topics [27, 29]. One typical approach is to optimize not only the relevancy but also the novelty of returned results. In other words, the search list is diversified by removing duplicate or highly similar documents. For multimedia search, in particular, this strategy is highly effective for preventing from the repeated watch of similar videos. A recent work which demonstrated the efficient elimination of duplicate web videos from search result is by [25], where both content and context features are integrated for real-time processing. Nevertheless, the approach in [25] can only effectively deal with full duplicates. For partial near-duplicates [20], which commonly exist in web videos, have not been seriously explored for re-ranking or sub-topic retrieval. Incorporating human interaction to refine search results is also a popular way of re-ranking. Well-known systems include InforMedia XVR [8], VisionGo [17] and RotorBrowser [5]. These systems, nevertheless, mainly focus on rapid localization of relevant items for targeted search, while lacking of mechanism for

presenting different views of search results. In [6], thread-based video browsing was developed for visualizing the results from various search dimensions such as time, high-level concepts and visual proximity. However, as threads are defined as the linked sequences of shots and each thread is viewed separately, it is difficult to infer event-level information from threads for describing search topics.

Multi-video summarization has been studied recently to condense large amount of videos for providing a gist overview of search results [9, 14, 23, 21]. A straightforward approach is by extending the algorithms for single-video summarization [23], where the similarity notion is not restricted to within the shots of a video but the pool of shots from multiple videos. Techniques such as maximal marginal relevancy (MMR) and minimum description length (MDL) were employed to select the best set of shots from multiple videos for summarization [14, 23]. The recent work in [9] presented an approach for mining the "key shots" to summarize the search results. The key shots are identified by near-duplicate keyframe detection. With key shots, informative tags are further extracted to generate visual-textual storyboard. Meanwhile, key shots are also optimally assembled based on their time order in the original clips for sequential skimming. Another work by [21] shown the use of external knowledge such as Google Trends to chronologically order web videos for timeline based visualization.

3. VISUAL SNIPPETS GENERATION

Web videos of a topic, though visually diverse, are always somewhat correlated. Given a large set of videos associated with sparse text, the task is to mine the snippets that are capable of depicting the major patches of information in these videos. For instance, given the videos about "Beijing Olympics 2008", the candidate snippets are the videos describing major events such as "torch relay", "Fuwa Mascots" and "Michael Phelps swimming". A commonly observed phenomenon in social media is that the important events of a topic are often discussed by a large group of Internet users and in a short period of time. For video websites, this phenomenon turns into the fact that similar videos of an event, possibly captured from different viewpoints and in different snapshots of time, are massively uploaded. In addition, some video segments are cut from a video and then pasted into other videos with modifications such as supplement of captions and sound effects [13, 20]. This phenomenon gives clue that a practical way of mining visual snippets without the employment of semantics is by detecting the partial near-duplicate segments among videos. With this assumption, we define two terminologies here. *Thread* refers to as a near-duplicate segment that is commonly shared by videos. *Visual snippet* refers to as the set of videos which are interlinked, either directly or transitively, by at least one thread. In the remaining of this section, we present our framework for partial near-duplicate detection, which integrates and refines several existing works in the literature.

Figure 2 shows an overview of our framework which is composed of three stages. The first stage processes videos by uniformly sampling keyframes and extracting local interest points [16]. The keyframes are represented with bag-of-visual-words (BoW). For efficiency, BoW of $20K$ words is built on a vocabulary tree of two layers and indexed with inverted file structure [19]. To minimize the effect of quantization loss due to the use of BoW, hamming embedding [12] is

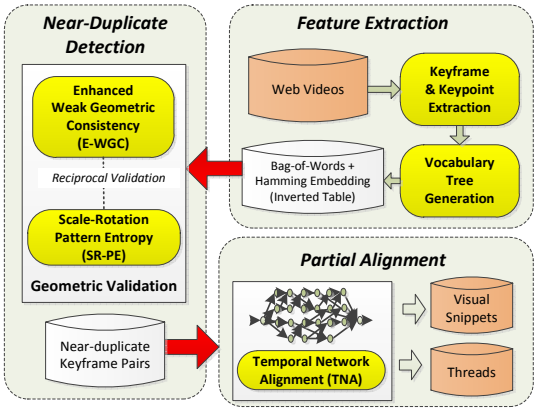


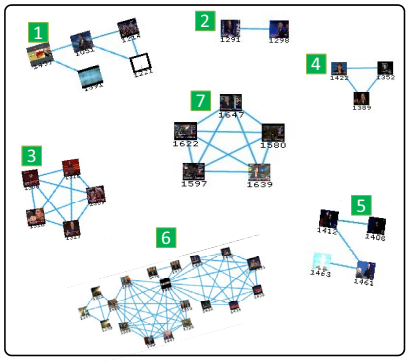
Figure 2: Framework of visual snippet generation.

employed to generate a short signature for each word. In the second stage, near-duplicate (ND) keyframes are retrieved based on BoW matching and signature verification. Basically each keyframe is treated as a query, and the set of candidate ND keyframes are retrieved. To verify the candidate NDs, reciprocal validation is performed by cross-checking the geometric parameters estimated by two algorithms: enhanced weak geometric consistency (E-WGC) [12, 31] and scale-rotation invariant pattern entropy (SR-PE) [30]. Both E-WGC and SR-PE estimate the scale and rotation parameters of two keyframes but with different approaches. E-WGC utilizes the by-product of local interest point detection to estimate parameters, and thus is sensitive to the robustness of point detection. SR-PE relies directly on the BoW matches for estimation, and thus is sensitive to the quality of BoW retrieval. We leverage this fact by performing candidate pruning whenever E-WGC and SR-PE give inconsistency estimation. In the third stage, the set of ND keyframes are consolidated to generate threads. We employ temporal network [20] which efficiently aligns the ND keyframes between videos by imposing temporal constraints. The set of aligned keyframes are then extracted from videos to form threads. Meanwhile, the videos which are linked through threads are also grouped to form visual snippets.

Figure 3 shows an example of visual snippets mined from the search topic "US presidential election 2008" which consists of 738 videos. Our framework extracted a total of 54 visual snippets. For illustration purpose, some of the snippets are aligned with the events in wiki timeline shown in Figure 3(c).

4. TEMPORAL ALIGNMENT

The temporal alignment aims to arrange the important web videos of a topic according to milestone events, and in addition, create hyperlinks for cross-referencing the videos and timeline information available online. One knowledge source which includes the event timeline is wikipedia [2], as shown in Figure 3(c). An intuitive approach for alignment is by directly matching the meta-data of web videos and the text description of event timeline. However, such matching may not be reliable since both ends contain only the snapshot description about events, of either a sparse set of texts from a video or a few short sentences from wiki page. In this section, we present a novel way of alignment by considering



(a) Visual snippet

(b) Keyframes

1. **2004 July 27** – *Barack Obama delivers the keynote address at the Democratic National Convention.*
2. **2008 August 25–28** – *The Democratic National Convention convenes in Denver, Colorado.*
3. **2008 September 1–4** – *2008 Republican National Convention convenes in Minneapolis-St. Paul, Minnesota.*
4. **2008 September 24** – *The first segment of Sarah Palin's interviews with Katie Couric airs on CBS News.*
5. **2008 October 2** – *Joe Biden and Sarah Palin appear at the vice presidential debate at Washington University in St. Louis.*
6. **2008 October 7** - *John McCain and Barack Obama appear at the second presidential debate at Belmont University, Tennessee.*
7. **2008 November 4** – *Election Day: Barack Obama and Joe Biden win 52.92 percent of the popular vote and 365 electoral votes to John McCain and Sarah Palin's 45.66 percent and 173 electoral votes.*

(c) Timeline

Figure 3: Examples of visual snippets for the search topic "US presidential election 2008". The correspondences between snippets (a) and events in wiki timeline (c) are labeled with numbers. The keyframes for snippet "3" are listed in (b).

four different sources: wiki page, news articles, videos and threads.

4.1 Problem Formulation

We employ a graph model, as illustrated in Figure 4, for temporal alignment. The model is composed of four layers of heterogeneous nodes. The nodes residing at the first layer represent the event entries extracted from a wikipedia page, while the second-level nodes are news ariticles related to a search topic. The third and forth layers are respectively the web videos and their threads. The edges between adjacent layers specify the similarity between nodes. Note that there are no edges between nodes of a layer. The problem of alignment is to find a path, which traverses from an event entry and reaches a thread, that optimizes the total similarity between any two adjacent nodes along the path. Denote the graph as $\mathbf{G}(\mathbf{N}, \mathbf{E})$, where $\mathbf{N}$ is the set of nodes and $\mathbf{E}$ is the set of edges. We formulate the problem as *network flow optimization*, and the objective function is to maximize:

$$\text{maximize} \sum_{e_{ij} \in \mathbf{E}} f(e_{ij}) w(e_{ij}) \qquad (1)$$

where the flow $f(e_{ij})$ is a binary indicator with value equals to 1 if nodes i and j are connected, and 0 otherwise. The term $w(e_{ij})$, which will be described in Section 4.1.1, signifies the importance (or weight) of an edge. Specifically, the optimization seeks for a path that has the largest capacity for carrying the flow through different layers (or information sources) of the graph. The capacity is characterized by the edge weights along the path. The result of optimization is a set of heterogeneous documents cross-referencing each other.

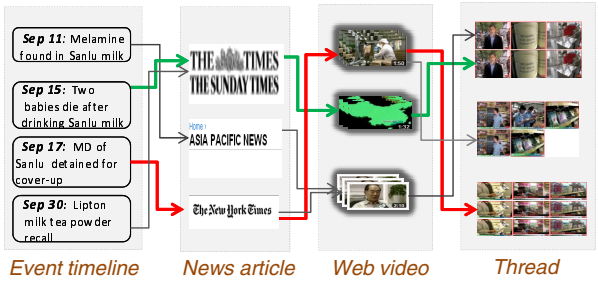


Figure 4: An illustration of graph model for aligning heterogeneous information sources.

There are three major significances with the introduction of the graph model in Figure 4. First, the second-layer news articles serve as "intermediate agents" before matching the meta-data of a web video and the short passage of an event entry in wikipedia. This layer densifies and enriches the amount of informative text for matching. In other words, both video and event entry can align with each other even in the extreme case where there are no common words between them, as long as there is a news article which bridges them. Second, only the videos belong to threads are considered for alignment. These videos are likely to be representative since they contain near-duplicate segments which are repeatedly reused for signifying major events and changes of perspective. Third, the result of aligning heterogeneous information sources can be directly utilized to develop the browsing interface which will be described in Section 6.

4.1.1 Edge Weighting

There are two kinds of edges in $\mathbf{G}$ for inter-relating different natures of nodes: binary and weighted edges. The former is for specifying visual similarity while the later is for text similarity. The binary edges basically group together the set of web videos according to the threads which are mined using the approach described in Section 3. Or equivalently, the edges link videos to their corresponding threads. The value of an edge is either 1 or 0 depending on the belonging of a video to a thread. The weighted edge, on the other hand, is used for characterizing node proximity, for example, between an event entry (or web video) $\mathcal{W}$ and a news articles $\mathcal{N}$. The edge weight is quantified using text cue as following:

$$Sim(\mathcal{W}, \mathcal{N}) = \sum_{w \in \mathcal{W} \cap \mathcal{N}} \text{TF-IDF}(w) \qquad (2)$$

Both nodes $\mathcal{W}$ and $\mathcal{N}$ are represented by a set of keywords. The similarity basically measures the number of common keywords in both nodes. Nevertheless, direct counting the number of common keywords may overlook the contribution of certain words to the similarity value. This is because the matching is usually conducted between a long article (news) and a short passage (wikipedia event entry) or a sparse set of texts (meta-data of a web video). In Equation (2), instead, the importance of a keyword w is weighted by TF-IDF computed based on the news article $\mathcal{N}$.

4.1.2 Network Flow Algorithm

The formulation in Equation (1) is typically a constrained binary programming problem that can be solved using network flow maximization [3]. By network flow, each edge e_{ij} is characterized by two terms, weight $w(e_{ij})$ and flow $f(e_{ij})$. Two artificial nodes, source n_{src} and sink n_{sink} nodes, are introduced. The source node is connected to all nodes of event entries such that all paths are originated from this node. Meanwhile, sink node is connected to all threads such that this is a node where all paths terminate. The solution of network flow should obey the equilibrium constraint so that the optimal path must not be broken at any point from the source to sink. This constraint can be characterized by network flow conservation where the net inflows and the outflows at a particular node must be equal to zero. In other words, Equation (1) can be effectively solved subject to the following conditions:

$$\sum_{e_{in} \in E_{in}(n)} f(e_{in}) - \sum_{e_{out} \in E_{out}(n)} f(e_{out}) = 0, \quad \forall n \in \mathbf{N} \quad (3)$$

$$\sum_{e_{out} \in E_{out}(n_{src})} f(e_{out}) = 1 \quad (4)$$

$$\sum_{e_{in} \in E_{in}(n_{sink})} f(e_{in}) = 1 \quad (5)$$

$$0 \leq f(e_{ij}) \leq 1, \quad \forall e_{ij} \in \mathbf{E} \quad (6)$$

where $E_{in}(n)$ and $E_{out}(n)$ denote the set of in-coming edges and out-going edges of node n respectively. Equations (3) to (6) impose the flow conservation constraints to control a well-behaved weight transfer from the source to sink node. The unimodality property [3] of network flow formulation ensures that the solution must be binary. Thus, the set of nodes traversed by the optimal path indicated by $f(.) = 1$ constitutes the solution. This path, which contributes the largest accumulated weights, corresponds to the various information sources that are best aligned. Since there always exists a valid path from the source to sink, by convexity property, Equation 1 will converge into a global solution [3].

In the implementation, we target to align each event entry in wikipedia page to a web video, as long as the accumulated weights of the path are significantly large. Thus, the network flow algorithm is iteratively run, in a way that the current optimal path is removed from the graph before the start of next iteration. The procedure stops when all the event entries are attempted for temporal alignment.

4.2 Efficient and Robust Alignment

The size of graph can become large when there are hundreds to thousands of web videos to be matched. On the other hand, the alignment also becomes difficult when the objective is to select one video out of this excessive number of noisy and diverse clips for matching against an event entry. An efficient solution is by detecting the potential hot times of a search topic, and then performing temporal alignment separately for each hot time. In other words, in each alignment, only the event entries, news articles and videos that fall within the hot time are considered for matching. This results in faster and more robust alignment, owing to a smaller pool of candidate videos which have higher likelihood to be matched is considered by the network flow algorithm.

We adopt the strategy in [21] to detect the hot times of a search topic. The detection is based on the statistics from Google Trends and Youtube video upload count. The former, especially, provides the search volume index to summarize the number of queries for a topic on Google search engine across time. The index reflects the temporal changes of user interest, and the peaks of index approximate the occurrence of milestone events under a topic. Similar observation is also found for Youtube video upload count, but there is normally a time lag between these two statistics. Similar to [21], we use month as the unit of hot time, and identify the hot months by first fusing the statistics from search volume index and video upload count. Thresholding technique is then employed to sample the hot months of a search topic. Temporal alignment is carried out sequentially for each detected hot month to align the event entries and video clips uploaded within the month.

To guarantee reliable matching, the recall of representative news articles is important. For example, using search query topic such as "US Presidential Election 2008" will easily result in miss of articles about critical events such as "Sarah Palin's teen daughter pregnant", which has impact for Palin's candidacy. To tackle this problem, the frequent keyword sets associated with each hot month are mined for retrieving online news articles. The keyword sets are obtained by the tri-gram frequent itemset mining algorithm. The examples of keyword sets include "democratic national convention", "barack obama debate" and "palin teen pregnant". These sets broaden the coverage of a search topic, and are more discriminative in recalling back the relevant and representative articles for temporal alignment.

5. SPATIAL LINKING

While timeline information provides valuable clue for event detection and alignment, not all the wiki pages have a well-structured time-event entries as shown in Figure 3(c). The wiki page of a topic is usually a free-style description of text composed of sections, paragraphs and sentences. The narration of an event could happen at a sentence, a paragraph, or even arbitrarily in different sections of a page. Furthermore, several events may be interchangeably reported, but altogether depicting the evolution of a topic. In most cases, the time information could be missing or is ambiguous to be resolved if without the analysis of contextual relationship among sentences. In short, temporal alignment between events and web videos could be difficult when the time cue is weakly described.

Instead of performing temporal alignment, this section describes the extraction of textual snippets directly from wiki pages by matching with web videos. We name this process as "spatial alignment" since time cue is not explicitly utilized, and more importantly, the matching is conducted directly at the sentence level. Intuitively, the matching could be difficult, since in contrast to temporal alignment, additional cues such as timeline and news articles are not available for exploitation. To guarantee robust alignment, we leverage the tag cloud pooled from the web videos of a visual snippet for matching. More specifically, rather than brute-force matching the sentences and videos for finding the best alignment, the matching is conducted between sentences and visual snippets.

The problem statement of spatial alignment is defined as follows. Denote $\mathcal{V}$ as the tag cloud of a visual snippet, and $\mathbf{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m\}$ as the set of m sentences from a wiki page. The task is to find the best sentence $\mathcal{S}^*$ that narrates

$\mathcal{V}$:

$$\mathcal{S}^* = \arg max_{\mathcal{S}_k \in \mathbf{S}} \{\mathcal{V} \cap \mathcal{S}_k\} \quad (7)$$

where the matching is based on counting the number of keywords in a sentence that also appear in the visual snippet. In our implementation, more than one sentence, specifically a passage of several sentences, might be extracted for a $\mathcal{V}$ if the retrieved sentences with high match are spatially close to each other. This also ensures that the contextual narration among sentences be well preserved for describing $\mathcal{V}$. The end result of alignment is a set of $(\mathcal{V}, \mathcal{S}^*)$ pairs. We further select one or several videos from a snippet which contribute the most number of words to tag cloud of $\mathcal{V}$ as the representatives. These videos are hyperlinked to the corresponding wiki page at the sentence-level for cross-reference browsing.

Using visual snippets as the unit for matching offers several advantages. First, a snippet, which is formed by videos inter-linked by threads, offers a denser and more comprehensive description of one or several events. Thus, the richer set of texts ideally will provide a better ground for more reliable sentence-level matching. Second, redundant matching, such as the repeated match of similar videos to different sentences, can be avoided. This is simply because the sets of partially near-duplicate videos have been clustered by snippets before matching. Third, the number of snippets is significantly less than videos, and thus the matching can be carried out more efficiently.

6. TOPICAL BROWSING

The techniques presented in sections 4 and 5 enable three types of browsing systems: (a) timeline based visualization, (b) wiki add-ons, and (c) galaxy browsing, as illustrated in Figure 5. Each system has its own features, and offers different experiences for exploring search topics. Timeline based visualization links together web videos and news articles of milestone events, while providing a gist overview of event evolution over time. Wiki add-ons, on the other hand, cross-references wikipedia pages and web videos, where a user can jump back and forth between two different media. Galaxy browsing supports the exploration and identification of information patches, by showing a galaxy view of visual snippets and how different videos are inter-linked with each other. A user can slide cursor to navigate different snippets and zoom in to view the detailed information.

Among these three systems, timeline based visualization explicitly links the major videos and news articles of events to time. The available key events offered by the pool of retrieved videos of a search topic are revealed to users by indicating the time of occurrence on the trend chart. In Figure 5(a), a user can click an event label on the trend chart, and the system will prompt an essential video together with textual snippets summarizing the event. The snippets are extracted from wiki timeline and the news articles which are most aligned with the web video based on the network flow algorithm as described in Section 4.1.2. By further clicking the link associated with news snippet, the corresponding online news article will be shown. The trend chart also gives clues to the relative importance of events, which provides another dimension for users to navigate the search result.

While timeline visualization is efficient in depicting the temporal evolution of a search topic, there are topics which are not well structured with time. Even there is, the time information may not be available online for use. Wiki add-ons gives another view of browsing search result when timeline is not available, and is especially suitable for users who have little or no knowledge about the events of a search topic. As shown in Figure 5(b), when reading a wiki page, users can click sentence for viewing one or multiple videos hyperlinked to the description.

Galaxy browsing bundles timeline visualization and wiki add-ons, in addition, visualizes the video relationship in the form of visual snippets. Each snippet is hyperlinked with text snippets which may include *when* and *what* information extracted from wikipedia pages based on temporal and spatial alignment. The snippets help users quickly understand the information they encounter, and make decision about which navigational path to follow. By clicking a snippet, the system will bring users to the set of available information. As shown in Figure 5(c), these include the timeline of event under investigation and the paragraphs extracted from wiki page with key sentences being highlighted.

7. EXPERIMENTS

We used the web video dataset in [24] consisting of 22 search topics and approximately 22,000 web videos for performance evaluation. The topics were selected based on the top 10 news happened during 2006 to 2009 as recommended by CNN, TIMES and Xinhua. The search topics, as listed in Table 1, were issued to YouTube and all the returned videos were crawled to construct the dataset. Table 1 shows the details of dataset for 22 search topics, along with the number of snippets and threads being generated by our approach in Section 3. We manually check the result. The accuracy of threading is approximately 80%. About 60% of snippets are semantically meaningful. Noisy snipppets are mostly formed due to anchor person shots and logos.

The collected dataset shown in Table 1 basically covers different characteristics of search topics for experimental evaluation. The topics differ in terms of duration persistence and content complexity. For example, the videos about topic-18 (Iran nuclear program) last for several years from 2006 to 2009, while the videos about topic-7 (Virginia tech massacre) only span for one month from April to May of year 2007. However, the persistence of a search topic does not necessarily imply the complexity of a topic. For example, topic-22 (Michael Jackson dead) lasts for few months but the videos comprise a series of sub-topics like "cardiac arrest", "moonwalk", "memorial tribute" and "last rehearsal". In contrast, the videos for topic-12 (California wildfire) happen periodically for several years, but the content is relatively homogeneous.

7.1 Objective Evaluation

We conduct evaluation separately for temporal alignment and spatial linking. The experiment for temporal alignment is carried out for 9 topics which have the timeline information as shown in Table 1. While for spatial linking, we filter few topics which have no English version of wiki pages from experiments. In both experiments, the video titles, descriptions and tags are extracted as the meta-data. Standard pre-processing such as stop word removal and Porter stemming is applied to the meta-data before matching.

Temporal Alignment. We compare our approach based on network flow algorithm with a baseline which matches the event entries extracted from wiki timeline and the meta-data of videos directly. The baseline is basically a greedy ap-

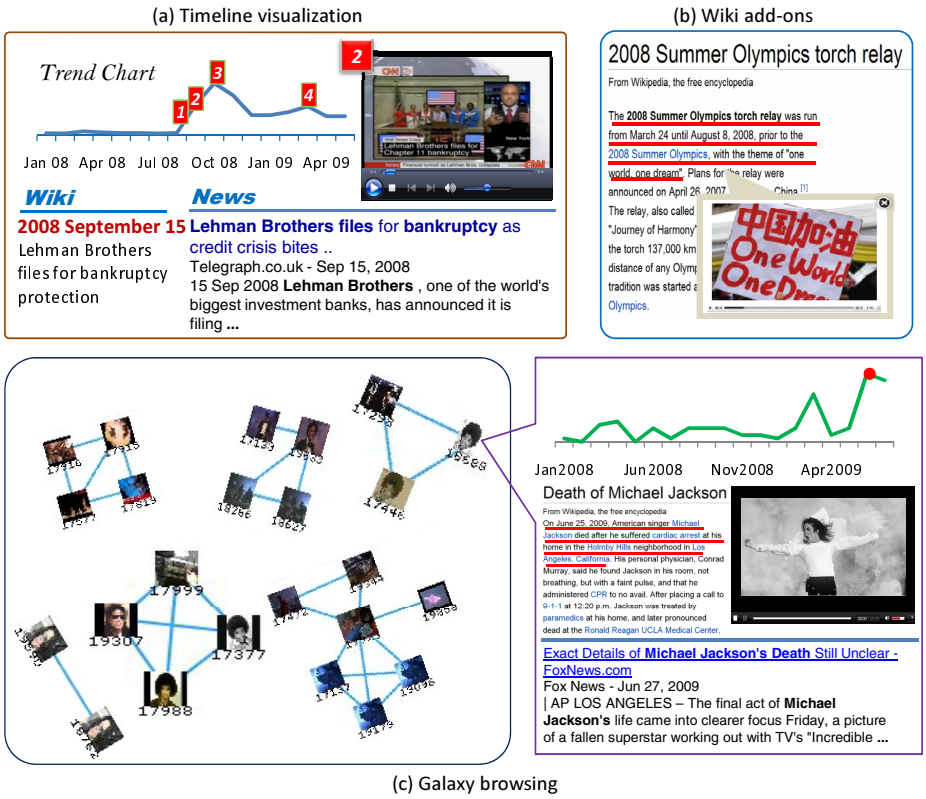


Figure 5: The three developed browsing systems based on media content synchronization.

proach similar to [21], where one-to-one matching between videos and events is performed within each hot month of a topic. Similar to our approach, the video-event similarity is measured based on the accumulated tf-idf score between them. The major difference between both methods are: our approach considers news articles as the bridge for matching event entries and videos, and in addition, threads mined from the set of returned videos are used to filter less important videos from matching. We evaluate the performance based on precision and recall. Recall measures the number of events being matched to relevant videos. Precision examines the ability of an approach in reducing the false alarm among the established matches. A match is regarded as relevant when the video is related or partially related to the target event. Since "relevancy" is subjective, as long as the event is the main focus of video, we regard the match as relevant. Since our approach also considers news articles, we impose an additional constraint that the matched news article should also be somewhat related in order to acknowledge an match as relevant.

Table 2 shows the result of temporal alignment. The result indicates that our approach outperforms baseline in terms of precision and recall. Especially, among the established alignments, approximately 75% of them are regarded as relevant, which is significantly better than the baseline. Few topics, however, are suffered from lower precision due to the fact that the underlying events are entangling and not discriminative. For instance, topic-1 (Economic collapse) contains the keywords "bank", "mortgage" and "federal" in several entries of events. Similarly for topic-10 (Melamine) where the common keywords in events are "Sanlu", "Milk" and "Melamine". The recalls for both approaches are not high. This is simply because wiki page tends to give a more detailed description of a topic than videos. For instance in topic-4 (Mumbai terror attack), there are event entries describing the preparation before the terror attack such as arming with weapon, leaving for Mumbai by ship and the route. We do not find any videos which could match to these events. Similar observation is also found for topic-1 (Economic collapse) and topic-21 (Swine flu). In general, compared to baseline, our approach benefits from the proposed graph model in the way that the news articles are able to densify the sparse text in video meta-data and lead to better alignment performance. The use of threads also enhances the chance of matching relevant videos to events.

Spatial Linking. To show the benefit of using visual snippets for matching sentences, we compare our approach to a baseline which matches the meta-data of popular videos to sentences. The popularity of a video is measured based on user view count. For each topic, we rank the videos by popularity and select a subset of them for matching. Table 3 shows the results in terms of precision. Because the result of baseline approach is noisy in general, the precision is calculated up to a depth of k, where k is the number of events detected by our approach. As indicated in Table 3, our approach using the tag cloud pooled from visual snippets offers a much better performance than the baseline. From our analysis, popular videos are mostly interesting videos but are not necessarily informative. In addition, without employing visual snippets to constraint the matching, some sentences are matched to videos of similar version. This in-

Table 1: Dataset

ID	Topic	Video #	Visual Snippet #	Thread #	Timeline (News #)
1	Economic collapse	1025	76	95	Yes (84)
2	US president election 2008	738	54	91	Yes (96)
3	Beijing olympics	1098	60	121	
4	Mumbai terror attack	423	42	52	Yes (65)
5	Russia Georgia war	749	63	112	Yes (42)
6	Somali pirates	410	47	60	
7	Virginia Tech massacre	682	59	72	Yes (32)
8	Israel attack gaza	802	72	113	
9	Beijing olympic torch relay	655	42	74	
10	Melamine	783	54	81	Yes (68)
11	Sichuan earthquake	1458	139	254	
12	California wildfires	426	10	14	
13	London terrorist attacks	784	58	122	
14	Oil price	759	49	65	
15	Myanmar cyclone	613	80	133	
16	Kosovo independence	524	24	38	
17	Russia presidential election	1335	96	120	
18	Iran nuclear program	1056	85	145	Yes (72)
19	Israeli-Palestine peace	586	31	54	
20	Korea nuclear test	1060	88	118	Yes (32)
21	Swine flu	1153	78	105	Yes (146)
22	Death of Michael Jackson	1865	139	252	

Table 2: Performance evaluation for temporal alignment. The parenthesis in the second column indicates the number of hot months.

ID	Event #	Detected events	*Our approach*		*Baseline*	
			Prec	Rec	Prec	Rec
1	30 (3)	6	0.33	0.07	0.17	0.03
2	26 (5)	8	0.62	0.19	0.25	0.08
4	4 (1)	2	1.00	0.50	1.00	0.50
5	17 (1)	3	1.00	0.18	0.33	0.06
7	6 (1)	2	1.00	0.33	0.50	0.17
10	31 (2)	10	0.40	0.13	0.50	0.16
18	3 (1)	2	1.00	0.67	0.50	0.33
20	3 (1)	3	0.67	0.67	0.67	0.67
21	22 (1)	4	0.50	0.10	0.25	0.05
Mean		5	0.73	0.31	0.46	0.23

deed indicates the significance of having visual snippets for minimizing redundant matches.

By investigating the hyperlinking results, among the correct alignments, basically a match between a visual snippet and sentence(s) indicates a different event. By comparing this result to temporal alignment, more events, of average 10 events per topic, are detected. This is because sentences are contextually more informative compared to the wiki timeline which summarizes events with short descriptions. Take the sentence from topic-10 (Melamine): "A number of countries have imposed blanket bans on Chinese milk products or its derivatives, among which are Bangladesh, ... and Malaysia which have imposed specific bans on mainland Chinese dairy products which have tested positive for Melamine" as an example. The context provides informative background for reliable matching of the sentence with video meta-data. For topics where timeline information is not available, for example topic-11 (Sichuan earthquake), the detected events include "when and where it happens", "the aftermath" and "the memorial". The extracted sentences corresponding to these events provide objective description for summarizing different facets of visual snippets mined from this topic.

Table 3: Precision for spatial alignment. The parenthesis indicates the number of near-duplicate videos being linked to a wiki page.

ID	Detected events	*Our approach*	*Baseline*
1	6	0.83	0.83 (2)
2	9	0.67	0.33 (2)
3	9	0.89	0.56 (4)
4	7	0.71	0.71
5	16	0.81	0.69 (2)
6	15	0.67 (2)	0.20
7	4	0.75	0.75 (3)
8	8	0.75	0.75
9	6	0.67	0.00 (4)
10	6	0.50	0.33 (3)
11	11	0.82	0.55 (2)
14	2	0.50	1.00 (2)
15	7	0.71	0.71 (2)
16	7	0.57	0.43
17	14	0.64	0.32
18	12	0.83	0.25 (3)
21	14	0.64	0.57
22	17	0.88	0.41 (4)
Mean	10	0.71	0.52

7.2 User Studies

We conduct subjective evaluation to compare five video browsing systems: timeline based visualization (TL), wiki add-ons (WA), galaxy browsing (GA), video skimming (VS) and static storyboard (SB). The aim of evaluation is to study the practicality of these systems for browsing and exploring the returned web videos of a search topic. VS and SB are two classical ways of browsing videos [22]. We adopt the implementation in [9] for developing VS and SB. VS concatenates

the key shots of web videos chronologically to generate a synopsis video. In [9], key shots are defined as the set of near-duplicate shots. In our implementation, we treat the set of threads mined using our framework as the key shots. SB, on the other hand, is a textual-visual board that puts together the keyframes and tags extracted from key shots as a static summarization of videos.

We define the following four criteria for evaluating the browsing systems:

- *Coverage*: To what degree the presented multimedia content retains the coverage of a search topic?
- *Precision*: Is the presented content always precise and relevant to a search topic?
- *Conciseness*: Is the presented content always concise and contains no redundant information?
- *Presentation*: To what degree the organization, presentation and functionality of the browsing system help in exploring the events in a topic?

The first three criteria basically assess the ability of a system in mining the essential information from web videos. The information should cover as much key events as possible, while being capable of minimizing the irrelevant and redundant events. The last criterion examines whether the mined content are presented in a manner that allows the efficient exploration of different events in a topic. We further define two criteria for evaluating the overall user experience in using a browsing system:

- *Engagement*: How useful the browsing system is in providing guidance for exposing the different sub-topics of a search topic?
- *Acceptance*: Will the browsing system lead to better user experience, if it is being used by social media websites such as YouTube and Wikipedia?

We invite 16 evaluators from different education backgrounds including computer science (12), linguistics (1), law (1), industry (1) and business (1). All the evaluators are familiar with video social media websites such as YouTube. The average age of evaluators is 28 and the standard deviation is 3.5. In the evaluation, topics are randomly assigned such that each topic is evaluated by four human subjects. Each evaluator is requested to use different browsing systems to explore videos under a topic. To minimize the carryover effect, evaluators are instructed to leave "wash-out" time between evaluating different systems. Furthurmore, the effect is also counterbalanced by assigning different order of topics for different subjects. After finishing exploring a topic, an evaluator is asked to rate the five[1] systems based on the aforementioned criteria. The rating is in the scale of 1 to 7, with 7 being the highest score (highly confident), and 1 being the lowest (not acceptable).

We summarize the evaluation result by averaging the rating for each topic from all the 16 evaluators. Figures 6(a) to 6(d) show the statistics for the first four criteria. Overall, the evaluators give higher rating to our three proposed systems than video skimming (VS) and storyboard (SB). From

[1]For the case where a topic does not have timeline information, only four browsing systems will be evaluated.

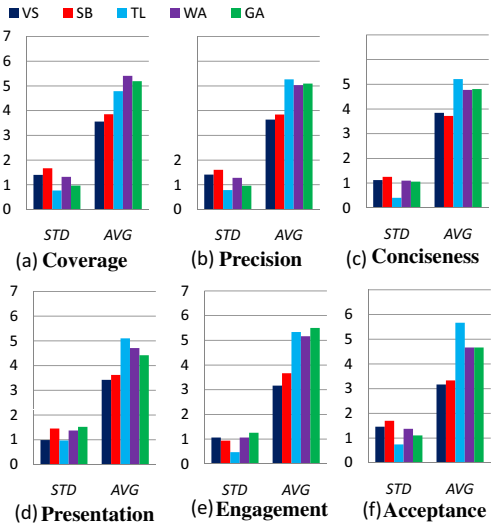


Figure 6: The average rating (AVG) and standard deviation (STD) of user studies for video skimming (VS), story board (SB), timeline visualization (TL), wiki add-ons (WA) and galaxy browsing (GA).

our analysis, evaluators indeed rely on the context description to understand the visual content. SB which provides a list of static images with sparse tag description is not informative in general. The concatenated keyshots in VS, on the other hand, do not preserve the event relationship and result in the poorest performance in terms of the criteria coverage and precision. Timeline visualization (TL), in contrast, encodes a wealth of descriptions extracted from wiki event timeline and news snippets, and achieves the highest score in the criteria presentation and conciseness. The trend chart, especially, offers a gist overview of key events for exploring search topics. Comparing all the systems, wiki add-ons (WA), which gives detailed description of a topic and is supplemented with aligned videos, obtains the highest score in coverage. However, the presented content is not as precise and concise as TL and galaxy browsing (GA). GA, interestingly, obtains high scores in all the four criteria. The feedbacks from evaluators indicate that the information presented by GA is comprehensive, but for ordinary users, GA is also relatively complicated to use than TL and WA.

Figures 6(e) and 6(f) show the rating for the overall user experience of using the five systems. The proposed TL, WA and GA show significantly better performance than VS and SB in terms of the two criteria engagement and acceptance[2]. GA achieves the highest score in search engagement. Most evaluators agree that GA is comprehensive and provides better guidance for exploring different details of a topic. The evaluators indicate that GA is especially suitable for browsing search topics of rich visual content such as topic-3 (Beijing Olympics), topic-4 (Mumbai terror attack), and topic-11 (Sichuan earthquake). For topics such as topic-1 (Economic collapse) and topic-18 (Iran nuclear program)

[2]The conclusion is based on AVONA (analysis of variance) test. First, the calculated F value exceeds the tabulated F value for p=0.001. Second, the 5% least significant difference between the means of any two systems also shows that the three proposed systems are significantly better.

which require understanding of background history, WA is still the best choice. TL, on the other hand, is benefited from the provision of timeline for browsing the event evolution of a search topic. Majority of evaluators give high rating that they would like to use TL for browsing videos in social media. In general, the feedbacks collected from evaluators indicate that the ability of synchronizing various online resources for content presentation is an interesting feature for multimedia information search.

8. CONCLUSIONS

We have presented our work for visualization and exploration of video search results by cross hyperlinking of different media. The visual content selection, based on automatic thread and snippet generation, provides a solid ground for elegant matching of various knowledge sources such as news articles and wikipedia pages. The proposed algorithms allow the effective extraction of events for augmenting search results with rich information. The user evaluation on the three developed browsing systems clearly indicates the advantages of search result authoring for information browsing and exploration.

Currently, we consider media content synchronization for three major knowledge sources: Youtube videos, Wikipedia and Google news. Future extension includes the generalization of alignment algorithms when more heterogeneous sources are incorporated. Potential issue to study is information extraction from contradictory content and diverse formats. In terms of computational complexity, our work does not support real-time synchronization of media in response to a user query. Processing one thousand videos could take about 1 to 2 days using a dual core CPU. Most of the computational time is consumed by the extraction of visual snippets. When distributed computing is considered, however, a search topic is expected to be available for browse within few hours. As "Google Wonder Wheel", it is suffice to show the search result augmenting and visualization for popular topics, which can be offline processed and mined from query log analysis.

ACKNOWLEDGEMENT

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 119610).

9. REFERENCES

[1] Google wonder wheel. http://www.googlewonderwheel.com/.

[2] Wikipedia. en.wikipedia.org/.

[3] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Reading, Massachusetts, 1993.

[4] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for YouTube: Taking random walks through the view graph. In *WWW*, 2008.

[5] O. de Rooij, C. G. M. Snoek, and M. Worring. Query on demand video browsing. In *ACM Multimedia*, 2007.

[6] O. de Rooij and M. Worring. Browsing video along multiple threads. *IEEE Trans. on Multimedia*, 12(2):121–130, Feb 2010.

[7] P. Duygulu, J. Y. Pan, and D. A. Forsyth. Towards auto-documentary: Tracking the evolution of news stories. In *ACM Multimedia*, 2005.

[8] A. G. Hauptmann, W. H. Lin, R. Yan, J. Yang, and M. Y. Chen. Extreme video retrieval: Joint maximization of human and computer performance. In *ACM Multimedia*, 2006.

[9] 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*, 2011.

[10] I. Ide, T. Kinoshita, T. Takahashi, S. Satoh, and H. Murase. mediawalker: A video archive explorer based on time-series semantic structure. In *ACM Multimedia*, 2007.

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

[12] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

[13] J. R. Kender, M. L. Hill, A. Natsev, J. R. Smith, and L. Xie. Video genetics: A case study from YouTube. In *ACM Multimedia*, 2010.

[14] Y. Li and B. Merialdo. Multi-video summarization based on AV-MMR. In *CBMI*, 2010.

[15] L. Liu, L. Sun, Y. Rui, Y. Shi, and S. Yang. Web video topic discovery and tracking via bipartite graph reinforcement model. In *WWW*, 2008.

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), Nov 2004.

[17] H. B. Luan, S. Y. Neo, H. K. Goh, Y. D. Zhang, S. X. Lin, and T. S. Chua. Segregated feedback with performance-based adaptive sampling for interactive news video retrieval. In *ACM Multimedia*, 2007.

[18] S. Y. Neo, T. Ran, and et. al. The use of topic evaluation to help users browse and find answer in news video corpus. In *ACM Multimedia*, 2007.

[19] D. Nistĺer and H. Stewĺenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

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

[21] S. Tan, H. K. Tan, and C. W. Ngo. Topical summarization of web videos by visual-text time-depedent alignment. In *ACM Multimedia*, 2010.

[22] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM TOMCCAP*, 3(1), 2007.

[23] F. Wang and B. Merialdo. Multi-document video summarization. In *ICME*, 2009.

[24] X. Wu, Y. Lu, Q. Peng, and C. W. Ngo. Mining event structure from web videos. *IEEE Multimedia Maganize*, 18(1), Jan 2011.

[25] X. Wu, C. W. Ngo, A. G. Hauptmann, and H. K. Tan. Real-time near duplicate elimination for web video search with content and context. *IEEE Trans. on Multimedia*, 11(2):196–207, Feb 2009.

[26] X. Wu, C. W. Ngo, and Q. Li. Threading and autodocumenting in news videos. *IEEE Signal Processing Magazine*, 23(2):59–68, Mar 2006.

[27] C. Zhai, W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *ACM SIGIR*, 2003.

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

[29] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma. Improving web search results using affinity graph. In *ACM SIGIR*, 2005.

[30] W. L. Zhao and C. W. Ngo. Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection. *IEEE Trans. on Image Processing*, 19(2), Feb 2009.

[31] W. L. Zhao, X. Wu, and C. W. Ngo. On the annotation of web videos by efficient near-duplicate search. *IEEE Trans. on Multimedia*, 12(5), July 2010.