Hierarchical Visualization of Video Search Results for Topic-Based Browsing

Yu-Gang Jiang, Jiajun Wang, Qiang Wang, Wei Liu, and Chong-Wah Ngo

***Abstract*—Existing video search engines return a ranked list of videos for each user query, which is not convenient for browsing the results of query topics that have multiple facets, such as the "early life," "personal life," and "presidency" of a query "Barack Obama." Organizing video search results into semantically structured hierarchies with nodes covering different topic facets can significantly improve the browsing efficiency for such queries. In this paper, we introduce a hierarchical visualization approach for video search result browsing, which can help users quickly understand the multiple facets of a query topic in a very well-organized manner. Given a query, our approach starts from the hierarchy of its textual descriptions normally available on Wikipedia and then adjusts the hierarchical structure by analyzing the video information to reflect the topic structure of the search result. After that, a simple optimization problem is formulated to perform the video-to-node association considering three important criteria. Furthermore, additional topic facets are mined to complement the contents of the existing semantic hierarchies. A large YouTube video dataset is constructed to evaluate our approach both quantitatively and qualitatively. A demo system is also developed for users to interact with the proposed browsing approach.**

***Index Terms*—Search result visualization, video search, hierarchical structure, visual analysis.**

I. Introduction

THE number of videos on the Internet is growing explosively, which poses a great challenge on automatic video search, a technique urgently needed in many applications. Existing video search engines, such as YouTube, visualize the search results with a ranked list. The simple list structure may be suitable for locating a movie trailer or a music video, but cannot handle queries with complex topic structures. For example, a search of the famous film director "Steven Spielberg" returns more than 200,000 videos, covering different aspects of his films, interviews, achievements, personal life, etc. A ranked list mixed with videos in different topic facets is difficult for the users to locate their information needs. In this case, a semantically structured hierarchy with nodes covering all the critical topic facets becomes a much better choice, which is the focus of this paper. With an informative hierarchy, users can easily interact with the interface to browse the query topic according to their interests. For instance, one may first have a coarse understanding of the query topic by skimming through the high-level nodes, and then delve into more details by following a specific node downwards.

A good browsing experience of videos is closely related to the quality of the topic hierarchy, which should be semantically cohesive, and each node of it should be labeled with human-readable phrases. Obviously, manually creating topic hierarchies for every query is impractical simply due to the great number of the search topics. For the problem of visualizing search results, several previous methods of dividing search results into multiple groups include classification and clustering, but straightforwardly using the two kinds of methods is not suitable as the grouping results are not reliable. Specifically, classification also requires a large number of training samples that are not available in this scenario. Meanwhile, the purely data-driven clustering needs no training samples, but a major drawback is that the extracted clusters lack human-readable labels and are extremely hard to be interpreted semantically.

Due to the significant difficulty of creating a high-quality topic hierarchy manually from scratch, in this work we follow a Prototype-based Hierarchical Clustering (PHC) method [1]. The PHC incorporates a predefined prototype hierarchy as the basis to visualize the results, with which the quality of the hierarchy can be ensured. We use Wikipedia as the source of our prototype hierarchies as most queries with complex topic structures have related description page on Wikipedia. With crowdsourcing techniques, the pages on Wikipedia are very informative and highly organized with semantic hierarchies from coarse levels to fine-grained levels, which is a good match to our needs. It is worth noting that our visualization approach is not designed to be used for every user query. Instead it is only suitable for a set of queries that have multiple facets. The special query set can be offline determined by mining search logs and the related results can be precomputed to be used for online search. This is a common practice in modern search engines, where different types of queries are responded by different techniques.

With the initial prototype hierarchy from Wikipedia, the main technical issue of our approach is to adapt the hierarchy based on the contents of the returned videos, which is highly important since the prototype hierarchies from Wikipedia are originally used to organize text contents. We address this problem in a

Manuscript received May 1, 2016; revised August 11, 2016; accepted September 21, 2016. Date of publication September 28, 2016; date of current version October 19, 2016. This work was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Grant CityU 11210514, in part by the NSF China under Grant 61572134 and Grant U1509206, and in part by the STCSM, Shanghai, China, under Grant 16QA1400500. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Yingcai Wu.

Y.-G. Jiang, J. Wang, and Q. Wang are with the School of Computer Science, Fudan University, Shanghai 201203, China (e-mail: ygj@fudan.edu.cn; jiajunwang13@fudan.edu.cn; qiangwang14@fudan.edu.cn).

W. Liu is with Tencent, Inc., Shenzhen 518057, China (e-mail: wliu@ee.columbia.edu).

C.-W. Ngo is with the Department of Computer Science, City University of Hong Kong, Hong Kong, China (e-mail: cscwngo@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2614233

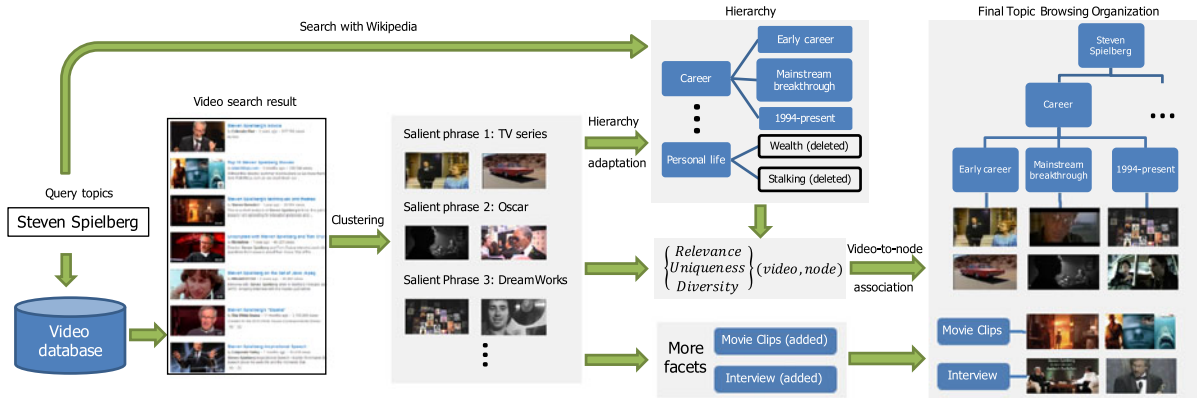


Fig. 1. Illustration of our approach. Given a query topic, we first retrieve relevant videos using online video search engines and download a hierarchy for the topic from Wikipedia. The hierarchy is then adapted by analyzing the retrieved videos, and more topic facets are mined from the results to complement the topic hierarchy. Finally, videos are assigned to the nodes of the hierarchy by performing optimization w.r.t. three criteria. See texts for more explanations.

two-step method. First, we remove the redundant nodes from the prototype hierarchies that lack the accordant describing videos. Second, additional clusters of videos are created solely from video information to complement the prototype hierarchies. In order not to impair the semantic cohesiveness of the hierarchies, the additional video clusters are listed apart from the hierarchies instead of directly adding into them.

To assign a good set of videos to each node of the hierarchy, we introduce an optimization framework to integrate three critical criteria: relevance, uniqueness and diversity. Relevancy measures how a video is semantically related to a specific node, while uniqueness reflects to what extent a video is focused on a single node (not to other nodes), which is preferred in video selection. The last criterion diversity is introduced to reduce content redundancy in videos. Both textual and visual information of the videos are exploited in the process of video-to-node matching. Especially, near-duplicate video search is used to improve the diversity of the results by removing visually similar videos.

Fig. 1 illustrates the framework of our approach. First, we discover the hidden topic structure of a query topic by clustering the textual information of the retrieved videos to get a set of salient phrases. Next, the salient phrases are assigned to the leaf nodes of the prototype hierarchy extracted from the corresponding Wikipedia page, and the redundant nodes with no phrases attached are pruned from the hierarchy. After that, additional topic facets are further mined from video information to complement the topic hierarchy. Finally, an optimization problem is formulated to select videos for each node, considering the three criteria of relevance, uniqueness and diversity. This work extends upon a previous conference version [2] with a new strategy of adding nodes/facets, new experiments and analysis on additional query topics, a demo system to show the user interactions with our interface, and extra discussions throughout the paper.

The rest of the paper is organized as follows. Section II reviews the related works. Section III elaborates the techniques involved in our approach. Experimental validations are discussed in Section IV and the visualization interface of our demo system is briefly introduced in Section V. Finally, Section VI concludes the paper.

II. Related Works

Many researchers have used the method of classification to organize Web search results into more user-friendly visualization interfaces, especially in the text search domain. In order to achieve a better classification result, some textual metadata was incorporated into the traditional classifiers, such as hyper-link relationships embedded in Web pages [3], [4] and user annotated tags associated with the Web pages [5]. Besides flat classification, hierarchical document classifiers have also been built to organize documents into a tree structure [6]–[8]. However, a major disadvantage of the classification methods is that training multiple-category classifiers requires a large number of labeled samples. In the domain of image and video search, clustering is a more popular method. By grouping visually similar samples together, a better visualization of search results may be achieved [9]. Similarly, clustering can also be divided into flat clustering [10]–[12] and hierarchical clustering [13]–[15]. In [16], the authors attempted to make the image clusters more discriminative and diverse, and only one representative image was selected for each image cluster to form a diverse result set.

As aforementioned, building a user-friendly hierarchical structure from scratch in a purely data-driven manner [17] is not very practical since the results can be semantically uninterpretable. One method to address this problem is to incorporate off-the-shelf knowledge such as WordNet [18], [19]. As a high-quality online encyclopedia created by crowdsourcing, Wikipedia is perhaps the most popular knowledge source. In [20], Wikipedia was used to label document clusters. Especially, candidate phrases were first extracted from Wikipedia pages and then assigned to clusters based on textual similarity. However, this method is not reliable if the initial clustering results are already of low quality. Apart from just providing labels, Hu *et al.* [21] mined concepts and category information from Wikipedia to enrich document representation for clustering. Furthermore, Ming *et al.* [1] directly used Wikipedia hierarchies to cluster

and organize Web collections such as Yahoo! Answers. In this work, predefined prototype hierarchies were used as the basis to associate related Web contents. The inputs of the method are a semantic hierarchy extracted from Wikipedia and a set of question-answer pages, and the output is a hierarchical structure with each node assigned with related pages. Due to the high-quality of the prototype hierarchies, the output organization structure can be very helpful for Web content navigation. In our work, we extend this idea to organize and visualize video search results. Since videos contain multi-modal information (e.g., textual and visual cues), how to analyze and utilize the information to arrive at a good visualization interface is the main focus of this paper.

Two recent works most related to ours are [22] and [2], both of which assigned videos to Wikipedia hierarchies for visualization. However, this work has two main differences. First, we try to process four kinds of popular queries that have multiple facets of related contents (i.e., celebrities, cities, companies and events), while the work of [22] only adopted news videos which have clear semantic structures and [2] tested the approach on queries of celebrities and cities. Second, since the Wikipedia hierarchies are not intended to be used for organizing video contents, we spend significant efforts in adapting the hierarchies to reflect the topic structures of videos, including removing redundant nodes and adding more complement nodes. In contrast, there is no hierarchy adaptation process in [22] and [2] only applied node removal.

Research on near-duplicate video search is also weakly related to this work. Near-duplicate videos are commonly seen on the Internet. According to the statistics in [23], on average there are 27% redundant videos in the search results of their set of queries. Thus, many related applications have been developed based on this technique. For example, near-duplicate video search has been used to mine event structures from Web videos [24], [25], to create storyline from news videos [26], and to track visual memes in social network videos [27]. In our work, this technique is used to reduce content redundancy in the video search results.

III. Hierarchical Result Visualization

Our hierarchical visualization method organizes the video search results into a hierarchical form for efficient browsing. The semantic hierarchies extracted from the Wikipedia pages are used as the prototype hierarchy. A Wikipedia page is organized as a hierarchy with nodes covering different facets of a topic. The description texts attaching to each node are used as the "representative documents" of each topic facet. We adopt YouTube as the basic video search engine to gather initial candidate videos, so each video has its own multi-modal information: video titles and description texts as textual cues and the video itself as visual cue. All the aforementioned information is used to construct the video hierarchy.

In this section, we first introduce the way we compute the textual and visual similarities, which is a core component of the entire framework. After that, the prototype hierarchy adaptation and the video-to-node association methods are elaborated. Finally, we introduce the method for mining more topic facets.

A. Textual and Visual Similarities

Both the nodes in the Wikipedia hierarchy and the videos have surrounding texts, so their relevance can be estimated through textual matching. First, we remove the nodes intending for references in the end of the semantic hierarchies, which include "Notes", "References", "See also", "Further reading" and "External links". After that, the texts of both nodes and videos are pre-processed by removing stop-words. The classical vector space model is used to represent texts using the tf-idf weighting scheme. The similarities between the nodes and the videos are calculated with the Cosine similarity. Note that the single-word based vector space model is not perfect since meaningful phrases are destroyed. Fortunately, many important phrases or words are themselves Wikipedia entries and are embedded with hyper-links on Wikipedia pages. These phrases with hyper-links (i.e., links to other Wikipedia entries) are extracted and treated as single units in the tokenization process. In order to emphasize the importance of the phrases, they are empirically assigned higher weights by multiplying with a constant factor of 5 in the bag-of-words representation.

The textual similarity is used between the nodes and the videos, while the visual similarity is only adopted among the videos for redundant information removal, since the visual contents on Wikipedia pages (e.g., the embedded images) are too few to be used. To compute the visual similarity between a pair of videos, we adopt a state-of-the-art system of near-duplicate video search. We first uniformly sample keyframes from the videos and extract the SIFT [28] local features. Similar to the text representation, a bag-of-visual-words (BoV) model is then employed to quantize the SIFT features points into a frame-level feature vector. We use hierarchical k-means clustering to construct a visual dictionary of size 50 k. The inverted index is then used to efficiently retrieve similar keyframes. To alleviate the quantization loss in BoW, Hamming embedding and weak geometric verification [29] are utilized to achieve a better keyframe matching performance. Finally, we use a temporal network method [30] to align frame sequences. The outputs of the near-duplicate video search component are the timestamps of the matched video segment pairs, from which the visual similarity between the input video pair can be easily computed based on the proportion of their shared near-duplicate segments.

B. Prototype Hierarchy Adaptation

The prototype hierarchies extracted from Wikipedia pages are constructed to cover textual descriptions of the topic, so there may be some nodes unsuitable for video assignment. In order to remove the redundant nodes, the hidden facets in the video search results need to be discovered first. We use the clustering method in [31] to reach this goal by formulating the clustering process into a salient phrase ranking problem. Given a set of candidate phrases, five textual properties are calculated to model the likelihood of each phrase as a topic facet. The textual properties include tf-idf, phrase length, intra-cluster similarity,

cluster entropy and phrase contextual entropy. The first two properties are common parameters associated with phrases, and next two properties involve the video cluster attached with each phrase. The video cluster of a candidate phrase is defined as all videos with that phrase appearing in surrounding texts. The intra-cluster similarity is to measure if the video cluster is close enough, and the cluster entropy is to measure the uniqueness of the video cluster. The final property of contextual entropy is used to ensure the independence of the phrases. After that, a linear regression model is then trained to combine all five properties into a single salient score by following the training method in [31]. More specifically, three annotators were asked to label positive and negative phrases as the ground-truth, and the target value is the salient score. Please see [31] for more details. The top-ranked phrases are then defined as salient phrases to represent the topic facets. Finally, video cluster pairs with an overlapping rate (i.e., the percentage of shared videos between the pair) higher than 50% are merged together as a single cluster.

The following procedure is used to generate the candidate phrases. The n-gram ($n \leq 3$) phrases are first extracted from the video titles. Then, the mutual information (MI) [32] between phrases and video texts is used to filter noisy phrases. The MI value quantifies how informative each phrase is to all the videos, which is computed as

$$MI(t) = \sum_{v} p(v,t) \log \frac{p(v,t)}{p(v)p(t)}$$

in which t denotes a phase and v stands for a video. The probability term in the equation is derived from phrase co-occurrence statistics among the video texts. The remaining phrases with MI value larger than 0.3 are used as the candidates for salient phrase selection.

The prototype hierarchy adaptation is performed by assigning video clusters to the leaf nodes of the hierarchy. It is achieved based on cluster-to-node similarity which is simply defined based on the textual similarity in Section III-A as

$$Sim_t(C,n) = \frac{1}{|C|} \sum_{v \in C} Sim_t(v,n)$$

where $Sim_t(v,n)$ is the video-to-node similarity and the subscript t indicates that the similarity is based on texts. After all the clusters are assigned to the leaf node with the highest similarity score, the nodes with no video clusters assigned are deemed as redundant ones and removed from the hierarchy. Redundant nodes have no sufficient related videos to support a topic facet, thus the removal of them will significantly improve the quality of the semantic hierarchy.

C. Video-to-Node Association

Selecting the best videos for each node of the adapted hierarchy is directly related to the users' browsing experience. For this we adopt a similar method to that in [22], where an optimization problem is formulated based on three criteria: relevance, uniqueness and diversity. The criteria are all defined between videos and nodes. Note that this is slightly different from the definitions in [22], where the uniqueness is a standalone metric only related to videos. Specifically, the relevance concerns the semantic relatedness between videos and nodes; the uniqueness is used to highlight that videos related to only one node rather than to several nodes are preferred (i.e., highly focused contents); the diversity ensures that there is no content redundancy in videos (i.e., near-duplicate videos should be removed). All the criteria are defined based on textual and visual similarities as follows:

$$\begin{aligned} Rel(v,n) &= Sim_t(v,n) \\ Uniq(v,n) &= \frac{Sim_t(v,n)}{\sum_{n_j \in Leaf} Sim_t(v,n_j)} \\ Div(v,n) &= \max_{v_j \in Path} (Sim_v(v,v_j)) \end{aligned}$$

where subscript v stands for visual similarity in the above equations. Videos relevant to a child node is assumed to be related to the parent nodes, so the calculation of the uniqueness criterion is limited to videos under the leaf nodes. The $Path$ in the definition of diversity means the set of videos on the top-down browsing path, which includes the videos in the node n and the videos in all the parent nodes of n. Some videos can contain information of several topic facets, so we only remove near-duplicate videos on the top-down browsing path.

Following [22], we now define an optimization function for the video-to-node association using the three criteria as

$$\mathcal{F} = \sum_{v,n} \beta \cdot Rel(v,n) * Uniq(v,n) - (1-\beta) \cdot Div(v,n).$$

Videos are selected for each node based on the above function. For the sake of efficiency, we adopt a greedy method which inserts one video at each step node by node. Although this is not globally optimal, we find it sufficient for this task. The candidate videos are only selected from the corresponding video clusters (attached to a salient phrase) assigned to the node to further speed up the process. The video clusters are highly related to the node, so using the salient phrase clustering outcomes also helps improve the precision of the selected videos, compared with the brute-force selection from the entire collection. The entire process terminates after all the nodes are added with sufficient videos. With the videos assigned, our proposed hierarchical visualization is already realized. Next, more topic facets are mined to complement this semantic hierarchy.

D. Mining More Facets

It is not guaranteed that the Wikipedia-based hierarchy can cover all the aspects of a given topic. Therefore, in this subsection, we try to mine more facets from the video information that is not covered in the hierarchy. Since adding more nodes into the already well-organized hierarchy without damaging its semantic structure is very difficult, we only intend to identify some omitted yet highly important facets to complement the hierarchical visualization. In other words, the mined nodes are placed besides the hierarchy for the users to browse.

Technically, we use the salient phrases and their video clusters in the process of hierarchy adaptation as the candidate facets. Then, for a given video cluster C and a hierarchy node n, the

TABLE I
STATISTICS OF THE COLLECTED DATASET. THE FOUR GROUPS OF TOPICS (CELEBRITIES, CITIES, COMPANIES, AND EVENTS) ARE LISTED FROM TOP TO BOTTOM. THE NUMBERS IN THE PARENTHESES ARE THE AMOUNT OF LEAF NODES IN THE CORRESPONDING WIKIPEDIA HIERARCHIES

ID	Topic name	#Videos	#Nodes (leaf)	Wikipedia URL
1	Barack Obama	1887	31 (22)	en.wikipedia.org/wiki/Barack_Obama
2	Steven Spielberg	1618	23 (20)	en.wikipedia.org/wiki/Steven_Spielberg
3	Michael Jackson	1404	25 (22)	en.wikipedia.org/wiki/Michael_Jackson
4	Michael Schumacher	1371	26 (19)	en.wikipedia.org/wiki/Michael_Schumache
5	Steve Jobs	1870	31 (24)	en.wikipedia.org/wiki/Steve_Jobs
6	London	2144	41 (33)	en.wikipedia.org/wiki/London
7	Chicago	2020	36 (28)	en.wikipedia.org/wiki/Chicago
8	Los Angeles	2020	27 (20)	en.wikipedia.org/wiki/Los_Angeles
9	New York City	1834	31 (23)	en.wikipedia.org/wiki/New_York_City
10	Paris	1936	45 (35)	en.wikipedia.org/wiki/Paris
11	Apple	1872	32 (26)	en.wikipedia.org/wiki/Apple_Inc.
12	Google	1474	20 (16)	en.wikipedia.org/wiki/Google
13	Coca cola	2029	28 (21)	en.wikipedia.org/wiki/Coca-Cola
14	BMW	1745	39 (32)	en.wikipedia.org/wiki/BMW
15	Monsanto	1844	57 (40)	en.wikipedia.org/wiki/Monsanto
16	2012 Summer Olympics	3070	30 (26)	en.wikipedia.org/wiki/2012_Summer_Olympics
17	Fukushima Daichi Nuclear Disaster	2032	31 (23)	en.wikipedia.org/wiki/Fukushima_Daiichi_nuclear_disaster
18	Manila Hostage Crisis	1656	19 (16)	en.wikipedia.org/wiki/Manila_hostage_crisis
19	Shooting of Trayvon Martin	2140	35 (27)	en.wikipedia.org/wiki/Shooting_of_Trayvon_Martin
20	Edward Snowden Event	2163	22 (15)	en.wikipedia.org/wiki/Edward_Snowden

cluster-to-node similarity $Sim_t(C, n)$ is calculated based on the aforementioned text matching method. After that, the standalone cluster similarity to the whole hierarchy is defined as follows:

$$Sim_t(C) = \max_{n \in H} (Sim_t(C, n))$$

where H is the node set of the hierarchy. A suitable similarity range is decided to pick the complement facets, which is discussed later in the experiments. The video clusters accompanying the mined facets are orderless, so we need to rank the video clusters to choose the representative videos for each facet. The optimization formulation in the above section is not suitable due to the lack of description text for the mined facets. In order to overcome this obstacle, we train a Latent Semantic Indexing (LSI) model from the surrounding texts of all the videos for each topic. By mapping the salient phrase and video texts to a common semantic space, video-to-facet proximity is calculated using Cosine similarity and the videos are ranked according to their similarity scores. Only the top-ranked videos are displayed in the interface.

IV. EXPERIMENTS

In this section, we start from introducing our dataset and parameter settings, followed by detailed discussions of quantitative evaluations, subjective user studies, as well as computational efficiency.

A. Dataset

A total number of 20 query topics are adopted in this study, covering four very popular search query groups: celebrities, cities, companies and events. We use these queries to search on YouTube and download the top-ranked videos and their surrounding texts. There are more than 1,300 videos for each topic in the final dataset. We also use the queries to search on Wikipedia to extract their corresponding semantic hierarchies, the texts under each node of the hierarchies, and the terms with hyper-links on the Wikipedia pages. The XML dump of Wikipedia is more uniformly organized than the HTML Web pages, so all the Wikipedia-related data are obtained by parsing

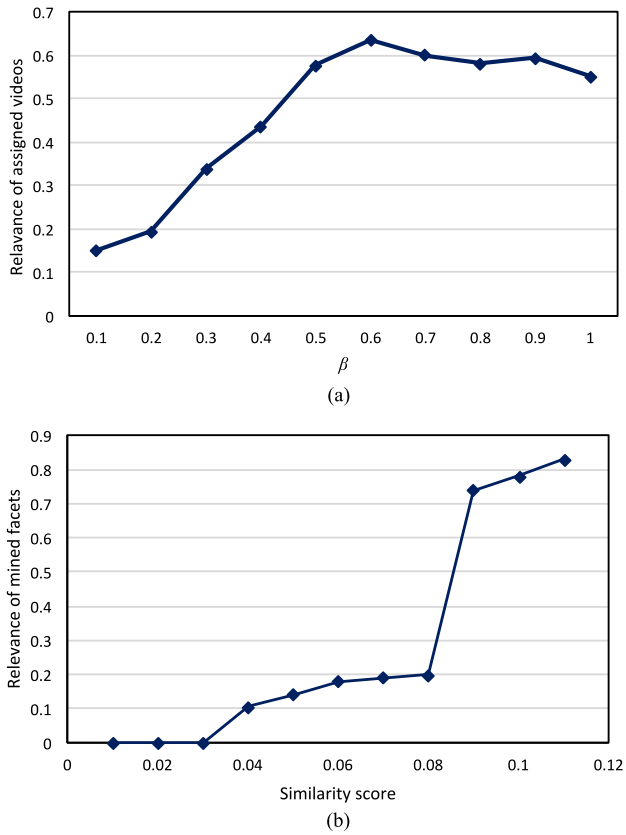


Fig. 2. Impact of two key parameters (a) β and (b) similarity score, evaluated on a separate validation set. See text for discussions.

TABLE II
RESULTS OF QUANTITATIVE EVALUATIONS. "PROTOTYPE" AND "ADAPTED" IN THE TABLE REPRESENT THE EVALUATION RESULTS ON THE ORIGINAL PROTOTYPE HIERARCHY AND THE ADAPTED HIERARCHY, RESPECTIVELY. NOTE THAT, FOR DIVERSITY, A LOWER VALUE IS BETTER

ID	Salient Phrase	Video-to-Node Association						Facet Mining	
	Precision (%)	Relevance (%)		Uniqueness (%)		Diversity (%)		Relevance (%)	
		Prototype	Adapted	Prototype	Adapted	Prototype	Adapted		Recall (%)
1	78	41	70	41	47	5.8	1.0	60	68
2	55	39	59	36	42	8.7	1.7	50	65
3	50	31	51	27	31	10.4	1.6	66	62
4	46	15	42	32	44	7.7	0.7	60	58
5	61	21	45	29	41	6.5	0.0	60	61
6	67	33	58	35	56	5.4	0.0	70	67
7	78	44	53	39	45	4.4	0.0	60	62
8	82	43	61	53	65	1.5	0.0	60	77
9	70	39	51	36	40	2.0	0.6	53	71
10	43	26	35	29	34	0.9	1.1	80	51
11	72	58	63	41	50	8.5	0.4	50	51
12	46	52	66	50	62	8.7	0.0	70	70
13	52	60	64	42	57	5.0	1.1	50	60
14	54	36	52	27	37	4.2	0.6	55	74
15	70	45	60	28	32	12.6	1.5	60	57
16	64	50	66	32	36	0.8	0.0	50	43
17	56	49	54	40	55	10.5	0.0	60	73
18	46	36	50	39	46	4.1	0.5	50	60
19	52	52	65	31	34	6.8	0.4	52	71
20	68	60	70	24	36	12	0.3	60	69
Mean	61	39	54	36	45	6.3	0.6	59	64

the XML files. Table I summarizes the dataset, including the topic names, the number of videos per topic, the URL of the Wikipedia entries and the numbers of nodes in the semantic hierarchies.

Manual labels are needed to perform quantitative evaluations, which include the relevance of the salient phrases to the nodes in the hierarchies, the relevance of the assigned videos to the nodes, and the relevance of the assigned videos to the mined complement facets. Annotating the data is a very time-consuming process. We hired several annotators to perform the task. All the annotators were asked to carefully read the Wikipedia pages beforehand to ensure a good quality of the labels.

B. Parameter Settings

Five additional topics are chosen as a validation set to tune some key parameters in our approach. The trade-off parameter of β in the optimization function is set based on the precision of the assigned videos for the validation queries. Based on the results in Fig. 2(a), we set β as 0.6 and fix this value in all the experiments.

Another important parameter is the similarity range used in the mining of the additional facets. A too small similarity score indicates that the facet is not related to the topic, while a too large score implies that the facet is probably already covered in the original hierarchy. We adjust the similarity score and manually annotate the relevance of the salient phrases of the mined facets. As shown in Fig. 2(b), there are no related phrases lower than 0.03 and too many potentially overlapping phrases higher than 0.08. Therefore, the range of 0.03–0.08 is adopted.

Other parameters like the MI value and the weight of phrases in the bag-of-words representation are set in a similar fashion, whose values have been given in early discussions of this paper.

C. Quantitative Evaluation

We quantitatively evaluate the performance of the following three critical parts in our approach: salient phrase selection, video-to-node association and complement facet mining. After that, the overall recall of selected videos is also evaluated and discussed.

1) Salient Phrase Selection: The salient phrases are used to adapt the prototype hierarchy by assigning them to related nodes, so we evaluate the precision of phrase-to-node assignment (i.e., the percentage of the correctly assigned phrases). The results are summarized in the second column of Table II, where we can see that the precision varies across topics. This metric depends highly on the quality of the retrieved videos from YouTube. If the retrieved videos are not very relevant to the whole semantic hierarchy, the precision of the assigned salient phrases will be low. Likewise, the whole quality of the retrieved videos can also influence the precision of the assigned videos in the hierarchy, as will be discussed in the following subsection, which is very easy to understand. Overall, with a mean precision of 61%, the selected salient phrases are fairly accurate.

2) Video-to-Node Association: As elaborated earlier, the video-to-node association is performed by optimization on three criteria: relevance, uniqueness and diversity. We therefore design the following three metrics to evaluate the performance separately.

TABLE III
SALIENT PHRASES OF ALL THE MINED FACETS FOR EACH SEARCH TOPIC

Topic	Salient Phrases of the Mined Facets
Barack Obama	Human rights, Climate change
Steven Spielberg	Movie clip, Interview
Michael Jackson	Michael Jackson molestation, Honor, Legacy, Tribute
Michael Schumacher	Michael Schumacher interview, Michael Schumacher tribute, Michael Schumacher funeral
Steve Jobs	Apple computers, Teen Choice Awards
London	London streets, London fashion week
Chicago	Social media, Chicago tourism, Chicago tribune, Chicago Teacher Union
Los Angeles	California Science Center
New York City	Mexico City, Law, New York accent
Paris	Paris street, Disneyland Paris, Eiffel Tower
Apple	Apple success, Apple valley
Google	Google Maps, Wall Street, Google Glass, Google Chrome, Google Finance
Coca cola	Coca Cola logo, Earth hour
BMW	BMW Group, Top Gear, Performance
Monsanto	Food supply, GMO food
2012 Summer Olympics	Drugs test, Track field
Fukushima Daichi Nuclear Disaster	Workers, World concerns
Manila Hostage Crisis	Hostage kill, Manila hijack, Police force, Manila hostage drama, Metro Manila
Shooting of Trayvon Martin	Defense attorneys, Social media, Charges Zimmerman, Geroge Zimmerman attorneys
Edward Snowden Event	Human rights, Surveillance program

Relevance: The precision of assigned video is used to measure the relevance criterion.

Uniqueness: The average of uniqueness value of the videos, i.e., the mean value of $Uniq(v, n)$, under the leaf nodes is computed.

Diversity: Diversity is measured by the percentage of the selected videos that are redundant. If the visual similarity between two videos is over 0.5, one of them is then determined to be redundant. Lower value means better performance.

In the video-to-node association, at most $2 \times \#Salient\ Phrases$ videos are selected for the leaf nodes. Since the salient phrases are only assigned to the leaf nodes, we select at most 5 videos for the non-leaf nodes. A leaf node with more salient phrases is semantically more diversified, so more videos should be picked to cover the semantic diversity. The video-to-node association is performed on both the prototype hierarchy and the adapted hierarchy to compare their performances, using the same parameters in the two methods.

The results are listed in columns 3–8 in Table II. We can see that, compared with the prototype hierarchy, the results of all the three criteria on the adapted hierarchy are largely improved. The significant increase of relevance is due to the fact that the redundant nodes in the hierarchy are pruned in adaptation. For example, in the semantic hierarchy of the topic "London", unsuitable nodes with few related videos like "Toponymy", "Prehistory and antiquity" have been removed, while the nodes with sufficient support videos like "Accent" and "Leisure and entertainment" are preserved. It is a bit out of our expectation that the adaptation of the hierarchy also improves significantly on uniqueness and diversity. This indicates that the adapted hierarchy is highly desired to visualize video search results.

By analyzing the results of different groups of queries, an interesting conclusion can be drawn. The five topics in the city group contain more semantic facets than that of the other query topics. From the results we can see that these queries have fewer near-duplicate videos (better diversity values). In contrast, all the rest three groups of queries contain a topic with a diversity value of higher than 10%. The event topic "2012 Summer Olympics" has a broad set of semantic facets, and hence a very low diversity value. In addition, the city group topics have lower average relevance values for a similar reason that it is difficult to assign relevant videos for a topic with scattered contents.

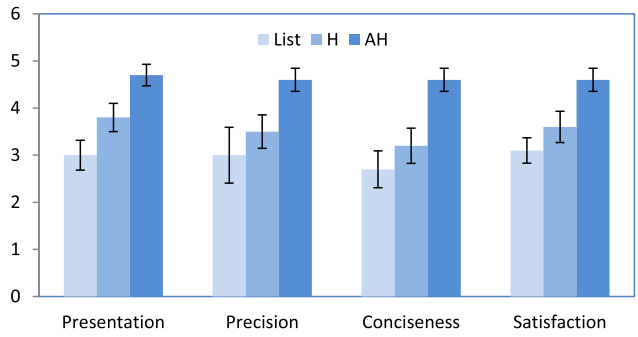


Fig. 3. Subjective evaluation scores of the List, H, and AH methods, ranging from 1 (worst) to 5 (best). Our AH achieves the highest scores.

3) Complement Facet Mining: We calculate the relevance of the selected videos for the complement facets. Same with the nodes in the hierarchy, at most 5 videos are selected for each facet. The evaluation results of the mined facets are given in column 9 of Table II. We can see that the relevance value varies across topics. Although there are no description texts for the mined facets, we still achieve a mean relevance of 59% by leveraging the LSI model, which is slightly better than the relevance score of videos in adapted hierarchy (column 4 in Table II).

Table III further shows all the salient phrases of the mined facets for each topic. The phrases are easy to be interpreted semantically and are highly related to their corresponding topics. It is interesting to see that some phrases are highly relevant, e.g.,

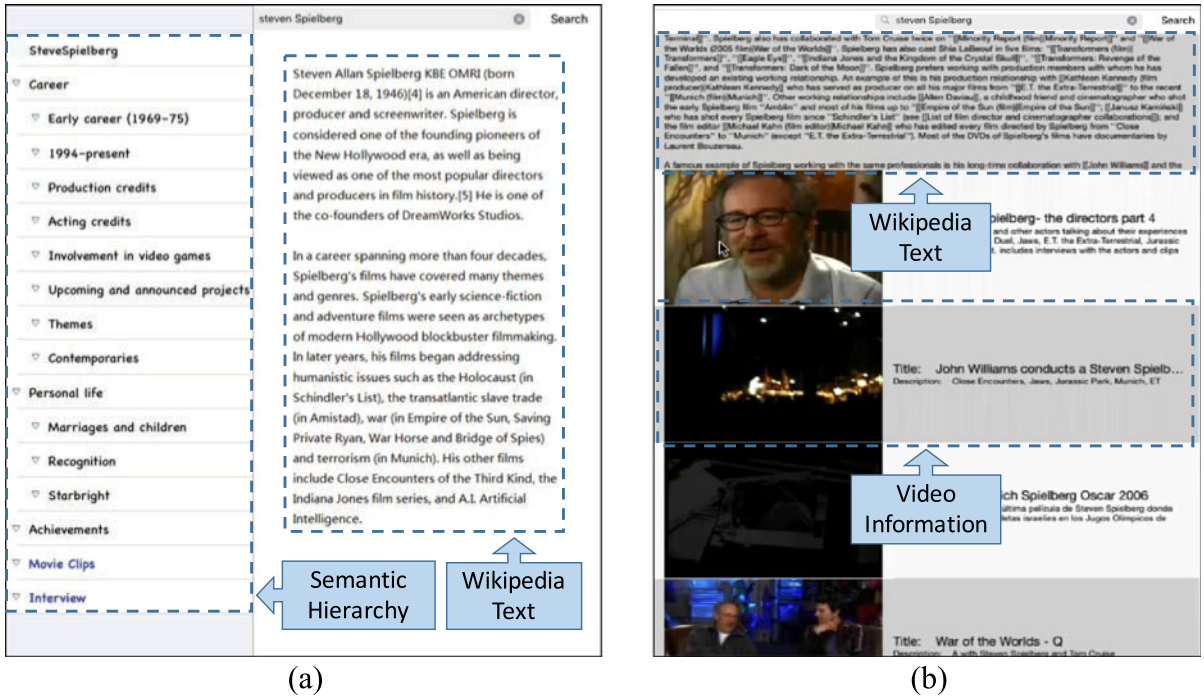


Fig. 4. Demo interfaces: (a) a hierarchical browsing interface and (b) a video browsing interface.

"Food Supply" and "GMO food" in the topic "Monsanto", and "Human rights" and "Surveillance program" in the topic "Edward Snowden Event". By browsing these complement facets, users can obtain some important information that is not covered by the semantic hierarchy.

4) Recall of the Selected Videos: Finally, we calculate the recall of all the selected videos for each topic. To evaluate recall, all the videos are annotated with respect to not only the nodes in the hierarchy but also the additionally mined facets. The results are listed in the last column of Table II, where we can see that many videos are not selected possibly due to the lack of suitable textual descriptions or the existence of repeated visual contents. Nonetheless, our system still captures around 60% of the related videos, which are sufficient for most information retrieval tasks where precision/relevance is generally more important.

D. User Study

We hired ten students to perform a subjective user study. These students were asked to answer the following four questions (one for each topic group) using three kinds of browsing methods. The questions are: "What is the biggest achievement of Michael Schumacher in his F1 career?"; "Where do you want to visit when traveling in Chicago?"; "Why Monsanto was involved in many legal cases?"; "What are China's responses to the Manila Hostage Crisis?". The students were selected on the condition that they are not familiar with the topics about which they are going to answer. Considering browsing efficiency, only five minutes are allowed to answer one question using one method, so the entire study is carried out in 60 minutes ($3 \times 4 \times 5$). The order of the questions and the browsing methods were provided randomly to minimize bias. The three methods are as follows:

List: Videos are visualized in a ranked list.

H: Videos are visualized in the original hierarchy.

AH: Videos are visualized in the adapted hierarchy with complement nodes by our approach.

We directly use the YouTube search results as the *List* method. For *H* and *AH*, a demo system is used in the user study, which will be introduced in the next section. After answering the questions, the students were asked to rate the three methods in the scale of one to five stars in the following four aspects:

Presentation: Is the organization of the videos useful in quickly locating the needed information?

Precision: To what extent are the displayed videos related to the topic?

Conciseness: Are there repeated contents (i.e., near-duplicate videos) in the results?

Satisfaction: The overall satisfaction of the video search result visualization.

Results of the user study are shown in Fig. 3. The traditional list-based system has the lowest score mainly because mixing videos of all the facets creates difficulties in quickly understanding the topic. Our approach *AH* achieves the best scores in all four aspects. For presentation and precision, the high ratings are because the adaptation of the hierarchy makes it more focused for video organization, while in the meantime improves the precision of the assigned videos (see Table II). The conciseness of *AH* is higher because of the near-duplicate removal function. Only 0.6% of videos are redundant as reported in Table II. Based on its good performance in all the first three aspects, the best overall satisfaction is also achieved by the *AH* method.

E. Computation Efficiency

The experiments are all carried out on an Intel Xeon E5-2690 3.00 GHz CPU. The computation can be split into two kinds

of operations: textual and visual. The textual part includes the indexing of videos and Wikipedia texts, the computation of the textual criteria, and the selection of videos via optimization. The whole process takes around 30 seconds for each topic on average using one thread of the CPU. The visual part includes the detection of near duplicate videos, which takes around 8 hours for one topic. The method of near duplicate detection is based matching of local visual features, so it takes a very long time on feature extraction and matching. We realize that there are more efficient solutions for the visual part [33] and will improve this part in our future work.

V. Demo Interface

We have developed a demo system to evaluate our proposed method.[1] The system is implemented as an app for iPad using Objective-C in the IDE Xcode, which can be easily extended to other mobile platforms. The visualization interface mainly consists of two parts: a hierarchical browsing interface and a video browsing interface. The former is to browse the semantic hierarchy and the additional nodes so that the users may identify their interested facets of the topic, after which the latter is used to browse videos under the selected nodes.

As shown in Fig. 4(a), after typing in a query topic in the top search box and clicking on "Search", a semantic hierarchy is shown to the users with arrows indicating the structure of the nodes. The additional facets out of the hierarchy are listed below in a different color. Brief introductory texts from the corresponding Wikipedia page are also presented to offer some background information of the search topic. The users are free to expand or collapse the node levels. After clicking on an interested node, the interface of video browsing is presented, as shown in Fig. 4(b). The video browsing interface contains both the Wikipedia texts of the selected node and the related videos. The Wikipedia texts of the chosen node is presented in the top area which can be hidden when scrolling upwards. After clicking on the video thumbnail, a video player will be invoked to play the selected video. By swiping in from the left edge of the screen, the hierarchical browsing interface will reappear for selecting other nodes.

VI. Conclusion

We have introduced a hierarchical visualization approach for video search result browsing, where videos are associated to an adapted Wikipedia hierarchy. Useful topic facets that are not covered by the Wikipedia hierarchy are also mined automatically to ensure a good coverage of the search results. Experiments on a large YouTube video dataset have clearly validated the effectiveness of our approach. In particular, the result comparison between the prototype hierarchies and the adapted hierarchies indicated that our hierarchy adaptation method is helpful for removing the nodes unsuitable for video search. The results also indicated that the additionally mined facets are of similar quality to the nodes in the hierarchies. The user studies conducted on our interaction interfaces have confirmed the superiority of our approach over the traditional list-based browsing method. By deploying this novel visualization approach for queries with complex topic structures, we believe that the user satisfaction of a video search engine can be greatly improved.

[1]Demo video available at:http://bigvid.fudan.edu.cn/demo/VideoSearchDemo.mov

References

[1] Z.-Y. Ming, K. Wang, and T.-S. Chua, "Prototype hierarchy based clustering for the categorization and navigation of web collections," in *Proc. Prototype Hierarchy Based Clustering Categorization Navigation Web Collections*, 2010, pp. 845–848.

[2] J. Wang, Y.-G. Jiang, Q. Wang, K. Yang, and C.-W. Ngo, "Organizing video search results to adapted semantic hierarchies for topic-based browsing," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 845–848.

[3] P. Calado *et al.*, "Combining link-based and content-based methods for web document classification," in *Proc. ACM 12th Int. Conf. Inf. Knowl. Manage.*, 2003, pp. 394–401.

[4] Y.-H. Kuo and M.-H. Wong, "Web document classification based on hyperlinks and document semantics," in *Proc. Pacific Rim Int. Conf. Artif. Intell.*, 2000.

[5] M. G. Noll and C. Meinel, "Exploring social annotations for web document classification," in *Proc. ACM Symp. Appl. Comput.*, 2008, pp. 2315–2320.

[6] G.-R. Xue, D. Xing, Q. Yang, and Y. Yu, "Deep classification in large-scale text hierarchies," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 619–626.

[7] C.-C. Huang, S.-L. Chuang, and L.-F. Chien, "LiveClassifier: Creating hierarchical text classifiers through web corpora," in *Proc. 13th Int. Conf. World Wide Web*, 2004, pp. 184–192.

[8] S. Dumais and H. Chen, "Hierarchical classification of web content," in *Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2000, pp. 256–263.

[9] F. Jing *et al.*, "iGroup: Web image search results clustering," in *Proc. 14th ACM Int. Conf. Multimedia*, 2006, pp. 377–384.

[10] J. Goldberger, S. Gordon, and H. Greenspan, "Unsupervised image-set clustering using an information theoretic framework," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 449–458, Feb. 2006.

[11] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, and Q.-S. Cheng, "Web image clustering by consistent utilization of visual features and surrounding texts," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 112–121.

[12] A. Hindle, J. Shao, D. Lin, J. Lu, and R. Zhang, "Clustering web video search results based on integration of multiple features," *WWW J.*, vol. 14, no. 1, pp. 53–73, 2011.

[13] S. Krishnamachari and M. Abdel-Mottaleb, "Image browsing using hierarchical clustering," in *Proc. IEEE Int. Symp. Comput. Commun.*, 1999, pp. 301–307.

[14] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen, "Hierarchical clustering of WWW image search results using visual, textual and link information," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 952–959.

[15] H. Ding, J. Liu, and H. Lu, "Hierarchical clustering-based navigation of image search results," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 741–744.

[16] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol, "Visual diversification of image search results," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 341–350.

[17] X. Zhu, Z.-Y. Ming, X. Zhu, and T.-S. Chua, "Topic hierarchy construction for the organization of multi-source user generated contents," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2013, pp. 233–242.

[18] J. Sedding and D. Kazakov, "WordNet-based text document clustering," in *Proc. Workshop Robust Methods Anal. Natural Lang. Data*, 2004, pp. 104–113.

[19] D. R. Recupero, "A new unsupervised method for document clustering by using WordNet lexical and conceptual relations," *Inf. Retrieval*, vol. 10, no. 6, pp. 563–579, 2007.

[20] D. Carmel, H. Roitman, and N. Zwerdling, "Enhancing cluster labeling using wikipedia," in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2009, pp. 139–146.

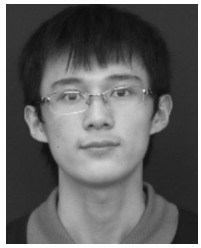
[21] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, "Exploiting Wikipedia as external knowledge for document clustering," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 389–396.

[22] S. Tan, Y.-G. Jiang, and C.-W. Ngo, "Placing videos on a semantic hierarchy for search result navigation," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 10, 2014, Art. no. 37.
[23] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Practical elimination of near-duplicates from web video search," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 218–227.
[24] X. Wu, Y.-J. Lu, Q. Peng, and C.-W. Ngo, "Mining event structures from web videos," *IEEE Multimedia*, vol. 18, no. 1, pp. 38–51, Jan.–Mar. 2011.
[25] I. Ide *et al.*, "Exploiting the chronological semantic structure in a large-scale broadcast news video archive for its efficient exploration," in *Proc. APSIPA Annu. Summit Conf.*, 2010, pp. 996–1005.
[26] X. Wu, C.-W. Ngo, and Q. Li, "Threading and autodocumenting news videos: A promising solution to rapidly browse news topics," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 59–68, Mar. 2006.
[27] L. Xie, A. Natsev, J. R. Kender, M. Hill, and J. R. Smith, "Visual memes in social media: Tracking real-world news in YouTube videos," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 53–62.
[28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[29] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 304–317.
[30] H.-K. Tan, C.-W. Ngo, R. Hong, and T.-S. Chua, "Scalable detection of partial near-duplicate videos by visual-temporal consistency," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 145–154.
[31] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma, "Learning to cluster web search results," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2004, pp. 210–217.
[32] W. H. Hsu and S.-F. Chang, "Topic tracking across broadcast news videos with visual duplicates and semantic concepts," in *Proc. IEEE Int. Conf. Image Process.*, 2006, pp. 141–144.
[33] Y.-G. Jiang and J. Wang, "Partial copy detection in videos: A benchmark and an evaluation of popular methods," *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 32–42, Mar. 2016.

Yu-Gang Jiang received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, China.

From 2008 to 2011, he was with the Department of Electrical Engineering, Columbia University, New York, NY, USA. He is currently a Professor of computer science with Fudan University, Shanghai, China. His research interests include computer vision and multimedia retrieval.

Prof. Jiang was the recipient of many awards, including Early Career Faculty Awards from Intel and China Computer Federation, the 2014 ACM China Rising Star Award, and the 2015 ACM SIGMM Rising Star Award.



Jiajun Wang received the B.Sc. degree in physics and M.Sc. degree in computer science from Fudan University, Shanghai, China, in 2013 and 2016, respectively.

He is currently a Software Engineer with Didi Chuxing, Inc. His research interests include computer vision and multimedia retrieval.

Mr. Wang was the recipient of the Best Poster Award at ACM Multimedia 2014.

Qiang Wang received the B.Sc. degree in computer science from Liaoning Normal University, Dalian, China, in 2014, and is currently working toward the M.Sc. degree at the School of Computer Science, Fudan University, Shanghai, China.

His research interests include computer vision and multimedia retrieval.

Wei Liu received the Ph.D. degree in electrical engineering from Columbia University, New York, NY, USA.

He is currently a Computer Vision Director with Tencent AI Lab, Shenzhen, China. Prior to joining Tencent, he was a Researcher with IBM T. J. Watson Research Center, Yorktown Heights, NY, USA, since 2012. His research interests include computer vision, machine learning, data mining, and information retrieval.

Dr. Liu was the recipient of the 2011–2012 Facebook Fellowship, 2013 Jury Award for best thesis of Columbia University, 2014 CVPR Young Researcher Support Award, and 2016 SIGIR Best Paper Award Honorable Mention.

Chong-Wah Ngo received the Ph.D. degree in computer science from the Hong Kong University of Science and Technology, Hong Kong, China.

He is currently a Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong, China. His recent research interests include large-scale multimedia information retrieval, video computing, and multimedia data mining.

Prof. Ngo was an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA.