2-2009

Real-time near-duplicate elimination for web video search with content and context

Xiao WU

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

Alexander G. HAUPTMANN

Real-Time Near-Duplicate Elimination for Web Video Search With Content and Context

Xiao Wu, *Student Member, IEEE*, Chong-Wah Ngo, *Member, IEEE*, Alexander G. Hauptmann, *Member, IEEE*, and Hung-Khoon Tan

***Abstract*—With the exponential growth of social media, there exist huge numbers of near-duplicate web videos, ranging from simple formatting to complex mixture of different editing effects. In addition to the abundant video content, the social web provides rich sets of context information associated with web videos, such as thumbnail image, time duration and so on. At the same time, the popularity of Web 2.0 demands for timely response to user queries. To balance the speed and accuracy aspects, in this paper, we combine the contextual information from time duration, number of views, and thumbnail images with the content analysis derived from color and local points to achieve real-time near-duplicate elimination. The results of 24 popular queries retrieved from YouTube show that the proposed approach integrating content and context can reach real-time novelty re-ranking of web videos with extremely high efficiency, where the majority of duplicates can be rapidly detected and removed from the top rankings. The speedup of the proposed approach can reach 164 times faster than the effective hierarchical method proposed in [31], with just a slight loss of performance.**

***Index Terms*—Content, context, copy detection, filtering, near-duplicates, novelty and redundancy detection, similarity measure, web video.**

I. Introduction

With the exponential growth of social media in Web 2.0, the huge volume of videos being transmitted and searched on the Internet has increased tremendously. Users can capture videos by mobile phones, video camcorders, or directly obtain videos from the web, and then distribute them again with some modifications. For example, users upload 65 000 new videos each day on video sharing website YouTube and the daily video views were over 100 million in July 2006 [30]. Among these huge volumes of videos, there exist large numbers of duplicate and near-duplicate videos.

Near-duplicate web videos are identical or approximately identical videos close to the exact duplicate of each other, which have similar time duration/length, but different in file formats, encoding parameters, photometric variations (color, lighting changes), editing operations (caption, logo and border insertion), and certain modifications (frames add/remove). A video is a duplicate of another, if it looks the same, corresponds to approximately the same scene, and does not contain new and important information. For videos having remarkably different time durations, e.g., full version and short version, by definition, they will not be regarded as near-duplicates. A user would clearly identify the videos as "essentially the same." Two videos do not have to be pixel-wise identical to be considered duplicates. A user searching for entertaining video content on the web, might not care about individual frames, but the overall content and subjective impression when filtering near-duplicate videos. Exact duplicate videos are a special case of near-duplicate videos, which are frequently returned by video search services.

Based on a sample of 24 popular queries from YouTube [37], Google Video [7] and Yahoo! Video [35], on average there are 27% redundant videos that are duplicate or nearly duplicate to the most popular version of a video in the search results [31]. For certain queries, the redundancy can be as high as 93% (see Table I). As a consequence, users are often frustrated when they need to spend significant amount of time to find the videos of interest, having to go through different versions of duplicate or near-duplicate videos streamed over the Internet before arriving at an interesting video. An ideal solution would be to return a list which not only maximizes precision with respect to the query, but also novelty (or diversity) of the query topic. To avoid getting overwhelmed by a large number of repeating copies of the same video in any search, efficient near-duplicate video detection and elimination is essential for effective search, retrieval, and browsing.

Due to the large variety of near-duplicate web videos ranging from simple formatting to complex editing, near-duplicate detection remains a challenging problem. Among existing content based approaches, many focus on the rapid identification of duplicate videos with global signatures, which are able to handle almost identical videos. However, duplicates with changes in color, lighting and editing artifacts can only be reliably detected through the use of more reliable local features. Local point based methods have demonstrated impressive performance in a wide range of vision-related tasks, and are particularly suitable for detecting near-duplicate web videos having complex variations. However, its potential is unfortunately underscored by matching and scalability issues. In our recent work [31], good performance has been achieved by combining the global signature derived from color histogram, and local point based pairwise comparison among keyframes. Coupled with the use of a sliding

Manuscript received April 15, 2008; revised October 01, 2008. Current version published January 16, 2009. This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119508). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiebo Luo.

X. Wu, C.-W. Ngo, and H.-K. Tan are with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: wuxiao@cs.cityu.edu.hk; cwngo@cs.cityu.edu.hk; hktan@cs.cityu.edu.hk).

A. G. Hauptmann is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: alex@cs.cmu.edu).

Digital Object Identifier 10.1109/TMM.2008.2009673

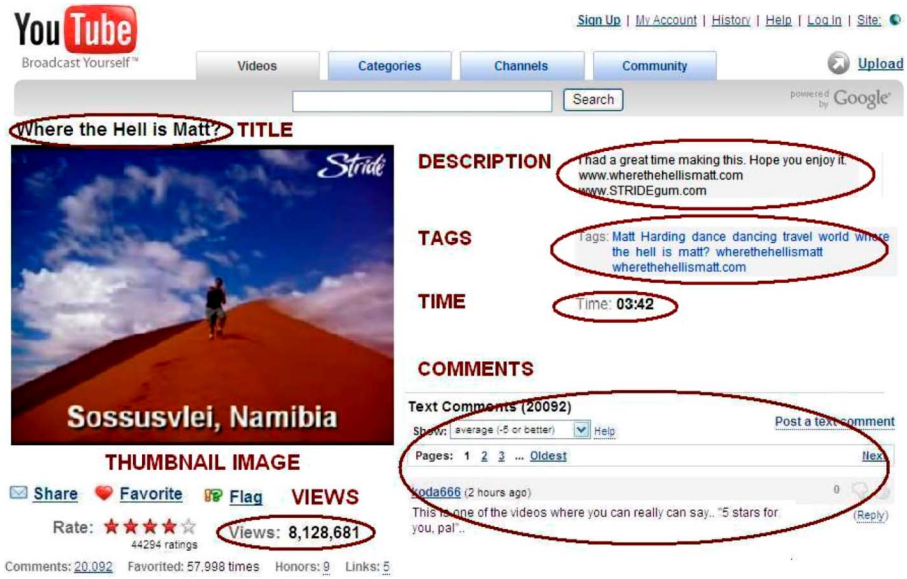

Fig. 1. Rich context information is associated with web videos, e.g., Title, Tags, Thumbnail Image, Description, Time, Views, Comments, etc.

TABLE I
TWENTY-FOUR VIDEO QUERIES COLLECTED FROM YOUTUBE, GOOGLE VIDEO AND YAHOO! VIDEO (#: NUMBER OF VIDEOS)

	Queries		Near-Duplicate	
ID	**Query**	**#**	**#**	**%**
1	The lion sleeps tonight	792	334	42 %
2	Evolution of dance	483	122	25 %
3	Fold shirt	436	183	42 %
4	Cat massage	344	161	47 %
5	Ok go here it goes again	396	89	22 %
6	Urban ninja	771	45	6 %
7	Real life Simpsons	365	154	42 %
8	Free hugs	539	37	7 %
9	Where the hell is Matt	235	23	10 %
10	U2 and green day	297	52	18 %
11	Little superstar	377	59	16 %
12	Napoleon dynamite dance	881	146	17 %
13	I will survive Jesus	416	387	93 %
14	Ronaldinho ping pong	107	72	67 %
15	White and Nerdy	1771	696	39 %
16	Korean karaoke	205	20	10 %
17	Panic at the disco I write sins not tragedies	647	201	31 %
18	Bus uncle (巴士阿叔)	488	80	16 %
19	Sony Bravia	566	202	36 %
20	Changes Tupac	194	72	37 %
21	Afternoon delight	449	54	12 %
22	Numa Gary	422	32	8 %
23	Shakira hips don't lie	1322	234	18 %
24	India driving	287	26	9 %
	Total	12790	3481	**27 %**

window scheme, considerable savings have been made to relieve the computational burden. Unfortunately, the local point based method remains expensive considering the prohibitively large number of keyframe pairs within two videos and local points within a single keyframe. The number of keyframes could exceed one hundred for a four minute music video with fast changing scenes, while the number of local points could reach hundreds to thousands per keyframe. Accurate detection generally comes at the cost of time complexity particularly in a large scale web video corpus. Timely response to user queries is one important factor that fuels the popularity of Web 2.0. Therefore, computation becomes the bottleneck for real-time near-duplicate elimination.

Fortunately, the social web provides much more than a platform for users to interact and exchange information. This has resulted in the rich sets of context information associated with web videos, such as thumbnail images, tags, titles, time durations, views, comments, and so on (see Fig. 1). These context resources provide complementary information to the video content itself, which could be exploited for improving the detection efficiency. Fig. 2 shows the actual search results for two queries ("The lion sleeps tonight" and "White and Nerdy") at YouTube. For the first query (the left part of Fig. 2), redundancy is fairly obvious, judging purely by the thumbnail images and the time duration information. However, for a music video with rich and diverse video content as illustrated by the second query "White and Nerdy" (the right part of Fig. 2), it is not easy to judge the video redundancies purely based on thumbnail images and time durations. In this example, thumbnail images fail to align even when the time difference between the two videos is minimal (as short as 1 second interval). The situation highlights the challenges of employing context information to perform near-duplicate video detection especially for videos with complex scenes. In this paper, we exploit contextual cues, integrated with content information, for the fast filtering of near-duplicate web videos. We show that the integration of content and context analysis can reach *real-time* novelty re-ranking with extremely high efficiency, in which the majority of duplicates can be swiftly detected and removed from the top rankings.

Our contribution in this paper includes the novel utilization of Web 2.0 context information to skip expensive content processing in near-duplicate video detection. While this task is traditionally achieved by pure content-based analysis, we demonstrate that near-duplicate identity is easier and cheaper to acquire, though not adequate, by context analysis. Careful cou-

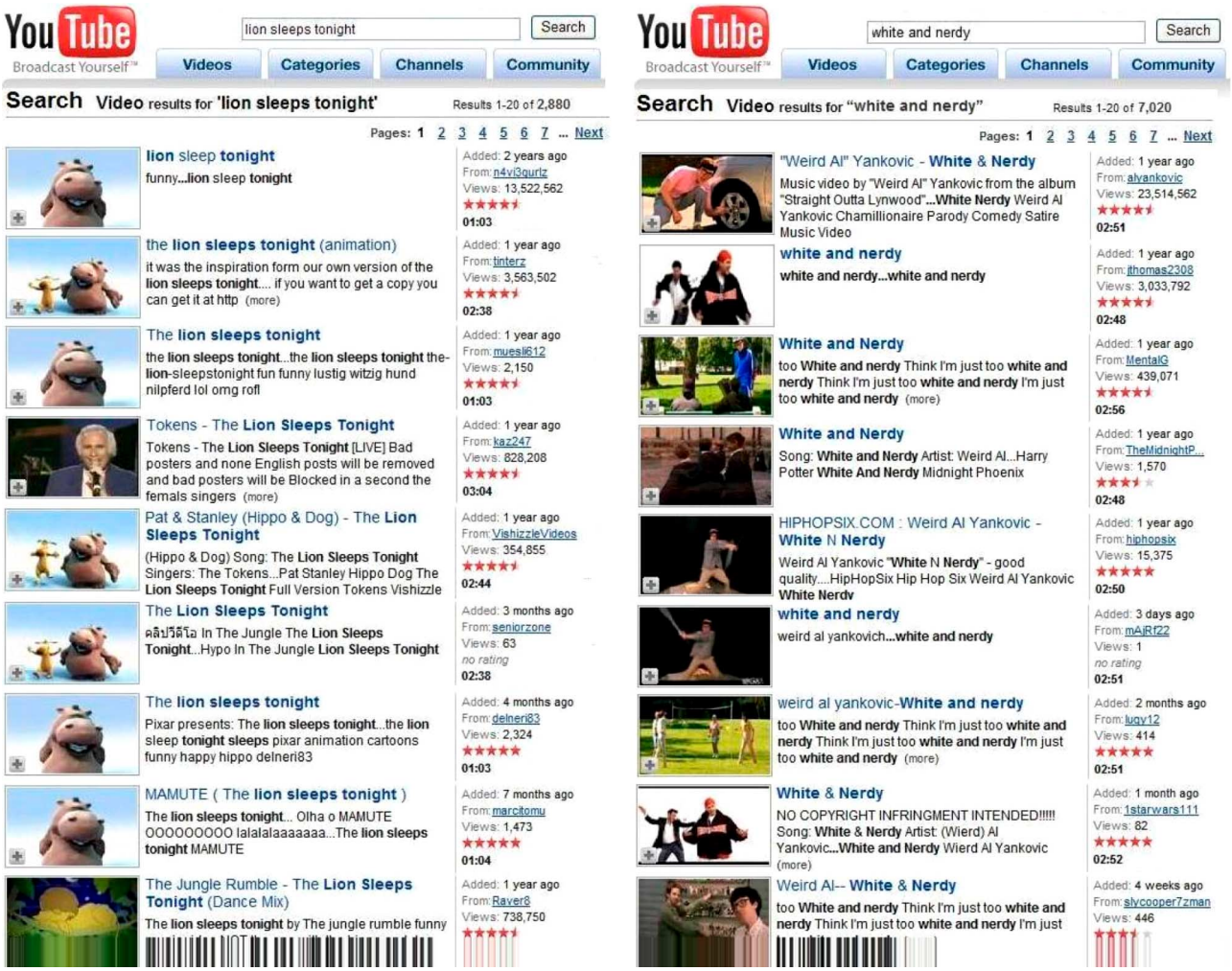

Fig. 2. Search results for the queries "The lion sleeps tonight" and "White and Nerdy" from YouTube. (a) For the first query, the redundancy is fairly obvious purely judging by the thumbnail images and the time duration information. (b) For the second query, these videos are near-duplicate except one. However, it is not easy to judge the redundancy of videos purely based on the thumbnail images and time duration.

pling of context and content is a promising way to hit a balance between speed and accuracy for practical elimination of near-duplicate web videos. To this end, a framework emphasizing the integration of context and content is also proposed. The rest of this paper is organized as follows. In Section II, we give a brief overview of related work. The proposed approach integrating content and context information for near-duplicate elimination is elaborated in Section III. Section IV presents experiments and results. Finally, we conclude the paper with a summary.

II. Related Work

A. Video Copy and Similarity Detection

Video copy and similarity detection has been actively studied for its potential in search [6], topic tracking [34] and copyright protection [16]. Various approaches using different features and matching algorithms have been proposed.

Among existing approaches, many emphasize the rapid identification of duplicate videos with compact and reliable global features. These features are generally referred to as signatures or fingerprints which summarize the global statistic of low-level features. Typical features include color, motion and ordinal signature [8], [38] and prototype-based signature [5], [6]. These global feature based approaches are suitable for identifying almost identical videos, and can detect minor editing in the spatial and temporal domain [5], [6], [8], [9], [38]. Our analysis of a diverse set of popular web videos shows that there are around 20% exact duplicate videos among all near-duplicate web videos [31]. It is common for web users to upload exact duplicate videos with minimal change. This highlights the need of an approach for fast detection of duplicate videos.

However, global features become ineffective when dealing with video copies buried in layers of editing cosmetics. For these more difficult groups, low-level features at the segment or shot level are helpful to facilitate local matching [1], [19], [22], [27], [38]. Typically the granularity of the segment-level matching, the changes in temporal order, and the insertion/deletion of frames all contribute to the similarity score of videos. Compared to signature based methods, segment level approaches are slower but capable of retrieving approximate copies that have undergone a substantial degree of editing.

At a higher level of complexity, more difficult duplicates with changes in background, color, and lighting, require even more intricate and reliable features at region-level. Features such as color, texture and shape can be extracted at the keyframe level, which in turn could be further segmented into multiple region units. However, the issue of segmentation reliability and the granularity selection brings into question the effectiveness of these approaches. Recently, local interest points (keypoints) are shown to be useful for near-duplicate and copy detection [13], [15], [16], [18], [26], [41]. Local points are salient local regions (e.g., corners) detected over image scales, which locate local regions that are tolerant to geometric and photometric variations [23]. Salient regions in each keyframe can be extracted with local point detectors (e.g., DOG [23], Hessian-Affine [25]) and their descriptors (e.g., SIFT [24]) are mostly invariant to local transformations. Keypoint based local feature detection approach (e.g., OOS [26]) avoids the shortcoming of global and segment-level features, and gives an accurate measurement even for images undergone drastic transformations. Although local points have largely been acknowledged as reliable and robust features, efficiently matching large amount of local points remains a difficult problem. Recent solutions include using indexing structure [15], [16] and fast near-duplicate tracking with heuristics [26].

Fundamentally, the task of near-duplicate detection involves the measurement of redundancy and novelty, which has been explored in text information retrieval [2], [42]. The novelty detection approaches for documents and sentences mainly focus on vector space models and statistical language models to measure the degree of novelty expressed in words. Query relevance and information novelty have been combined to re-rank the documents/pages by using Maximal Marginal Relevance [3], Affinity Graph [40] and language models [39]. Recently, multimedia based novelty/redundancy detection has also been applied to cross-lingual news video similarity measure [32] and video re-ranking [10], [11], [20] by utilizing both textual and visual modalities. To the best of our knowledge, there is little research on near-duplicate video detection and re-ranking for large scale web video search [6], [21], [31].

B. Context Analysis for Retrieval

Contextual information has been actively discussed from different viewpoints, ranging from the spatial, temporal, shape context, to pattern and topical context. Tags and locations are two commonly used context information for image retrieval [17]. Social links have attracted the attention which are used to study the user-to-user, and user-to-photo relations. The user context and social network context were also used to annotate images [29].

In some recent studies [11], [17], the fusion of content and context is a common way to improve the performance. The semantic context induced from the speech transcript surrounding a keyframe was combined with the visual keywords to improve the performance of near-duplicate image retrieval [33]. A context graph was constructed at the document level for video search reranking [11], in which context refers to the attributes describing who, where, when, what, etc, of web documents. Tags, notes, geolocations and visual data have been exploited to improve the retrieval information for Flickr [28]. In [17], tags, location and content analysis (color, texture and local points) were employed to retrieve the images of geographical related landmarks.

Most of the mentioned works are mainly based on the image sharing website Flickr. However, there has been little research exploring the context information for video sharing websites, such as YouTube. It remains unclear whether contextual resources are also effective for web videos. In particular, the integration of content and context information for near-duplicate web video elimination has not been seriously addressed.

III. Near-Duplicate Elimination with Content and Context

This section details our proposed approach for real-time near-duplicate elimination. We begin by briefly describing the available context information for web videos (III-A). The context cues are integrated with content information and ultimately lead to the proposal of our real-time near-duplicate elimination framework (III-B). The details of framework are further outlined in the remaining Sections III-C–III-E.

A. Context Cues for Web Videos

One attractive aspect of social media is the abundant amount of context metadata associated with videos as shown in Fig. 1. The metadata includes different aspects of information such as tags, time durations, titles, thumbnail images, number of views, comments, usernames, and so on.

Different from still image and text document, video describes scene through displaying a sequence of consecutive frames within a time frame. An interesting observation is that near-duplicate web videos more or less have similar *time duration*, with a difference of only a few s. As such, time duration could be a critical feature for efficient filtering of dissimilar videos. A potential risk, nevertheless, is when dealing with complex near-duplicate videos. These web videos could undergo content modification by dropping small parts of the videos or adding any arbitrary video segment at the beginning or end of the videos. This results in near-duplicate videos of different lengths which could not be dealt with if using only time duration.

Thumbnail image is another critical context metadata associated with videos. In social media, thumbnail images are simply extracted from the middle of videos. These images give users a basic impression of the video content. In most cases, it is sufficient to categorize two videos as near-duplicates if: a) their thumbnail images are near-duplicate and b) their time durations are close to each other. This dramatically reduces computation time compared to conventional content-based approaches, which involve all pairwise permutations of keyframes/shots. However, near-duplicate videos with complex variations might sample different thumbnail images due to different video length, particularly when frame insertions or deletions corrupt the time-line of the original video clips. An example illustrating the problem is shown in Fig. 2(b). This might lead to

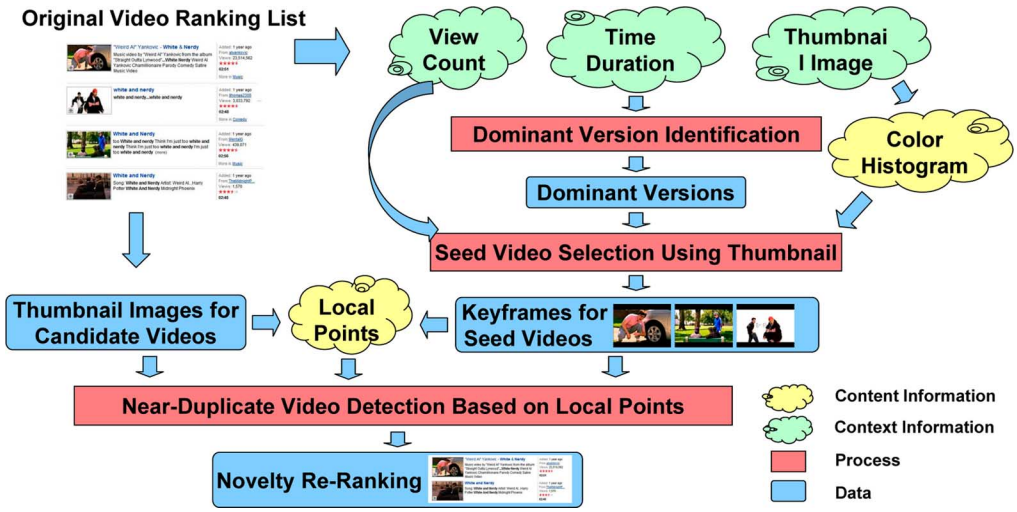

Fig. 3. Framework of the real-time near-duplicate elimination with the integration of content and context information.

incompetent near-duplicate detections. In summary, thumbnail images provide a useful but not always sufficient cue to identify duplicate videos. However, it can become effective if compensated by content information. Extracting a set of keyframes from the middle of the video could overcome the randomness encountered during thumbnail sampling.

View count is an important indication of the popularity of videos. The larger number of views indicates that it is either a relatively important video or an original video. The popularity count of a video carries the implicit information about the intention and interest of the users. We can thus identify the common interest of the public at large. Therefore, when presented with a group of potentially near-duplicate videos, view count provides clue to identify the original video sources from which other near-duplicate videos are transformed.

In addition to time duration, thumbnail image and view count, other frequently employed context features include tag and title. These features, while known to be useful for video retrieval, are not particularly helpful for near-duplicate elimination. For novelty reranking, all the returned videos are relevant to the text query, among which duplicate and non-duplicate videos are mixed together. Near-duplicate web videos may have inconsistent titles and tags, while non-near-duplicate videos could have the same ones. Although titles and tags could bring some hints of the novelty, e.g., person name and episode number, these cues are buried within the noisy text information. The *description* context contains more details, but they are not discriminative to determine the novelty and redundancy. Social-oriented context such as the *comment* context is extremely noisy and in most cases irrelevant to the video content. Generally, the user-supplied text resources are noisy, inaccurate, ambiguous, and even wrong.

Context information can accelerate the detection of near duplicates in two manners. First, context can be used jointly with content as a basis to perform near-duplicate decision. Second, context information can be used to uncover the set of dominant near-duplicate groups. In Section III-B, we explore the use of *time duration*, *thumbnail images* and *view counts* for real-time near-duplicate elimination. These contextual cues can be exploited in a complementary way. Time duration provides useful hint for rapid identification of dissimilar (novel) videos. On the other hand, thumbnail images extracted from social media reveal the near-duplicate identity of web videos. View count indicates the popularity of a web video and provides cue for the selection of source videos which being popularly modified into different versions of near-duplicate videos.

B. Proposed Framework

With the fact that contextual cue is cheaper to acquire from social media while content is expensive to process, a framework exploiting both cues is proposed as illustrated in Fig. 3. As a preprocessing step, time duration is used to rapidly but coarsely identify the preliminary groups of near-duplicate videos, where we term each group as a *dominant version*. For each dominant version, a seed video, which potentially is the original source from which other videos are derived, is selected. To be efficient, the selection is based on the color histograms of thumbnail images and their view counts. Since seed videos are potential sources, the final step of near-duplicate detection is reduced to compare thumbnail images of candidate videos to the selected seed videos. To ensure accurate detection, this step is performed by matching local points. In brief, the framework integrates context and content information, seeking a balance between real-time processing and detection accuracy. The three main processes (dominant version identification, seed video selection, near-duplicate video elimination) are briefly described as below. Further details are elaborated in Sections III-C, III-D, and III-E, respectively.

Dominant Version Identification: In an entertainment web video search, there usually exist a couple of dominant videos. Other videos in the search result are usually derived from them. For example, there are a short version and a full version in the query "The lion sleeps tonight." This observation is evident in Table I where the most popular version takes up a large portion of the search result. For each query, dominant version identification is performed by analyzing the distribution of time duration.

Seed Video Selection: Seed video selection is performed to pick one seed video for each dominant time. *Seed video* is defined as a potential source from which most near-duplicate videos derived. Videos falling within a specified time range are then filtered by matching the thumbnail images. Color histogram (CH) is extracted for every thumbnail image. Without delving into the content detail inside the videos, the color histogram of thumbnail images and the number of views are combined to select the seed videos. For each seed video, the representative keyframes are extracted from the middle part of this video. We call these representative keyframes as *prototype keyframes* since they are the potential "prototypes" from which other near-duplicates are transformed, either directly or indirectly. While each video is represented by one thumbnail image, the seed video has multiple prototype keyframes extracted from the middle part. Therefore, the matching of videos is implemented as the comparison of thumbnail images and prototype keyframes, which significantly reduces computational time.

Near-Duplicate Video Elimination: According to the original ranking from the search engine, each video is compared with the seed videos and every novel video to see whether they are duplicates. The comparison is carried out in two ways, by using context and content information respectively. First, time duration information is treated as a filter to avoid the comparison between videos with considerably different length. Second, content features based on local points extracted from the thumbnail images and prototype keyframes are matched using [26]. The first step speeds up detection by avoiding unnecessary comparison, while the second step keeps the expensive content processing as low as possible. After all, if the thumbnail image of a video is found to be a duplicate of the seed video or some previous novel videos, it will be labeled as a near-duplicate video and filtered out. Otherwise, the video is considered novel and will be returned to the users.

C. Dominant Version Identification Using Time Duration

A content-based approach for discovering the dominant groups of videos is by performing video clustering. Ideally, the clusters with larger size can be regarded as the dominant groups. However, such approach is infeasible since exhaustive pair-wise matching between videos is normally required. In contrast, the time duration context coupled with thumbnail image context can effortlessly identify the set of dominant videos with equally competitive precision. Assuming that the editing done by users does not involve significant frame additions or removals, dominant videos often fall into the same bin in the time duration. We refer to such time duration bin as the *dominant time duration*.

For each query, the time duration of all videos in this query is first collected, from which we can have a basic understanding of the time distribution and potential peaks. Fig. 4 shows the time duration distribution of "The lion sleeps tonight" (Query 1), "Sony Bravia" (Query 19) and "Free hugs" (Query 8). There might exist several dominant versions for a query. For Query 1, we can see that there are two peaks: 63 and 158 s, which are observed to be populated by videos corresponding to the two dominant versions, a short and long version of the video clip. Similarly, it is easy to tell the dominant versions for Query 19. However, there exist cases that the dominant version is not obvious, e.g., "Free hugs" (Query 8). For these categories of query, the time distribution is relatively flat, and the one with the highest frequency will be chosen as the dominant version.

However, the time durations for near-duplicate videos may not be exactly equal. The editing operations imposed by users may cause slight variation in time duration. When the time duration distribution is plotted as a histogram, videos that fall close to a certain time duration may not be distinguishable as a conspicuous peak. To increase robustness towards editing effects and quantization errors, in addition to the frequency of a single bin, neighboring bins could also be taken into account when computing the significance of a time duration. A simple strategy to consider the neighboring bins is

$$T = \begin{cases} t_i \left| \left(\sum_{j=i-d}^{j=i+d} |t_j| \right) > \theta \right. \\ t^* | \forall t_i : |t_i| \leq |t^*| \end{cases} \quad \text{otherwise}$$

where $|t_j|$ is the cardinality of the time duration bin t_j, d is the length of neighborhood window, and θ is a threshold. The window d can be set within the range of 3 to 5 s, so that it can tolerate the near-duplicate variation caused by editing. Practically, the parameter should not be large to avoid the inclusion of excessive neighboring bins. Accordingly, the setting of window d will also affect the choosing of θ. The threshold θ can be set proportional to the number of videos for each query, e.g., 5% of the number of videos. Since dominant versions usually form apparent peaks in the histogram, the dominant time durations are not difficult to pick. In case a peak is still not observed with this strategy, the duration with the highest frequency will be heuristically picked as the dominant version.

In addition to finding the peaks, the time duration distribution demonstrates another interesting property. If the time distribution has prominent peaks, this translates into a large collection of near-duplicate videos in the search result accordingly. Otherwise, the near-duplicate videos only take up a small portion in the search result and most of the videos are novel videos. For example, we can see from Fig. 4 that Query 1 and Query 19 have obvious dominant versions. The redundant videos that are near-duplicate to the most popular version of Query 1 and Query 19 are 42% and 36%, respectively, which are listed in Table I. It is also consistent with the assumption that near-duplicate videos have high possibility to fall into these peak regions. Videos having a large time difference, by definition, will not be regarded as near-duplicate videos. On the contrary, the time duration distribution for Query 8 is relatively flat, in which there is no prominent peak. From Table I, we can see that there are only 7% near-duplicate videos in the search result.

D. Seed Video Selection Using Thumbnail

Once the dominant time durations have been determined, the next step is to select one seed video for each dominant time duration. The objective is to pick one video that has the highest

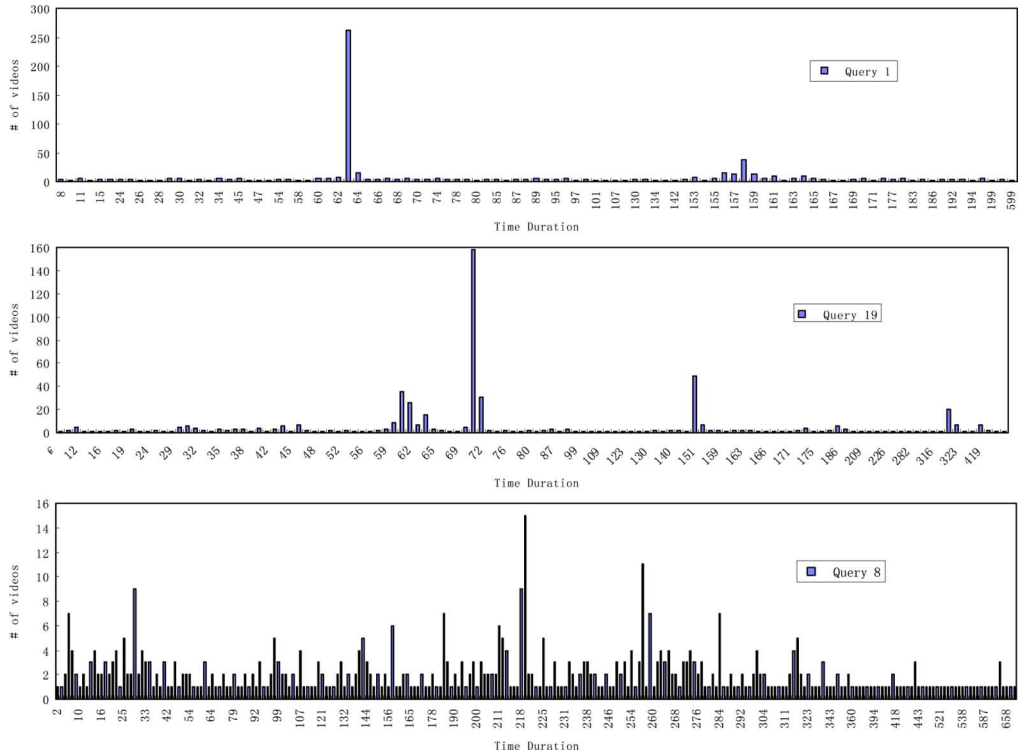

Fig. 4. Time duration distribution for the queries "The lion sleeps tonight" and "Sony Bravia" (Query 1 and Query 19) indicates that there might be multiple sets of duplicate videos in the search results. However, for query "Free hugs" (Query 8), the time distribution is relatively flat and the major time duration is not so prominent. The time with the largest number of videos will be picked as the dominant time duration.

occurrences for each dominant version. Videos having the similar time duration do not necessarily mean they are near-duplicates, as in the case where videos with the same background music might have totally different visual contents, which is a common scenario for web videos. Instead of employing pure content based method, here, we combine the content and context to swiftly select the seed video.

For efficiency, we use the color histogram of the thumbnail image and the number of views associated with videos to select the seed videos. A *relative distance distribution* is first built where an arbitrarily selected video is used as *reference video*. The distances between the thumbnail image of the reference video and all other videos in the set are computed, and the relative distance distribution in the form of a histogram is constructed. The distance of the color histogram is computed based on the *Euclidean* distance, formulated as

$$d(H_i, H_j) = \sqrt{\sum_{k=1}^{m} (x_k - y_k)^2}$$

where $H_i = (x_1, \ldots, x_m)$, and $H_j = (y_1, \ldots, y_m)$. The set of videos that fall into the largest distance bin thus forms the dominant near-duplicate group for the corresponding dominant time duration. Fig. 5 shows the relative distance distribution (histogram) based on the color histogram for videos of length 63 s in Query 1. From the distance distribution, we can find the largest set with the same color distance.

To pick the seed video, another context feature, view count, is employed. A higher number of clicks (views) indicates that a particular video is potentially the original version or it is more popular. The video with the biggest number of views will be picked as the seed video. In short, the video having the following two conditions will be selected as the seed video:

- its distance is within the largest relative distance distribution;
- within this distance, it has the largest number of views.

After that, the shot boundaries in the middle part of the seed video are detected, and representative keyframes (called *prototype keyframes*) are extracted for shots. Denote f_m as the frame at the middle of the video. Shot boundary detection is applied to the video segment of length $f_m \pm d$ for keyframe extraction, where d is the neighborhood window. The granularity of shot boundary detection should be as fine as possible so that scene

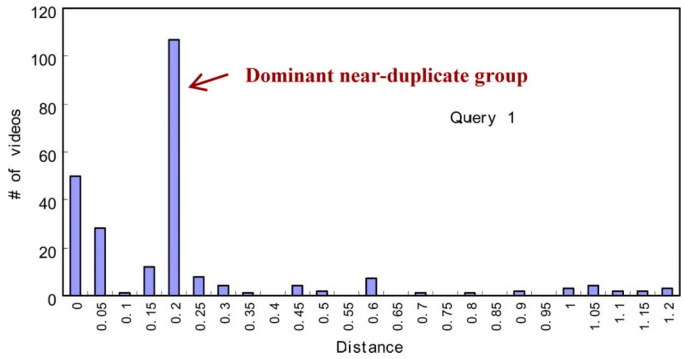

Fig. 5. Relative distance distribution based on color histogram for videos of length 63 s in Query 1.

Fig. 6. Middle part of prototype keyframes for the "White and Nerdy" seed video.

changes can be accurately detected, especially for videos with fast changing scenes, such as music videos. Sample keyframes for "White and Nerdy" are shown in Fig. 6. These prototype keyframes will be used for next step near-duplicate video elimination. Eventually, comparing an unprocessed video to a seed video is equivalent of matching its thumbnail image to the extracted prototype keyframes.

E. Near-Duplicate Video Elimination Using Content

Once the seed videos are selected, the elimination step will take place. The objective of search result novelty re-ranking is to list all the novel videos while maintaining the relevance order. To combine query relevance and novelty, each video V_i is computed through a pairwise comparison between V_i and every seed video S_j and previously ranked novel video N_j. The redundancy R is calculated by

$$R(V_i|S_1,\ldots,S_k,N_1,\ldots,N_m) = \mathrm{Max}(\max_{1\le j\le m} R(V_i|N_j), \max_{1\le j\le k} R(V_i|S_j)).$$

The comparison is performed according to the original ranking returned from the search engine. By default, the first video is a novel video. The preceding ranked video that is most similar to V_i determines the redundancy (R) of V_i. To determine if the video V_i is a novel video, it is matched to the sequence of novel videos and terminates when either: a) V_i is found to be visually duplicate to a video V_j or b) V_i is dissimilar to all videos. V_i is marked as a redundant video for the former case, while it is a new novel video for the latter.

During comparison, time duration information will act as a filter to avoid unnecessary computation. If the difference of time durations is large enough, the two videos will not be compared and they will be directly regarded as dissimilar. Otherwise, the thumbnail image of the video will be compared with the thumbnail images of previously novel videos and the prototype keyframes of the seed videos to evaluate the degree of redundancy. To increase efficiency, comparison is performed only on thumbnail images, instead of exhaustively between all keyframe pairs of two videos. The local point based matching method will be performed on the thumbnail images [26]. Similar to [31], we treat two images as near-duplicate if they have enough number of local point matching pairs. If a video is novel, it will be kept in the novel video list. Finally, the ranked list after removing all near-duplicate videos will be presented to the user.

IV. Experiments

A. Dataset and Evaluation

We selected 24 queries designed to retrieve the most viewed and top favorite videos from YouTube. Each text query was issued to YouTube, Google Video, and Yahoo! Video respectively and we collected all retrieved videos as our dataset, which is the same dataset used in [31]. The videos were collected in November, 2006. Videos with time duration over 10 minutes were removed from the dataset since they were usually documentaries or TV programs retrieved from Google, and were only minimally related to the queries. The final data set consists of 12 790 videos. Table II gives the statistics of web videos' format. The query information and the number of near-duplicates to the most popular version are listed in Table I. For example, there are 1 771 videos in Query 15 "White and Nerdy," and among them there are 696 near-duplicates of the most popular version in the result lists. Shot boundaries were detected using tools from CMU [12] and each shot was represented by a keyframe. Similar to [36], the words of title or tags that satisfy the condition will be treated as topic-specific stopwords: $tf(w,Q_i)/tf(Q_i) > \beta$, where $tf(w,Q_i)$ is the frequency of title/tag word w in videos of query Q_i, and $tf(Q_i)$ is the number of videos in query Q_i, and β is a threshold. In our experiment, $\beta = 0.05$. The setting is empirical and is simply used to remove the influence of frequent words. This preprocessing step is usually helpful as presented in [36]. Due to the noisy user-supplied text information (title and tags), special characters (e.g., $?, !, :, \#, >, |$) were first removed. Then the standard Porter stemming is applied to stem the text words. After a serial of data preprocessing (such as word stemming, special character removal, Chinese word segmentation, topic-specific stopword removal, and so on), there are 8 231 unique title words and 14 218 unique tag words, respectively.

TABLE II
Video Format Information

Format	# of videos	Percentage
FLV	10,925	85.4 %
MPG	45	0.3 %
AVI	1,714	13.4 %
WMV	98	0.7 %
MP4	8	0.1 %

Color histogram is based on the HSV color space, which is the same as [31], [34]. A color histogram is concatenated with 18 bins for *Hue*, 3 bins for *Saturation*, and 3 bins for *Value*. The local points of images were located by Hessian-Affine detector [25]. The local points were then described by PCA-SIFT [16], which is a 36 dimensional vector for each local point. With a fast indexing structure LIP-IS [26], local points were matched based on a point-to-point symmetric matching scheme [26].

Two non-expert assessors were asked to watch videos one query at a time. The videos were ordered according to the sequence returned by the video search engines. The assessors were requested to label the videos with a judgment (redundant or

novel) and to form the ground truth. To evaluate the re-ranking results, the assessors were requested to identify the near-duplicate clusters in an incremental way and the final ranking list was formed based on the original relevance ranking after removing near-duplicate videos.

In order to evaluate the performance, we use the novelty mean average precision (NMAP) to measure the ability to re-rank relevant web videos according to their novelty. The *novelty mean average precision (NMAP)* measures the mean average precision of all tested queries, considering only novel and relevant videos as the ground truth set. In other words, if two videos are relevant to a query but near-duplicate to each other, only the first video is considered as a correct match. For a given query, there are total of N videos in the collection that are relevant to the query. Assume that the system only retrieves the top k candidate novel videos where r_i is the number of novel videos seen so far from rank 1 to i. The NMAP is computed as

$$\text{NMAP} = \left(\sum_{i=1}^{k} r_i / i \right) / N.$$

The value of NMAP is in the range of 0 to 1. A value of 0 means all videos in the top-k list are near-duplicate of each other. In contrary, a value of 1 indicates that all top-k ranked videos are novel.

B. Performance Comparison

To evaluate the performance of novelty re-ranking, we compared the re-ranking results based on: a) context information only, where the time duration, title and tag context are evaluated separately; b) content information only, i.e., the hierarchical method (HIRACH) proposed in [31] and global signatures (Signature) using color histogram; and c) the proposed approach (CONT+CONX) that integrates the content and context information. HIRACH combines the global signature and pairwise comparison. A global signature from color histograms is first used to detect near-duplicate videos with high confidence and filter out very dissimilar videos. Then pairwise comparison among keyframes from two videos is performed with local points matching, which provides accurate analysis. The original ranking from the search engine serves as the baseline, while HIRACH serves as the upper limit performance.

For the time duration context, if the time difference between two videos is within an interval (e.g., 3 s), they will be treated as redundant. Similarly, for the global signature, two videos were regarded as duplicate when their signature difference is close enough (e.g., less than 0.15). Using the same empirical setting as in [31], we test different intervals (e.g., 0, 3, 5 s) and signature thresholds (e.g., 0.15, 0.2, 0.3), and the one with the best performance is reported. The titles and tags are compared using set difference measurement. If the titles or tags are same for two videos, they will be treated as duplicate.

Usually, the top search results receive the most attention for users. The performance comparison up to top 30 search results is illustrated in Fig. 7. It is obvious that the performance for original search results is not good because duplicate videos often

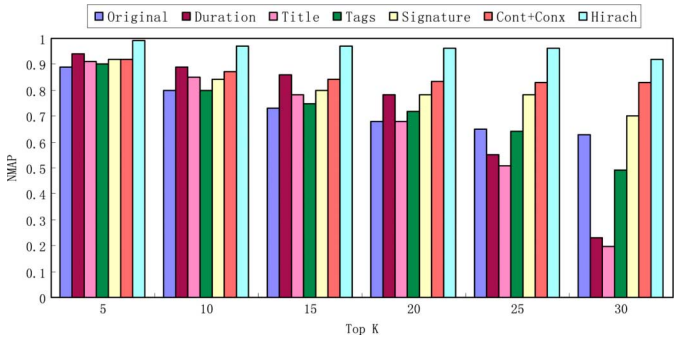

Fig. 7. Performance comparison of novelty re-ranking.

dominate the top list. In general, content-based methods, i.e., HIRACH and global signature, are able to consistently improve upon the baseline performance over all top k novel topics. As expected, HIRACH achieves the best performance. Although the global signature method can identify duplicate videos to some extent, its capability to discover more difficult duplicate videos is limited. A lot of near-duplicate videos cannot be correctly detected. Therefore the re-ranking list still consists of some duplicate videos and many novel videos are falsely removed.

Contextual information, when considered separately, can only maintain a good reranking performance for small values of k and they perform worse than the baseline when k grows larger. Although time duration information can distinguish novel videos initially, its usefulness is outlived as k increases. This shows the limitation of the contextual cue for web videos that have the same duration, especially for video queries accompanied with background music or music videos, e.g., Queries 1, 10, 23. As the number of videos increases, the information of time duration is inadequate, therefore the performance drops a lot. The approaches based on titles and tags have relatively poor performance. The titles and tags associated with web videos are not distinctive for the novelty and redundancy. The returned videos from search engines are all relevant to the text query, so they usually have the same titles and tags even though the content is totally new. Moreover, the titles and tags are rather noisy. Although, the methods based on titles and tags can identify novel stories at initial stage, the performance begins to deteriorate when k increases. A lot of novel videos are falsely filtered out simply because they are the same titles or tags, while duplicate videos are identified as novel due to additional or unusual tags are added. Generally, tags are more distinctive than the titles.

Our approach integrating content and context information achieves stable performance across all top k levels. The compensation from content and context makes this approach reliable. Certain videos can be filtered out simply based on time duration information. The local point based detection on thumbnail images guarantees the performance. Although the performance is slightly worse than the HIRACH method, the integration of content and context is able to compensate and overcome the limitations faced by each context as a separate entity.

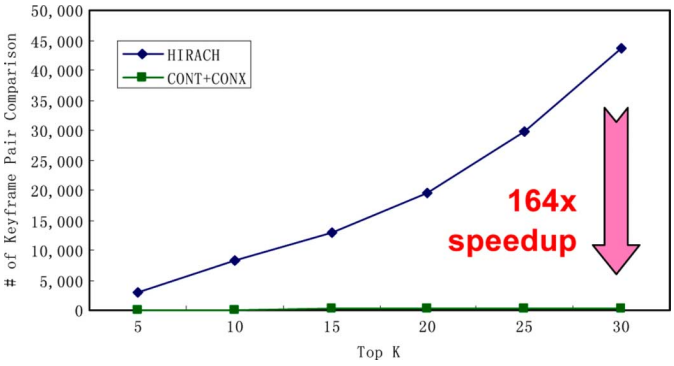

Fig. 8. Computation comparison between the hierarchical method (HIRACH) and the proposed method integrating content and context (CONT+CONX). CONT+CONX significantly improves the computation by 164 times speedup compared to HIRACH.

C. Computation Comparison

As search engines demand for quick response, the computation time is a critical factor for consideration. Mainly the following four factors affect the real-time computation cost: shot boundary detection for seeds videos, color feature extraction, local point extraction and the keyframe comparison. Because the thumbnail images are commonly extracted in the middle of videos, accordingly for the seed video, the shot boundary detection is not performed on the whole video, instead, only performed in the middle part (± 10 s). Usually a small number of prototype images are detected. The speed for shot boundary detection and color feature extraction is rather fast, which is negligible. Given the small number of prototype images for seed videos, the time for extracting local points is around 2 s per image, and local point extraction for thumbnail images is computed offline. Furthermore, the complexity for the mentioned processing is constant. However, the keyframe comparison will increase exponentially as the number of videos becomes larger, which is the main computation cost for the near-duplicate elimination. The keyframe comparison is the major computation factor in our evaluation.

The average number of keyframe pair comparison for top k re-ranking over 24 queries is showed in Fig. 8. Although HIRACH has been able to reduce the computation through the use of global signature filtering and the sliding window scheme, it is still infeasible for large scale duplicate elimination. The situation becomes worse as k becomes larger since it requires a large number of comparisons among keyframes.

In contrast, the proposed method integrating content and context is extremely efficient, achieving real-time elimination. A huge number of comparisons are filtered out with the time duration information. Furthermore, our proposed method only compares the thumbnail images or the prototype keyframes of seed videos, thus avoiding comprehensive keyframe pair comparisons. The speedup is especially evident for videos with fast changing scenes, e.g., Queries 15, 20, 23, which have over 100 keyframes. With the naive method, the exhaustive pairwise comparison between two videos is more than 100×100. In contrast, by using a small number of prototype keyframes for the seed videos and the thumbnail for the unprocessed video, the number of comparisons required by our method to match two videos has a constant time complexity. From Fig. 8, we can see that the average number of keyframe pair comparison grows almost exponentially using the hierarchical method. On the other hand, the method integrating content and context only has a linear increase. The speedup of our proposed method is around 164 times faster than the hierarchical method when re-ranking the top 30 novel videos. The acceleration will be more conspicuous as k becomes larger. Depending on the complexity of keyframes, the time for each keyframe pair comparison based on local points ranges from 0.01 to 0.1 second for a Pentium-4 machine with 3.4-GHz CPU and 1 G main memory. With the assistance of content and context information, our proposed approach achieves satisfactory results. The novelty re-ranking can respond in real-time, even for large scale web video platform, such as YouTube and Google.

V. Conclusion

Social web provides a platform for users to produce, share, view, and comment videos. Huge number of web videos is uploaded each day. Among them, there exist a large portion of near-duplicate videos. At the same time, rich context metadata are available with these uploaded videos, in the form of titles, tags, thumbnail images, time durations, and so on. Previous local point based methods have demonstrated promising performance. However, although multiple strategies have been adopted to accelerate the detection speed, it still cannot achieve real-time detection especially for large scale web video corpus.

From a different point of view, the context information provides constructive supplement for video content. Contextual information when considered separately is not effective while content information is accurate at the expense of time complexity. In our experiments, we show that the fusion of content and context is able to compensate the drawbacks of each method and deliver good precision and most importantly support online operations. Different sets of context information are considered jointly to perform different tasks at various stages of redundancy elimination. Time duration information is used to filter out dissimilar videos. Without digging into the complete video content, the detection is performed on the thumbnail images because the thumbnail image is a quick snapshot of the video content, which avoids the exhaustive pairwise comparison among keyframes of two videos. Experiments on 24 popular queries retrieved from YouTube showed that the proposed method that integrates the content and context information dramatically improve the detection efficiency. The speedup of performance is around 164 times faster than the effective hierarchical method proposed in [31]. Moreover, from experiments, we demonstrate that time duration and thumbnail image are useful contextual information to complement with content information. On the contrary, titles and tags are noisy, which turns to be less effective.

In this paper, time duration and thumbnail image are two critical context features used to eliminate the near-duplicate web

videos. User-supplied titles, tags and other text description attached to web videos are usually inaccurate, ambiguous, and even erroneous for video sharing websites. However, among the noisy text information, there exist some useful cues worth exploring, such as episode number, named entities, which provide useful information for the novelty detection. In the future, we plan to further explore the noisy text related metadata to find meaningful information. Furthermore, the social network formed in the social web is another interesting topic to study. User groups, usernames, related/relevant videos, user relationship, and relevance relationship among videos are interesting contextual and social information to explore, which will contribute to web video search and retrieval.

REFERENCES

[1] D. A. Adjeroh, M. C. Lee, and I. King, "A distance measure for video sequences," *J. Comput. Vis. Image Understand.*, vol. 75, no. 1, pp. 25–45, 1999.

[2] *Topic Detection and Tracking: Event-Based Information Organization*, J. Allan, Ed. Boston, MA: Kluwer, 2002.

[3] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. ACM SIGIR*, Melbourne, Australia, Aug. 1998, pp. 335–336.

[4] S.-F. Chang *et al.*, "Columbia University TRECVID-2005 video search and high-level feature extraction," in *Proc. TRECVID*, Washington, DC, 2005.

[5] S. C. Cheung and A. Zakhor, "Efficient video similarity measurement with video signature," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 59–74, Jan. 2003.

[6] S. C. Cheung and A. Zakhor, "Fast similarity search and clustering of video sequences on the world-wide-web," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 3, pp. 524–537, Jun. 2005.

[7] *Google Video*, [Online]. Available: http://video.google.com, Available

[8] A. Hampapur and R. Bolle, "Comparison of sequence matching techniques for video copy detection," in *Proc. Storage and Retrieval for Media Databases*, 2002.

[9] T. C. Hoad and J. Zobel, "Fast video matching with signature alignment," *Proc. MIR'03*, pp. 262–269, 2003.

[10] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *Proc. ACM Conf. Multimedia*, 2006, pp. 35–44.

[11] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," in *Proc. ACM Conf. Multimedia*, Augsburg, Germany, Sep. 2007, pp. 971–980.

[12] *Informedia*, [Online]. Available: http://www.informedia.cs.cmu.edu, Available

[13] A. Jaimes, "Conceptual Structures and Computational Methods for Indexing and Organization of Visual Information," Ph.D. dissertation, Columbia Univ., New York, 2003.

[14] A. K. Jain, A. Vailaya, and W. Xiong, "Query by video clip," *ACM Multimedia Syst. J.*, vol. 7, pp. 369–384, 1999.

[15] A. Joly, O. Buisson, and C. Frelicot, "Content-based copy retrieval using distortion-based probabilistic similarity search," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 293–306, Feb. 2007.

[16] Y. Ke, R. Sukthankar, and L. Huston, "Efficient near-duplicate detection and sub-image retrieval," in *Proc. ACM Conf. Multimedia*, Oct. 2004, pp. 869–876.

[17] L. Kennedy *et al.*, "How Flickr helps us make sense of the world: Context and content in community-contributed media collections," in *Proc. ACM Conf. Multimedia*, Augsburg, Germany, Sep. 2007, pp. 631–640.

[18] J. Law-To, B. Olivier, V. Gouet-Brunet, and B. Nozha, "Robust voting algorithm based on labels of behavior for video copy detection," in *Proc. ACM Conf. Multimedia*, 2006, pp. 835–844.

[19] R. Lienhart and W. Effelsberg, "VisualGREP: A systematic method to compare and retrieve video sequences," *Multimedia Tools Appl.*, vol. 10, no. 1, pp. 47–72, Jan. 2000.

[20] J. Liu, W. Lai, X.-S. Hua, Y. Huang, and S. Li, "Video search re-ranking via multi-graph propagation," in *Proc. ACM Conf. Multimedia*, Augsburg, Germany, Sep. 2007, pp. 208–217.

[21] L. Liu, W. Lai, X.-S. Hua, and S.-Q. Yang, "Video histogram: a novel video signature for efficient web video duplicate detection," in *Proc. Multimedia Modeling Conf.*, Jan. 2007.

[22] X. Liu, Y. Zhuang, and Y. Pan, "A new approach to retrieve video by example video clip," in *Proc. ACM Conf. Multimedia*, Orlando, FL, 1999, pp. 41–44.

[23] D. Lowe, "Distinctive image features from scale-invariant key points," *Int. J. Computer Vision*, vol. 60, pp. 91–110, 2004.

[24] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in *Proc. Comput. Vis. Pattern Recognit.*, 2003, pp. 257–263.

[25] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, pp. 63–86, 2004.

[26] C.-W. Ngo, W.-L. Zhao, and Y.-G. Jiang, "Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation," in *Proc. ACM Conf. Multimedia*, 2006, pp. 845–854.

[27] Y. Peng and C.-W. Ngo, "Clip-based similarity measure for query-dependent clip retrieval and video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 5, pp. 612–627, May 2006.

[28] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[29] B. Shevade, H. Sundaram, and L. Xie, "Modeling personal and social network context for event annotation in images," in *Proc. JCDL*, Vancouver, BC, Canada, Jun. 2007, pp. 127–134.

[30] *Wikipedia*, [Online]. Available: http://en.wikipedia.org/wiki/Youtube

[31] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Practical elimination of near-duplicates from web video search," in *Proc. ACM Conf. Multimedia*, Augsburg, Germany, Sep. 2007, pp. 218–227.

[32] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts," in *Proc. ACM Conf. Multimedia*, Augsburg, Germany, Sep. 2007, pp. 168–177.

[33] X. Wu, W.-L. Zhao, and C.-W. Ngo, "Near-duplicate keyframe retrieval with visual keywords and semantic context," in *Proc. ACM Conf. Image and Video Retrieval*, Amsterdam, the Netherlands, Jul. 2007, pp. 162–169.

[34] X. Wu, C.-W. Ngo, and Q. Li, "Threading and autodocumenting news videos," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 59–68, Mar. 2006.

[35] *Yahoo! Video*, [Online]. Available: http://video.yahoo.com, Available

[36] Y. Yang, J. Zhang, J. Carbonell, and C. Jin, "Topic-conditioned novelty detection," in *Proc. SIGKDD*, 2002.

[37] *YouTube*, [Online]. Available: http://www.youtube.com, Available

[38] J. Yuan, L.-Y. Duan, Q. Tian, S. Ranganath, and C. Xu, "Fast and robust short video clip search for copy detection," in *Proc. Pacific Rim Conf. Multimedia*, 2004, pp. 479–488.

[39] C. Zhai, W. Cohen, and J. Lafferty, "Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval," in *Proc. ACM SIGIR*, Toronto, ON, Canada, 2003, pp. 10–17.

[40] B. Zhang *et al.*, "Improving web search results using affinity graph," in *Proc. ACM SIGIR*, Salvador, Brazil, Aug. 2005, pp. 504–511.

[41] D.-Q. Zhang and S.-F. Chang, "Detecting image near-duplicate by stochastic attributed relational graph matching with learning," in *Proc. ACM Conf. Multimedia*, Oct. 2004, pp. 877–884.

[42] Y. Zhang, J. Callan, and T. Minka, "Novelty and redundancy detection in adaptive filtering," in *Proc. ACM SIGIR*, Tampere, Finland, Aug. 2002, pp. 81–88.

Xiao Wu (S'05) received the B.Eng. and M.S. degrees in computer science from Yunnan University, Yunnan, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Department of Computer Science, City University of Hong Kong, in 2008.

He is a Senior Research Associate at the City University of Hong Kong. From 2006 to 2007, he was with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, as a Visiting Scholar. He was with Institute of Software, Chinese Academy of Sciences, Beijing, China, from 2001 to 2002, and then the City University of Hong Kong as a Research Assistant between 2003 and 2004. His research interests include multimedia information retrieval and video processing.

Chong-Wah Ngo (M'02) received the B.Sc. and M.Sc. degrees in computer engineering from Nanyang Technological University, Nanyang, Singapore, and the Ph.D. degree in computer science from the Hong Kong University of Science and Technology in 2000.

Before joining City University of Hong Kong in 2002, he was with the Beckman Institute, University of Illinois at Urbana-Champaign. He was also a Visiting Researcher with Microsoft Research Asia in 2002. His research interests include video computing and multimedia information retrieval.

Alexander G. Hauptmann (M'92) received the B.A. and M.A. degrees in psychology from Johns Hopkins University, Baltimore, MD. He studied Computer Science at the Technische Universität Berlin, Berlin, Germany, from 1982 to 1984, and received the Ph.D. degree in computer science from Carnegie Mellon University (CMU), Pittsburgh, PA, in 1991.

He is a Senior Systems Scientist in the Computer Science Department of CMU and also a Faculty Member with CMU's Language Technologies Institute. His research interests have led him to pursue and combine several different areas: man-machine communication, natural language processing, speech understanding and synthesis, machine learning. He worked on speech and machine translation at CMU from 1984 to 1994, when he joined the Informedia project for digital video analysis and retrieval and led the development and evaluation of the News-on-Demand applications.

Hung-Khoon Tan received the B.Eng. degree in computer engineering from the University of Technology of Malaysia (UTM), the M.Phil. degree in computer science from the City University of Hong Kong, and the M.Eng. degree in microelectronics from Multimedia University (MMU). He is currently pursing the Ph.D. degree in the Department of Computer Science, City University of Hong Kong.

He was a Test Development and Senior Design Engineer in Altera's E&D Center, Penang, Malaysia, from 1999 to 2004. His research interests include multimedia content analysis, data mining, and pattern recognition.