10-2010

Data-driven approaches to community-contributed video applications

Xiao WU

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

Wan-Lei ZHAO

Citation

1

Data-Driven Approaches to Community-Contributed Video Applications

Xiao Wu
Southwest Jiaotong University

Chong-Wah Ngo and Wan-Lei Zhao
City University of Hong Kong

The authors demonstrate three social media applications that are built around data-driven and context-based analysis.

With the prosperity of video-sharing websites such as YouTube, the amount of community-contributed video has increased dramatically. Reportedly more than 65,000 new videos were uploaded to YouTube every day in July 2006 and it's estimated that 20 hours of new videos were uploaded to the site every minute in May 2009. In addition to the huge volume of video data, the social Web provides rich contextual and social resources associated with videos. These resources include title, tags, thumbnails, related videos, and user and community information, as illustrated in Figure 1. While billions of user-generated videos accompanied with rich-media information have enriched the Web-browsing experience, this scenario brings new opportunities and challenges for effective and efficient searching, mining, and organizing of large-scale Web videos.

Recent advances in machine learning have brought significant progress related to various building blocks in multimedia information retrieval, including annotation, classification, and high-level concept detection. The related approaches heavily rely on techniques such as support vector machines (SVMs) to map the low-level features to high-level semantics. The so-called model-based approaches normally involve learning a large set of classifiers. In general, these approaches become ineffective for managing the ever-increasing amount of Web videos for reasons such as scarcity of training examples, complexity in processing time, and unlimited vocabulary size.

On the other hand, with the explosive expansion of the social Web, developing data-driven approaches becomes feasible when huge amounts of Web videos are freely available, paving a new way to approach traditional applications with a data-driven methodology. With the overwhelming volume of community-contributed video data, many problems can be solved without the need for sophisticated algorithms. Recently, such data-driven techniques have been evident in the use of word similarity measure,[1] object recognition,[2] and video annotation.[3] In this article, we explore near-duplicate Web video detection, video annotation, and video classification from a new data-driven framework perspective, by exploiting different aspects derived from contextual and social resources.

Data-driven applications

The overall framework of our data-driven approaches is depicted in Figure 2. Various contextual and social features are exploited with different data-driven techniques in three applications:

- Time duration and thumbnail images are two contextual features explored for real-time detection of near-duplicates. Careful coupling of content and context provides a promising way to hit a balance between speed and accuracy for practical elimination of near-duplicate Web videos.
- With a huge volume of tagged Web videos freely accessible online, a novel classifier-free video annotation is proposed by first retrieving visual duplicates and then suggesting representative tags from the outcome of near-duplicate search.

1070-986X/10/$26.00 © 2010 IEEE

- The labels of related videos and user videos provide constructive clues for Web video categorization. The semantic meaning of title and tags, video relevance from related videos, and user interest induced from user videos, are integrated to determine the video category. The model-based classifiers built based on title and tags, and the data-driven approaches derived from the majority voting on related videos and user videos are combined to obtain a robust measure.

In the following subsections, we will elaborate the three applications.

Near-duplicate elimination

Near-duplicate Web videos are identical or approximately identical videos close to the exact duplicate of each other. They have similar time duration and length, but are different in terms of file formats, encoding parameters, photometric variations (color, lighting changes), editing operations (caption, logo, and border insertion), and certain modifications (frames added or removed).[4]

Near-duplicate video detection[4,5] and video copy detection[6,7] have been extensively studied in recent years. Many methods emphasize the rapid identification of duplicate videos with compact and reliable global features.[6] The searching and indexing methods become critical when the volume of data is huge, especially for Web videos. To achieve good performance, existing approaches usually adopt local keypoint features (for example, SIFT), which induce extensive computation. To meet the scalability requirement, a compact frame-level signature known as Global[7] is proposed to compare the frame similarity and avoid the large amount of comparison among local keypoints.

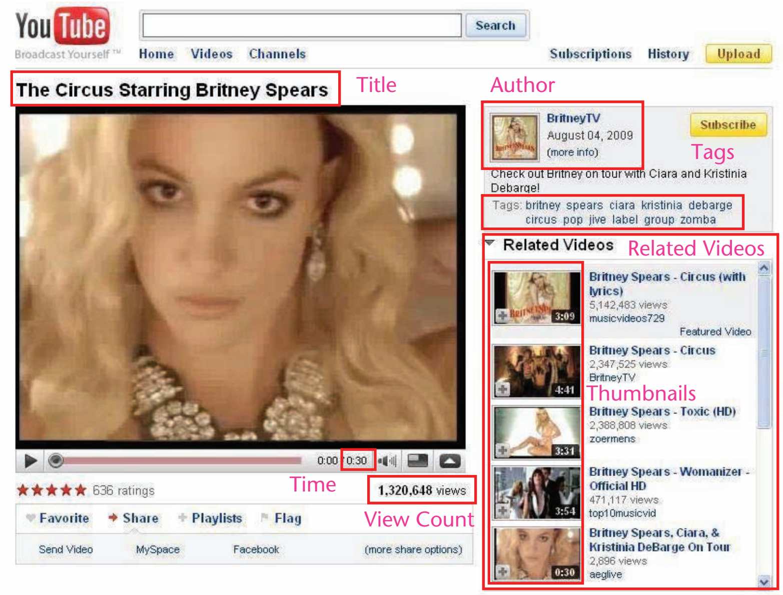


Figure 1. Contextual and social information associated with Web videos.

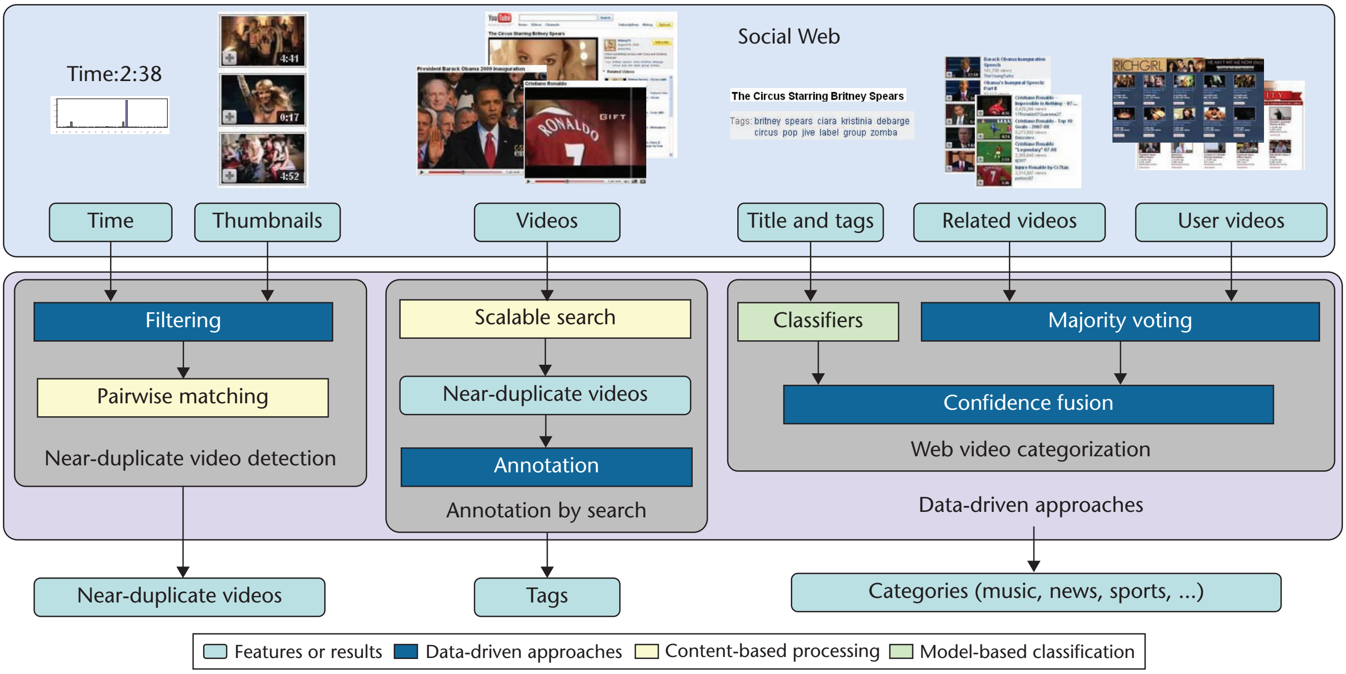
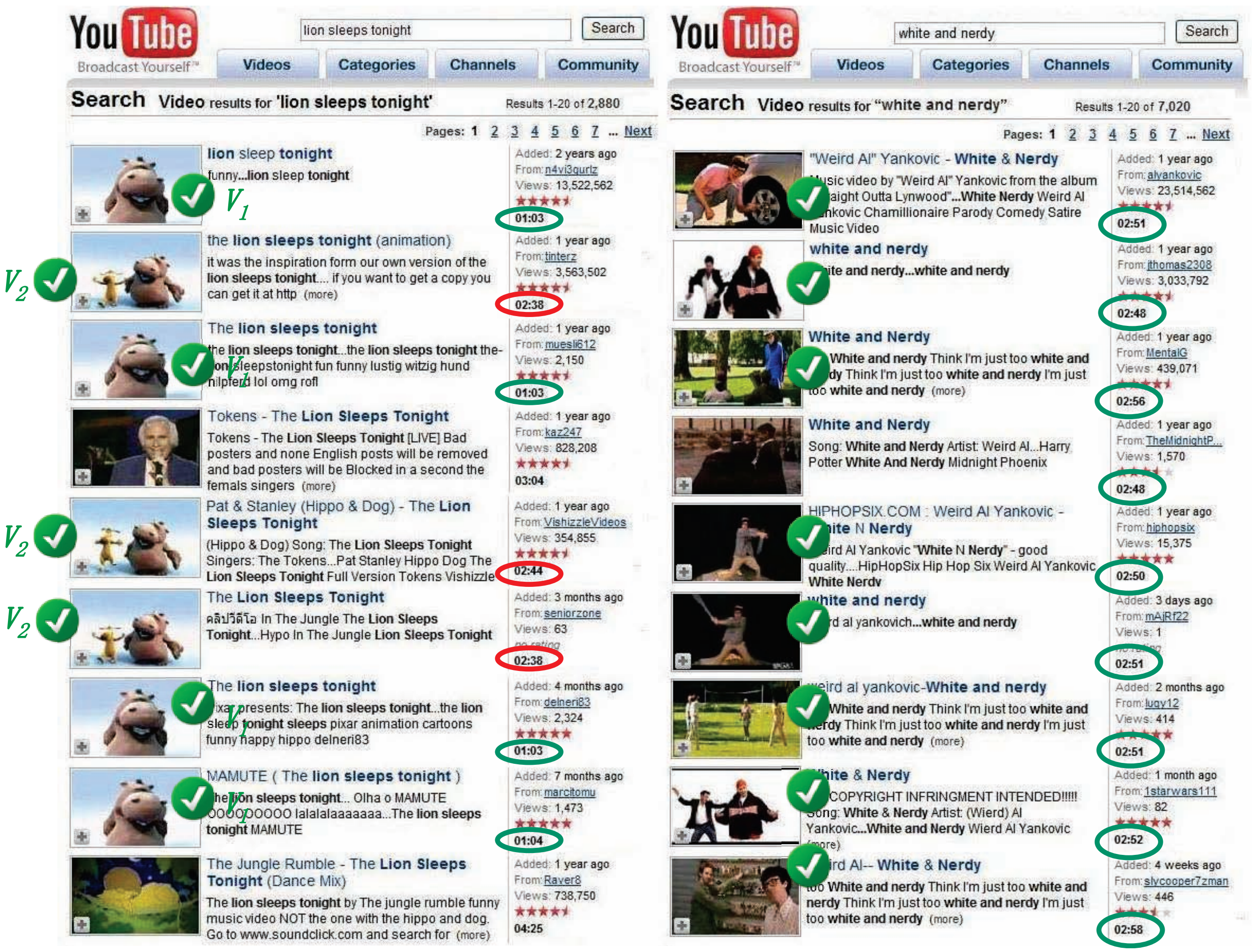


Figure 2. The framework for data-driven applications.



lion sleeps tonight | Search

Videos | Categories | Channels | Community

Search Video results for 'lion sleeps tonight' Results 1-20 of 2,880

Pages: 1 2 3 4 5 6 7 ... Next

lion sleep tonight
funny...lion sleep tonight
Added: 2 years ago
From: n4vi3qurlz
Views: 13,522,562
01:03

the lion sleeps tonight (animation)
it was the inspiration form our own version of the lion sleeps tonight.... if you want to get a copy you can get it at http (more)
Added: 1 year ago
From: tinterz
Views: 3,563,502
02:38

The lion sleeps tonight
the lion sleeps tonight...the lion sleeps tonight the-lionsleepstonight fun funny lustig witzig hund nilpferd lol omg rofl
Added: 1 year ago
From: muesli612
Views: 2,150
01:03

Tokens - The Lion Sleeps Tonight
Tokens - The Lion Sleeps Tonight [LIVE] Bad posters and none English posts will be removed and bad posters will be Blocked in a second the femals singers (more)
Added: 1 year ago
From: kaz247
Views: 828,208
03:04

Pat & Stanley (Hippo & Dog) - The Lion Sleeps Tonight
(Hippo & Dog) Song: The Lion Sleeps Tonight Singers: The Tokens...Pat Stanley Hippo Dog The Lion Sleeps Tonight Full Version Tokens Vishizzle
Added: 1 year ago
From: VishizzleVideos
Views: 354,855
02:44

The Lion Sleeps Tonight
คลิปวีดีโอ In The Jungle The Lion Sleeps Tonight...Hypo In The Jungle Lion Sleeps Tonight
Added: 3 months ago
From: seniorzone
Views: 63
no rating
02:38

The lion sleeps tonight
Pixar presents: The lion sleeps tonight...the lion sleep tonight sleeps pixar animation cartoons funny happy hippo delneri83
Added: 4 months ago
From: delneri83
Views: 2,324
01:03

MAMUTE (The lion sleeps tonight)
The lion sleeps tonight... Olha o MAMUTE OOOOOOOOO lalalalaaaaaaa...The lion sleeps tonight MAMUTE
Added: 7 months ago
From: marcitomu
Views: 1,473
01:04

The Jungle Rumble - The Lion Sleeps Tonight (Dance Mix)
The lion sleeps tonight by The jungle rumble funny music video NOT the one with the hippo and dog. Go to www.soundclick.com and search for (more)
Added: 1 year ago
From: Raver8
Views: 738,750
04:25

white and nerdy | Search

Videos | Categories | Channels | Community

Search Video results for "white and nerdy" Results 1-20 of 7,020

Pages: 1 2 3 4 5 6 7 ... Next

"Weird Al" Yankovic - White & Nerdy
usic video by "Weird Al" Yankovic from the album aight Outta Lynwood"...White Nerdy Weird Al nkovic Chamillionaire Parody Comedy Satire Music Video
Added: 1 year ago
From: alyankovic
Views: 23,514,562
02:51

white and nerdy
ite and nerdy...white and nerdy
Added: 1 year ago
From: jthomas2308
Views: 3,033,792
02:48

White and Nerdy
White and nerdy Think I'm just too white and dy Think I'm just too white and nerdy I'm just too white and nerdy (more)
Added: 1 year ago
From: MentalG
Views: 439,071
02:56

White and Nerdy
Song: White and Nerdy Artist: Weird Al...Harry Potter White And Nerdy Midnight Phoenix
Added: 1 year ago
From: TheMidnightP...
Views: 1,570
02:48

HIPHOPSIX.COM : Weird Al Yankovic - ite N Nerdy
ird Al Yankovic "White N Nerdy" - good quality....HipHopSix Hip Hop Six Weird Al Yankovic White Nerdy
Added: 1 year ago
From: hiphopsix
Views: 15,375
02:50

white and nerdy
rd al yankovich...white and nerdy
Added: 3 days ago
From: mAjRf22
Views: 1
no rating
02:51

eird al yankovic-White and nerdy
White and nerdy Think I'm just too white and erdy Think I'm just too white and nerdy I'm just too white and nerdy (more)
Added: 2 months ago
From: lugy12
Views: 414
02:51

hite & Nerdy
COPYRIGHT INFRINGMENT INTENDED!!!!! ong: White & Nerdy Artist: (Wierd) Al Yankovic...White and Nerdy Wierd Al Yankovic ore)
Added: 1 month ago
From: 1starwars111
Views: 82
02:52

ird Al-- White & Nerdy
oo White and nerdy Think I'm just too white and nerdy Think I'm just too white and nerdy I'm just too white and nerdy (more)
Added: 4 weeks ago
From: slycooper7zman
Views: 446
02:58

Figure 3. Near-duplicate Web videos can be identified by checking their time durations and thumbnails. They usually have similar time durations, and thumbnail images are commonly matched to the middle parts of seed videos.

Good performance has been achieved in our previous work[4] by combining the global signature derived from color histograms and pairwise comparison among keyframes based on local points. Coupled with the use of a sliding window scheme, considerable savings have been made to relieve the computational burden. Unfortunately, the local-point-based methods remain expensive, considering the prohibitively large number of keyframe pairs between two videos and local points within a single keyframe (usually hundreds to thousands). Therefore, computation becomes the bottleneck for Web-scale, near-duplicate video detection.

Near-duplicate Web videos more or less have a similar time duration, with a difference of only a few seconds. While thumbnail images are usually extracted from the middle of a video to roughly represent the main content, the thumbnails of near-duplicate videos can be commonly derived from a small set of keyframes extracted from the original versions, from which most videos are edited. Figure 3 illustrates two typical search results from YouTube in which we can determine the near-duplicate videos simply by time and thumbnails. Therefore, when there are a large number of relevant videos returned by search engines, it's sufficient to categorize two videos as near-duplicates if their time durations are close to each other and their thumbnail images are near-duplicate. Time duration acts as a critical feature for efficient filtering of dissimilar videos and locating near-duplicate ones. On the other hand, thumbnail images extracted from videos reveal the near-duplicate identity of Web videos.

Based on this observation, we explore the use of time duration and thumbnail images for real-time, near-duplicate elimination.[5] The method consists of three main processes: dominant version identification, seed video

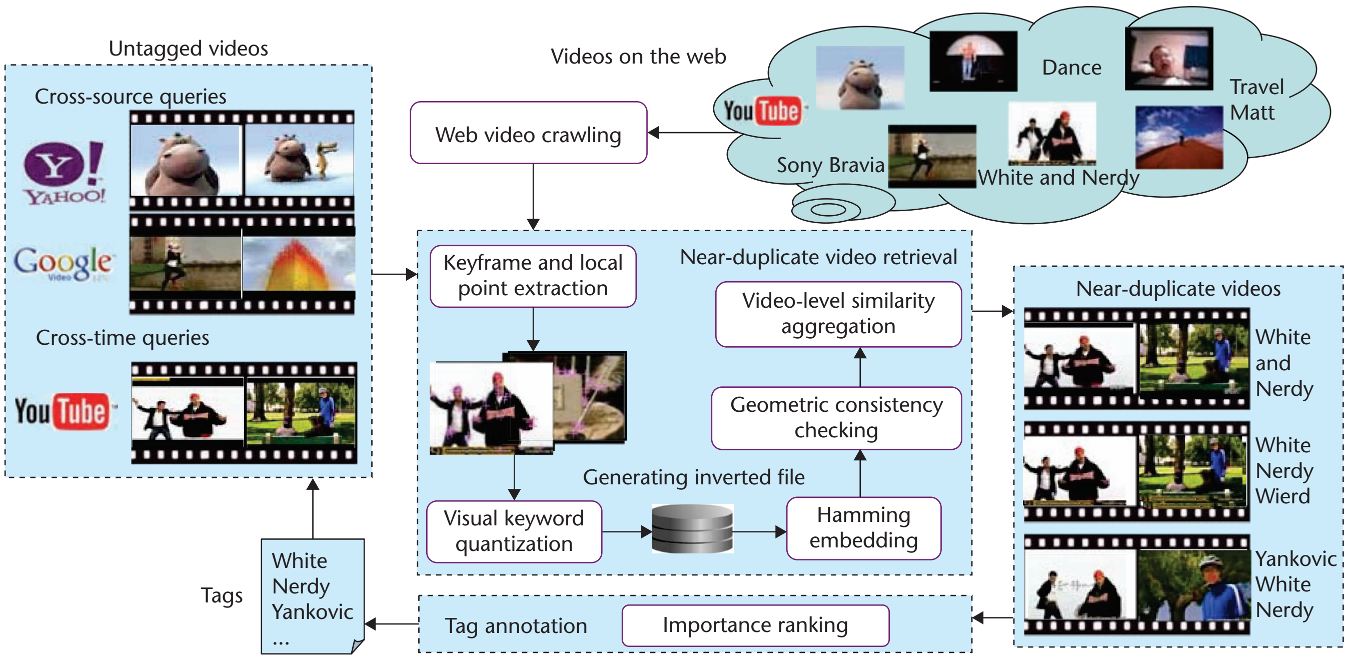


Figure 4. Annotation by scalable search.

selection, and near-duplicate video elimination. As a preprocessing step, we use time duration to rapidly but coarsely identify the preliminary groups of near-duplicate videos. We then select in a dominant group, which is most likely the set of duplicates. For each dominant version, a seed video is selected. This could potentially be the original source from which other videos are derived.

The final step of near-duplicate detection is comparing thumbnail images of candidate videos to the selected seed videos. To ensure accurate detection, this step is performed by matching local points. The integration of content and context analysis can reach real-time reranking speed with high efficiency, in which the majority of duplicates can be swiftly detected and removed from the top rankings. The novel use of context information lets us skip expensive content processing in near-duplicate video detection.

Annotation by scalable search

Automatic Web video annotation refers to the action of automatically associating relevant keywords to Web videos, which plays a vital role in many information retrieval tasks. Social tagging is widely adopted in various social media websites. Existing efforts are mostly devoted to tag recommendation and refinement, while the issue of annotation by scalable search has not been fully explored. One idea for annotation is to use the existing tags from community-contributed videos to label new videos. Basically, given an untagged video, similar videos are first retrieved from a database. The associated tags of similar videos are then examined and appropriate tags are picked for annotating the new video.

From a data-perspective point of view, the data-driven approach becomes possible when enough videos and tags are available to characterize any new incoming data. In a manner that is different from traditional approaches that adopt computer vision or machine learning techniques, we investigate the potential of a data-driven and model-free approach to annotate Web videos. Figure 4 illustrates the framework for Web video annotation. Two major components in this simple but effective data-driven solution are scalable near-duplicate video search and classifier-free tag annotation.

For the first issue, we propose a scalable near-duplicate retrieval framework by considering the indexing of local visual features, fast pruning of false matches at frame level, and localization of near-duplicates at video level. The keyframes that are similar to query keyframes are retrieved from video corpus via visual keywords, inverted file, and Hamming embedding. The retrieved keyframes are further reranked according to their geometric consistency with the query. Finally, the similarity of a video is determined by aggregating the scores of keyframes in the video while weakly considering their temporal consistency with the query, which is measured by a 2D Hough transform.

For the second issue, we adopt a voting scheme for recommending tags by observing the tagging behaviors in a pool of retrieved near-duplicates. Intuitively, if different users label visually similar videos with the same tags, these tags are likely to reflect an objective view of the video content. We propose a method that considers tag frequency, number of tags, and similarity weight of videos to rank the tags according to their relevance. We expect that partial near-duplicates will contribute less than full near-duplicates to the similarity. Eventually, the top few tags with higher rank are then recommended for annotating the query videos.

Classification by contextual information

Web video classification refers to the process of assigning Web videos to predefined categories such as sports, news, music, and so on. One attractive aspect of community-contributed social media is the abundant amount of context metadata. The related videos frequently have relevant content or similar category labels with the given video. At the same time, users (authors) upload and share videos based on their personal interests. Therefore the uploaded videos by the same author usually have a similar type. For example, videos from author "stanforduniversity" are tightly associated with the education category, while videos from author "BMWwebTV" mainly belong to the autos and vehicles category. Checking the category labels of related videos and user videos gives constructive clues for categorization.

For the sake of effectiveness, efficiency, and scalability, we combine the contextual information to classify Web videos under the scenario of the social Web.[8] Figure 2 illustrates the framework for Web video categorization. The semantic meaning of title and tags ($Semantics_i$), video relevance from related videos ($Relevance_i$), and user interest induced from user videos ($Interest_i$), provide complementary clues from three different viewpoints. Their combination gives a more accurate and confident judgment $Score_i(V_j)$, which is determined as follows:

$$Score_i\,(V_j) = Semantics_i\,(V_j) + Relevance_i\,(V_j) + Interest_i\,(V_j)$$

where $Semantics_i\,(V_j)$ is the probability score of video V_j belonging to the predefined category C_i, in the range of [0, 1]. $Semantics_i\,(V_j)$ is based on the model-based classifier trained by SVM on title and tags. $Relevance_i\,(V_j)$ and $Interest_i\,(V_j)$ are confidence scores according to class label distribution from related videos and user videos, respectively, defined as:

$$Relevance_i\,(V_j) = |R_{ij}|/|R_j|$$

and

$$Interest_i\,(V_j) = |U_{ij}|/|U_j|$$

where R_j is the set of related videos for video V_j, and R_{ij} is the set of related videos having category label C_i among R_j. Similarly, U_j is the set of user videos uploaded by the same user as video V_j. Note that the sum of distribution over different categories is equal to 1.

From a data viewpoint, this solution combines data-driven and model-based methods to achieve a robust measurement. Model-based classifiers are built based on text features (title and tags), while the data-driven majority voting is evaluated based on related videos and user videos. A skewed category distribution with a dominant peak indicates that the videos in this distribution have certain category preferences. The new video has a high possibility of having the same category label as the dominant category. While a flat distribution with low scores means that their relevance or interest is diverse, the incoming video could be any category.

Figure 5a shows the semantic scores according to classifiers built on text. Based on it, a video of guitar lesson will be falsely classified as "travel". As illustrated in Figures 5b and 5c, we can see that the skewed category distributions for both related videos and user videos are dominated by the "music" category. The method denotes that the video is tightly related to music and that the author "swanner" has a strong interest in music videos. Therefore, the new uploaded video has a high probability of being in the same category. The video category is ultimately determined by the fusion of the three components.

Evaluation

Until now, there have been few benchmark Web video datasets for large-scale social media applications. We use two datasets: CC_WEB_VIDEO (available at http://vireo.cs.cityu.edu.hk/webvideo/) and MCG_WEBV to test the performance of data-driven applications.

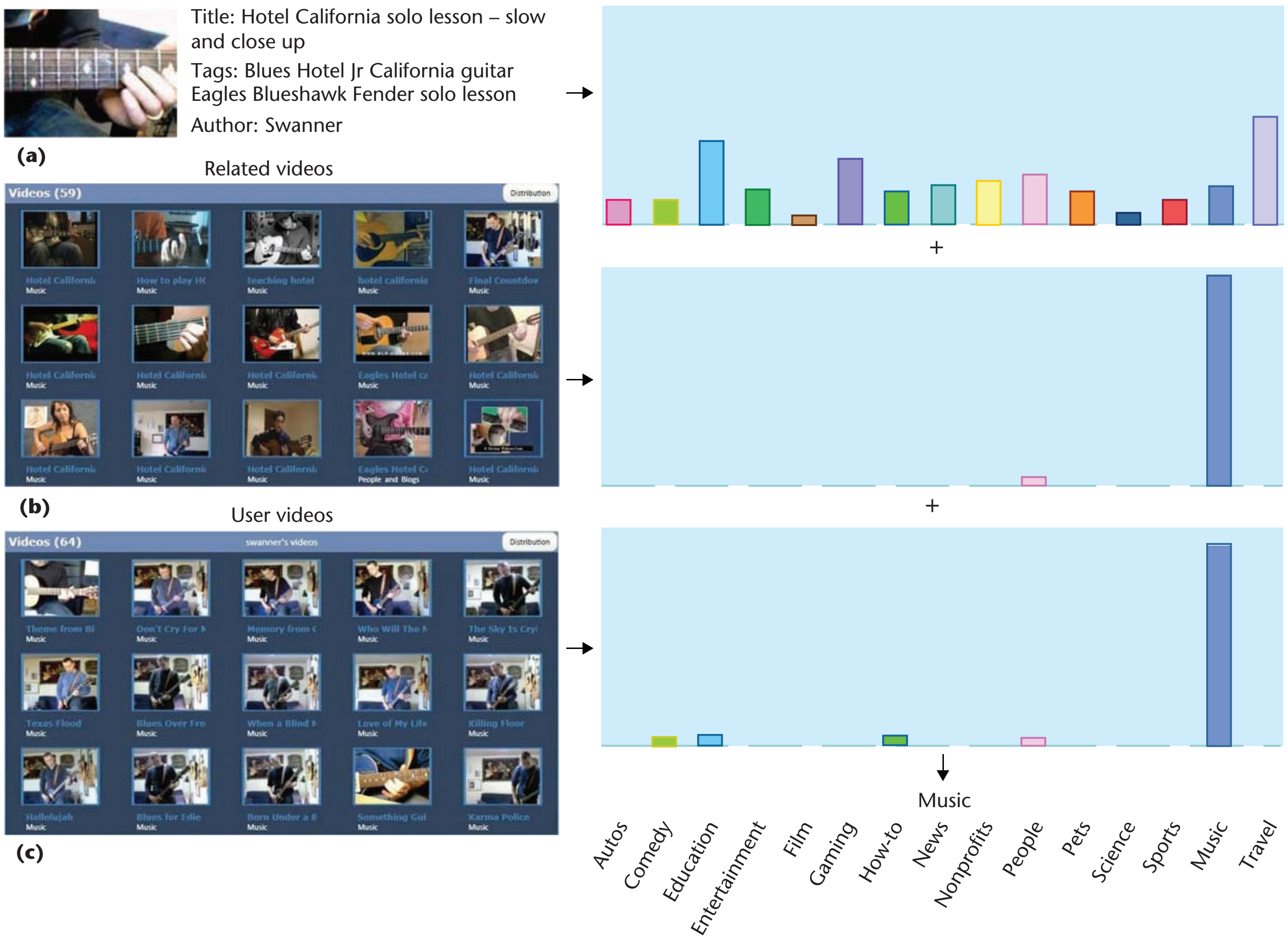


Figure 5. According to classifiers built on text (a) a video on guitar lesson is falsely suggested as "travel." According to the graph, the travel category has the highest score. When considering (b) video relevance (showing related videos) and (c) user interest (showing user videos), our method correctly recommends "music."

CC_WEB_VIDEO is the near-duplicate Web video dataset[4] and now publicly released. The main purpose of this corpus is for near-duplicate Web video detection and search reranking. It can also be used for Web video annotation, video copy detection, object recognition, and so on. The data set consists of 12,790 Web videos. A total of 24 representative queries are designed to retrieve the most viewed and top favorite videos from YouTube. The queries are issued to search engines (YouTube, Google Video, and Yahoo! Video) in the form of short text descriptions. The retrieved videos from different search engines are then pooled to form the dataset. The videos were collected in November 2006. Altogether, there are 398,009 keyframes, 8,231 unique title words, and 14,218 unique tag words.

Since the distribution of video data evolves over time, we expect that the uploaded videos might deviate from the original ones, and that the number of near-duplicate videos should diminish. To verify the robustness of a data-driven approach for Web video annotation, another new 8,459 videos were collected using the same queries as in CC_WEB_VIDEO but crawled at different time (in December 2008). We use the tags of videos in the original CC_WEB_VIDEO as historical data to annotate the newly downloaded videos.

Another recently released dataset, MCG_WEBV[9] from Institute of Computing Technology, Chinese Academy of Sciences, was collected mainly for video categorization. The core data was collected by retrieving the most-viewed videos of the month (from December 2008 to March 2009) from 15 predefined YouTube categories (such as music, sports, people, and so on). Each category has around 100 videos except the ''music'' category. The core data from MCG_WEBV consist of 4,610 Web videos, which we selected as our training dataset for Web video categorization.

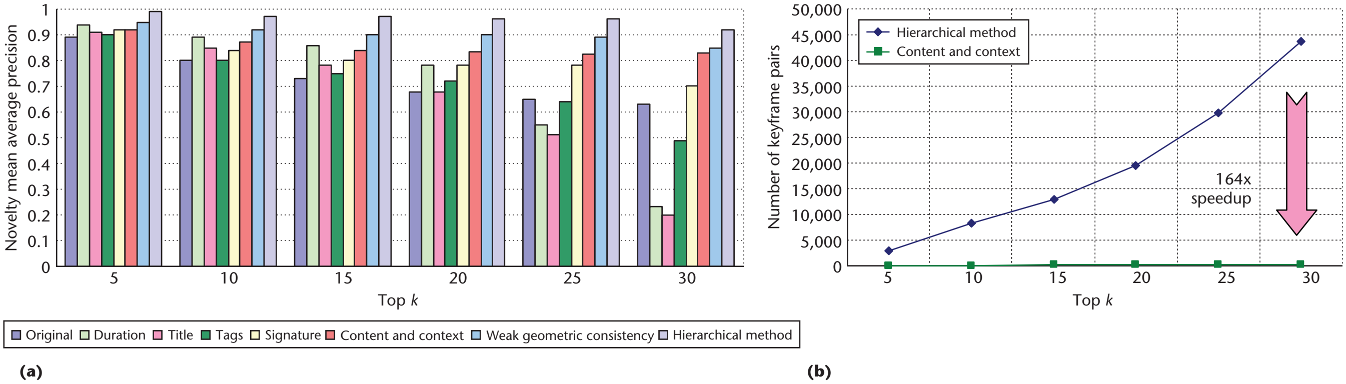


Figure 6. (a) Performance comparison and (b) computational improvement. Although the performance of the data-driven method integrating content and context is slightly worse than the hierarchical method, it significantly improves the computation.

Apart from MCG_WEBV core data, we further collected the most-viewed videos of all time from the predefined 15 YouTube categories as the testing data of Web video categorization. The testing dataset has 2,047 Web videos. In addition, we collect the related videos and the videos uploaded by the same users. There are 111,462 related videos, and 136,542 user videos. The original video category labels are treated as the ground truth. Altogether, there are 7,701 unique text words.

To verify performance, we tested near-duplicate video detection on CC_WEB_VIDEO dataset and video annotation on CC_WEB_VIDEO and its extension. We used the original MCG_WEBV dataset as the training data for Web video categorization and tested on the extension of MCG_WEBV.

Near-duplicate elimination performance

To evaluate the performance of near-duplicate video detection, we compared the novelty reranking results based on

- context information only, where the time duration, title, and tag context are evaluated separately;
- content information only, that is, the hierarchical method,[4] global signatures using a color histogram, and local-point-based approach with Hamming embedding and weak geometric consistency checking;[10] and
- the data-driven approach that integrates the content and context information.

The original ranking from the search engine serves as the baseline. We use novelty mean average precision (NMAP)[4] as the performance metric. The performance comparison up to the top 30 search results is illustrated in Figure 6.

The data-driven 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 novel videos can be directly detected simply based on time duration information. Verification by local-point-based detection on thumbnail images basically can guarantee satisfactory performance. Although the performance is slightly worse than hierarchical and weak geometric consistency (WGC) methods, the data-driven method is efficient. A huge number of comparisons are filtered out with the time duration information.

Furthermore, the data-driven method only compares the thumbnail images or the prototype keyframes of seed videos, thus avoiding comprehensive keyframe pair comparisons. From Figure 6, we can see that the average number of keyframe pair comparisons of content and context only has a linear increase, compared to the exponential grown using the hierarchical method. The speedup of the data-driven method is around 164 times faster than the hierarchical method when reranking the top 30 novel videos. Near-duplicate identity is easier and cheaper to achieve by context analysis. The data-driven framework achieves a balance between real-time processing and detection accuracy.

Annotation by scalable search performance

Since the average number of user-supplied tags for each video is around 4.6 in the dataset, we evaluate the performance of the top five tags in the experiment. We adopt precision at rank *k* as the performance metric to measure the proportion of retrieved tags that are relevant at rank *k*. The annotation results are 0.77. This means that around four out of the top five recommended tags are closely relevant to the video content, which exhibits satisfactory annotation accuracy. The good performance of near-duplicate video retrieval guarantees the high-quality tag annotation.

Video	Top five tag annotations
	simpson real life intro funny
	nerdy white al weird yankovic
	sony bravia ball commerce bounce
	anchorman delight afternoon ferrel funny
	numa gary brolsma numanumad dance

Figure 7. Examples for top five tag annotations.

The top five recommended tags of some examples are illustrated in Figure 7. We can see that the suggested tags are meaningful. For example, the second video belongs to "White and Nerdy." In addition to the common words "white" and "nerdy," the suggested tags provide new clues for this video, such as the author name "al yankovic," and his theme "weird al." The tags summarize the viewpoints from many users toward this set of similar videos. The third example also demonstrates interesting information, which is another commercial for a "Sony Bravia" TV set. Different from other Sony commercials, the principal content in this video is full of bouncing balls. Therefore, the suggested tags are constructive and help identify the major content of videos.

Web video categorization performance

We compare the categorization performance of the SVM classifier based on text feature (semantics), majority voting by related videos (relevance), majority voting by user videos (interest), and the fusion of these three sources (fusion). The SVM classifiers are trained using text features on the training set with a radial basis function RBF kernel. We use traditional precision to evaluate the individual performance and average precision to measure the overall performance, which is the average of the individual performance of 15 categories.

Figure 8 shows the performance comparison for Web video categorization. Because title and tags are noisy and ambiguous for Web applications, the general performance of semantics is poor. However, to some extent, text words have a sort of discriminative power. It's still a useful resource for classifying the videos. For certain categories such as "gaming," "science and technology," and "travel and events," our method achieves the best performance. Although the idea of majority voting from related videos and user videos is simple, it can provide a useful indication of the video categories. For most categories, this method demonstrates better performance than text classification. And the information from user videos is more useful for offering accurate hints through users' interests.

Semantics, relevance, and interest provide category confidence from different viewpoints. Their combination achieves better performance than the individual ones. Figure 5 shows an example using a user-uploaded video of a guitar lesson. Since the title and tags are ambiguous (for example, "hotel" and "California," the classifiers on semantics falsely classify this music-oriented video as "travel." Nevertheless, the majority voting from video relevance and user interest indicates that video relevance and the author "swanner" have a strong

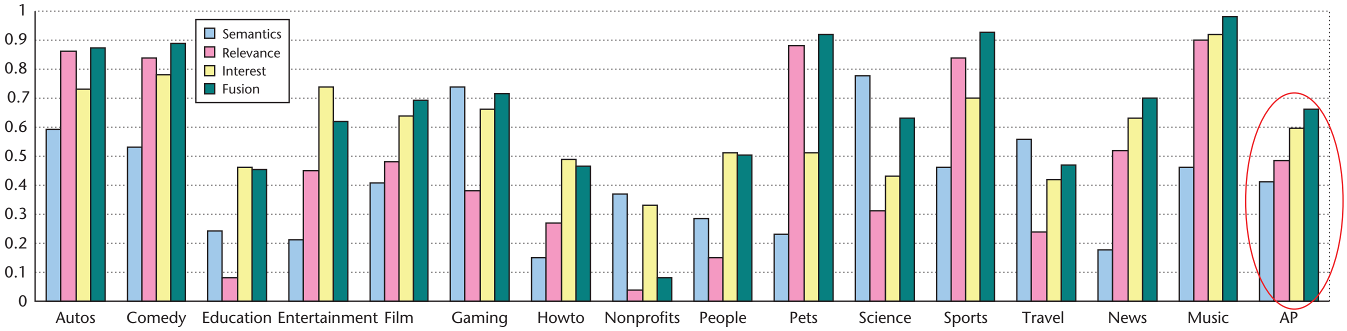


Figure 8. Performance comparison of 15 categories and overall performance.

Model-Based versus Data-Driven Approaches

Existing techniques in image and video annotation,[1,2] classification,[3-5] and high-level concept detection are dominated by model-based approaches, which depend heavily on the machine learning techniques (for example, SVMs and Gaussian mixture models) to train a large set of classifiers. There have been numerous efforts related to automatic tag suggestion or recommendation.[2,6,7] To bridge the semantic gap between low-level visual features and semantic concepts, many learning models[1] have been proposed to automatically assign keywords to videos or images and regions. Video classification takes place by building classifiers on textual, audio, and visual low-level features or their combination.[3-5] One comprehensive approach, demonstrated in Yang et al., treats Web video categorization by considering semantic modality (concept histogram and visual word representation) and text modality surrounding videos.[4] In X. Yuan et al., a video genre ontology is constructed and hierarchical SVM classifiers are designated to categorize video genres.[5] Unfortunately, considering the far-from-satisfactory performance of current, high-level concept detection and the cost of feature processing, these statistical models are still not acceptable for practical applications and are generally expensive to implement.

When facing the ever-increasing Web-scale video data, the state-of-the-art approaches encounter the following difficulties:

- A large amount of balanced labeled samples is often required for effective classifier learning. Nevertheless, the scarcity of training examples commonly exists in many applications. Collecting a large set of noise-free training examples with sufficient positive samples for learning is not always easy. Manual annotation of training examples can be laborious, and most labeling efforts are indeed spent in annotating negative examples.
- The extraction and selection of low-level features for generic concept learning remains an open issue. The discriminative power of features (for example, global features versus local features) often goes together with complexity in processing time, feature representation, and dimension.
- The size of vocabulary is huge and the meaning of concepts might change dynamically. In social media, for example, a word's meaning might evolve over time or change according to context. Novel words or phrases might emerge when new topics are being discussed. Learning classifiers in such a scenario is difficult to cope with and to completely model in the evolving Web environment.
- Web videos are diverse in visual quality and appearance. The types of videos can range from professional videos to low-quality home videos, making semantic annotation and categorization difficult.

Considering these challenges, building a set of classifiers that are scalable for concepts appearing in Web videos is beyond the state-of-the-art technologies. However, when the accumulation of video data reaches a certain degree, a quantity change will lead to a qualitative change. There are many examples of successful data-driven approaches. In one project, researchers measured the similarity between two words by querying the number of webpages containing words from Google.[8] In another project, researchers analyzed 80 million tiny images, approaching the object and scene recognition from the data-driven viewpoint.[9] When the underlying images are large enough, a simple scheme like nearest-neighbor matching can perform reasonably well. A similar idea was to annotate images first by

preference for music. Combining the recommendation from these three aspects, the video is correctly suggested as being in the ''music'' category, which compensates for the shortcoming of model-based classifiers.

Conclusion

In contrast to computationally demanding model-based techniques, data-driven solutions are simple to implement, and possess good scalability. Accompanying the opportunities and challenges induced by the continuous expansion of social media, data-driven approaches are expected to continue demonstrating promising performance and enlighten new directions for Web-scale applications. In our future research, we intend to explore more video applications based on data-driven approaches from different aspects, and combine content, context, and community information to facilitate Web video searching and mining. **MM**

Acknowledgments

The work described in this article was supported by the National Natural Science Foundation of China (NSFC grant no. 61071184), the Fundamental Research Funds for the Central Universities (project no. SWJTU09CX032 and SWJTU09ZT14); a grant from the Research Grants Council of Hong Kong Special Administrative Regions, China, (CityU 119508); and a grant from City University of Hong Kong (project no. 7002438).

discovering visually and semantically similar search results for annotation, and then identify salient terms from search results under a corpus with 2.4 million images.[7]

Such techniques, also referred to as *annotation by search*, have been demonstrated elsewhere for video annotation. In Moxley, Mei, and Manjunath, similar videos are ranked in a multimodal (auto correlogram, color moment, and edge distribution histogram) search, and graph reinforcement mining is proposed for propagating tags from similar documents to query videos.[6] Similarly, by exploiting visual content redundancy (overlapping and duplicated content), researchers used tag propagation to improve video annotation.[2] A neighbor voting algorithm is proposed in Li, Snoek, and Worring to predict the relevance of tags to images by accumulating votes from visual neighbors.[10] The algorithm is based on the idea that when different users label visually similar images with the same tags, these tags are likely to reflect objective views of the visual content.

In short, using community-contributed data, these works vividly demonstrate that data-driven approaches can mostly rely on simple approaches. Scalable solutions for various applications can be achieved especially with the aid of efficient data-indexing schemes. Hash encoding strategy is adopted in Wang et al. to map visual features into bitstreams, which significantly speeds up retrieval efficiency.[7] To implement real-time, near-duplicate clip detection, two schemes are proposed in Shen et al.[11] This work proposes a compact-representation model to summarize each video to a single vector, and then each video clip is mapped into a sequence of symbols.

References

1. G.-J. Qi et al., ''Correlative Multi-Label Video Annotation,'' *ACM Multimedia,* ACM Press, 2007, pp. 17-26.
2. S. Siersdorfer, J.S. Pedro, and M. Sanderson, ''Automatic Video Tagging using Content Redundancy,'' *Proc. Int'l ACM Sigir Conf. Research and Development in Information Retrieval,* ACM Press, 2009, pp. 395-402.
3. D. Brezeale and D.J. Cook, ''Automatic Video Classification: A Survey of the Literature,'' *IEEE Trans. Systems, Man, and Cybernetics,* vol. 38, no. 3, 2008, pp. 416-430.
4. L. Yang et al., ''Multi-Modality Web Video Categorization,'' *Proc. ACM Multimedia Information Retrieval,* ACM Press, 2007, pp. 265-274.
5. X. Yuan et al., ''Automatic Video Genre Categorization using Hierarchical SVM,'' *Proc. Int'l Conf. Image Processing,* IEEE Press, 2006, pp. 2905-2908.
6. E. Moxley, T. Mei, and B.S. Manjunath, ''Video Annotation through Search and Graph Reinforcement Mining,'' *IEEE Trans. Multimedia,* vol. 12, no. 3, 2010, pp. 184-193.
7. X.-J. Wang et al., ''Annotating Images by Mining Image Search Results,'' *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 11, 2008, pp. 1919-1932.
8. R.L. Cilibrasi and P.M.B. Vitanyi, ''The Google Similarity Distance,'' *IEEE Trans. Knowledge and Data Engineering,* vol. 19, no. 3, 2007, pp. 370-383.
9. A. Torralba, R. Fergus, and W.T. Freeman, ''80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition,'' *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 11, 2008, pp. 1958-1970.
10. X.-R. Li, C.G.M. Snoek, and M. Worring, ''Learning Social Tag Relevance by Neighbor Voting,'' *IEEE Trans. Multimedia,* vol. 11, no. 7, 2009, pp. 1310-1322.
11. H.-T. Shen et al., ''UQLIPS: A Real-Time Near-Duplicate Video Clip Detection System,'' *Proc. 33rd Conf. Very Large Data Bases,* ACM Press, pp. 1374-1377, 2007.

References

1. R.L. Cilibrasi and P.M.B. Vitanyi, ''The Google Similarity Distance,'' *IEEE Trans. Knowledge and Data Engineering,* vol. 19, no. 3, 2007, pp. 370-383.
2. A. Torralba, R. Fergus, and W.T. Freeman, ''80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition,'' *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 11, 2008, pp. 1958-1970.
3. E. Moxley, T. Mei, and B.S. Manjunath, ''Video Annotation through Search and Graph Reinforcement Mining,'' *IEEE Trans. Multimedia,* vol. 12, no. 3, 2010, pp. 184-193.
4. X. Wu, A.G. Hauptman, and C.W. Ngo, ''Practical Elimination of Near-Duplicates from Web Video Search,'' *ACM Multimedia,* ACM Press, 2007, pp. 218-227.
5. X. Wu et al., ''Real-Time Near-Duplicate Elimination for Web Video Search with Content and Context,'' *IEEE Trans. Multimedia,* vol. 11, no. 2, 2009, pp. 196-207.
6. H.-T. Shen et al., ''UQLIPS: A Real-Time Near-Duplicate Video Clip Detection System,'' *Proc. 33rd Very Large Data Bases,* ACM Press, pp. 1374-1377, 2007.
7. S. Poullot, M. Crucianu, and O. Buisson, ''Scalable Mining of Large Video Databases Using Copy Detection,'' *ACM Multimedia,* ACM Press, 2008, pp. 61-70.
8. X. Wu, W.-L. Zhao, and C.-W. Ngo, ''Towards Google Challenge: Combining Contextual and

Social Information for Web Video Categorization,'' *ACM Multimedia, Multimedia Grand Challenge,* ACM Press, 2009, pp. 1109-1110.

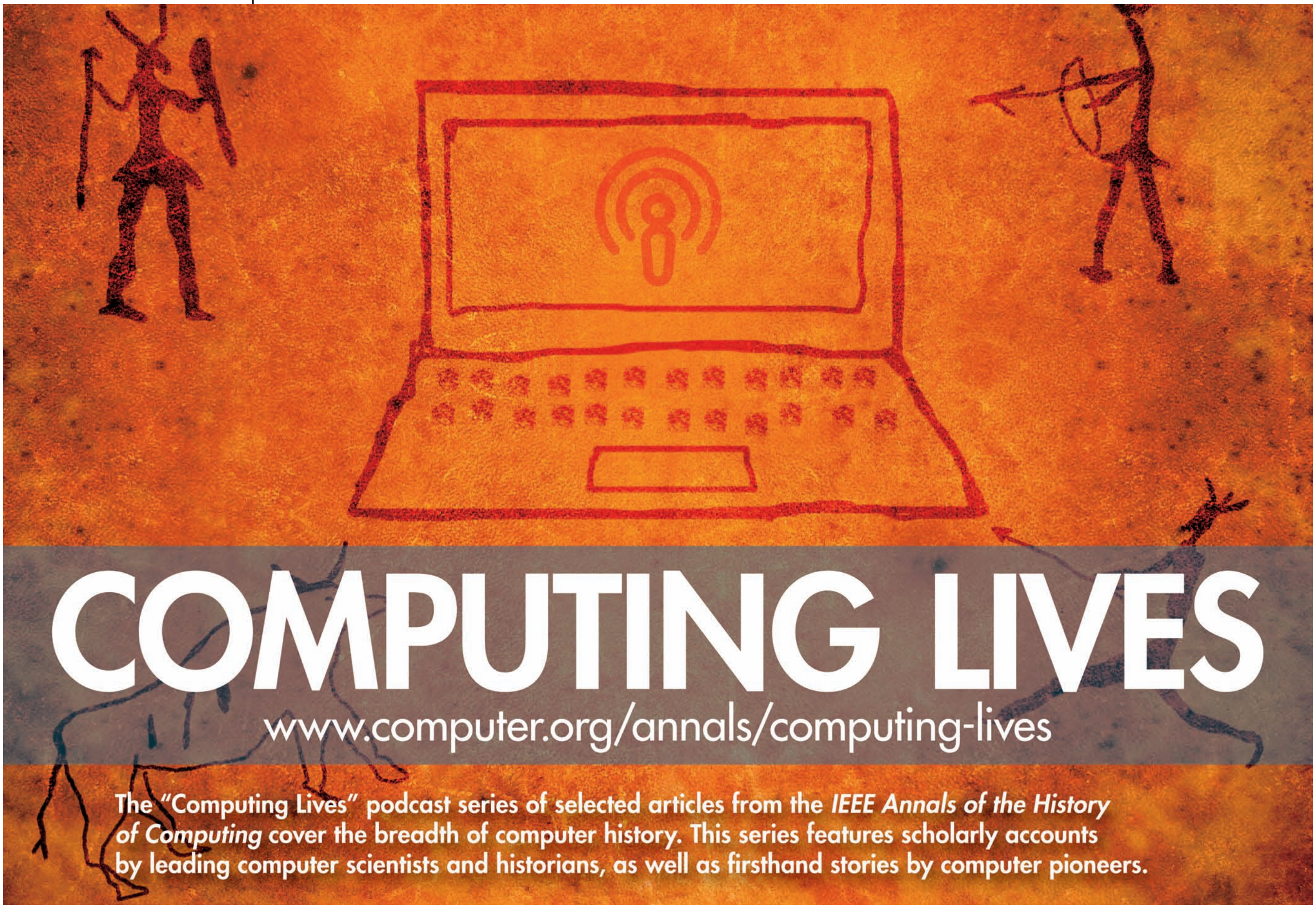
9. J. Cao et al., ''*MCG-WEBV: A Benchmark Dataset for Web Video Analysis,*'' tech. report ICT-MCG-09-001, Inst. Computing Tech, May. 2009; http://mcg.ict.ac.cn/chengguo1.html.
10. H. Jegou, M. Douze, and C. Schmid, ''Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search,'' *Proc. European Conf. Computer Vision,* Springer, 2008, pp. 304-317.

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

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

Wan-Lei Zhao is a senior research associate at City University of Hong Kong. His research interests include multimedia information retrieval and video processing. Zhao has a PhD in computer science from City University of Hong Kong. Contact him at wzhao2@cs.cityu.edu.hk.

cn *Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.*



IEEE computer society
Browse by Subject
MAGAZINES
TRANSACTIONS
The Ultimate Resource
Unlimited access to the IEEE Computer Society Digital Library!
IEEE CS DIGITAL LIBRARY
Digital Library RSS Feeds
The Berkian Institute of Technology