Video Retrieval Technology–From Content to Context (Postprint)

Authors: Cao Juan, Zhang Yongdong, Li Jintao

Date: 2017-03-10T00:00:00+00:00

Abstract

With the proliferation of video capture devices and the advent of Web2.0 technologies, video data on the Internet has experienced explosive growth. How to retrieve videos that meet user needs from large-scale video data is precisely the challenge that video retrieval technology seeks to address. This paper presents a comprehensive survey of video retrieval technology, focusing on content-based video copy detection technology, concept-based semantic video retrieval technology, and context information-based web video analysis technology. Additionally, this paper briefly reviews the research progress of our research group in video copy detection, semantic video detection, and web video analysis.

Full Text

Video Retrieval Technology: From Content to Context

Cao Juan, Zhang Yongdong, Li Jintao

Abstract

With the proliferation of video capture devices and the emergence of Web 2.0 technologies, video data on the Internet has grown rapidly. How to retrieve videos that users need from large-scale video data is precisely the problem that video retrieval technology aims to solve. This paper provides a comprehensive review of video retrieval technology, focusing on three main areas: content-based video copy detection technology, concept-based semantic video retrieval technology, and context-based web video analysis technology. Additionally, this paper briefly introduces the research progress of our research group in video copy detection, semantic video detection, and web video analysis.

Keywords: Visual features, Context information, Copy detection, Video retrieval, Video topic discovery and recommendation

1 Introduction

With the development of multimedia and network technologies, video has become one of the primary carriers for publishing and acquiring information in daily life. In September 2009, the popular video-sharing website YouTube saw approximately 20 hours of new video uploaded every minute. According to a report from the China Internet Network Information Center, the number of online video users in China reached 284 million in 2010, accounting for 62.1% of all internet users. Faced with this explosive growth in online video content and users, efficient video retrieval technology is urgently needed to help people quickly and accurately find desired video content within massive-scale web video data.

Video retrieval technology manifests differently at various levels. As shown in [Figure 1: see original paper], from a technical perspective, video retrieval encompasses core modules such as video structuring, feature extraction, high-dimensional indexing, similarity computation, and result ranking. For service providers, video retrieval can be categorized based on application modes into general video retrieval, specific video retrieval, and proactive video recommendation. For end users, video retrieval can be divided based on query input types into text keyword-based, mid-level semantic concept-based, and video example-based retrieval, as well as query-free proactive video recommendation technologies.

Current commercial video search engines such as Baidu, Google Video, and Blinkx primarily rely on text retrieval techniques. They extract information from video metadata—including titles, descriptions, tags, subtitle text, and speech recognition transcripts—to perform text-based video retrieval, with text keywords serving as the user query interface. However, the retrieval performance of these methods degrades significantly when video text is missing (e.g., in home videos) or when the text fails to accurately describe the video content (e.g., incorrect text tags).

Consequently, Content-Based Video Retrieval (CBVR) technology emerged in the 1990s [**?**][**?**][**?**]. These methods directly extract low-level visual features from videos for indexing and similarity computation, supporting example-based retrieval and sketch-based retrieval. Currently, content-based video retrieval methods cannot yet be applied to general-purpose video retrieval and are only used in small-scale experimental systems, such as IBM' s QBIC retrieval system [**?**], the JACOB system developed by the University of Palermo [**?**], Columbia University' s VideoQ video query system [**?**], and the web video search engine WebSEEk [**?**]. Notably, in certain specific domains, content-based video retrieval has demonstrated significant practical value, such as in detecting illegal copy videos for copyright protection, detecting duplicate videos in large-scale web video collections, and detecting specific semantic events in surveillance videos. Taking video copy detection as an example, due to its important application demands and value, the international video retrieval evaluation cam-

paign TRECVID [?] organized by the U.S. National Institute of Standards and Technology established a "video copy detection evaluation" task starting in 2008. Through annual evaluations, this field has achieved breakthrough progress [?][?].

The fundamental challenge facing content-based video retrieval is the "semantic gap." Smeulders et al. [?] defined this problem as "the lack of a one-to-one correspondence between low-level features extracted by machines from videos and high-level semantics understood by users." To bridge this semantic gap, a promising research direction has emerged in the multimedia field in recent years—concept-based video retrieval [?][?][?]. These methods introduce an intermediate semantic concept layer between low-level feature descriptions and user semantic queries. The concepts in this layer possess certain semantics while being trainable as concept detectors based on low-level features through automatic machine recognition, such as object class concepts (person, airplane, mountain, road, boat, building, etc.), scene class concepts (indoor/outdoor, waterscape, snowscape, desert, etc.), and event class concepts (takeoff, sports, walking, etc.). By establishing two layers of mapping—from low-level features to semantic concepts (i.e., semantic concept detection) and from user queries to semantic concepts (i.e., query analysis)—concept-based semantic video retrieval is ultimately realized. Results from the past three years of TRECVID evaluations reveal that this method's performance far exceeds that of purely text-based or purely visual-based video retrieval methods.

In recent years, with the development of Web 2.0 technologies, most video data is primarily stored and disseminated through network platforms such as YouTube, Tudou, and Youku. These platforms provide a network environment for user interaction around video data (social network). Beyond the relevance of the video content itself, rich contextual information also establishes connections between videos. For instance, videos uploaded by the same author share certain similarities, and videos commented on by the same user also exhibit certain associations. R. Jain [?] and X. Jin [?] strongly advocated for using contextual information for multimedia content analysis in their "brave new idea" presentations at ACM Multimedia 2010. First, content without context is meaningless [?]. For example, different users may have different understandings and annotations of the same video, and even the same user may have different understandings of the same video at different times. Second, rich contextual information holds important practical significance for overcoming problems such as sparse features and high noise in network multimedia data. Consequently, over the past two years, contextual information has gradually attracted the attention of multimedia researchers, with exploratory work emerging in areas such as image recommendation and prediction [?][?], video classification [?], and video topic discovery [?][?].

In summary, the evolution of video retrieval technology has progressed from text, to visual content, to semantic concepts, and finally to contextual information. Since text-based video retrieval primarily employs existing text information retrieval techniques, this paper will not elaborate on it further. In subsequent

sections, we will introduce the current state of video retrieval research using content-based video copy detection technology, concept-based video retrieval technology, and context-based web video topic mining and recommendation technology as examples. Finally, this paper will also provide a brief introduction to the progress achieved by our research group in related areas.

2 Content-Based Video Copy Detection Technology

The U.S. National Institute of Standards and Technology defines a video copy as: a video or its segment that, after certain editing processes, yields a homologous video version with identical content but not completely consistent visual appearance (e.g., brightness) [**?**]. Video copy detection refers to the process of matching content features of a query video with videos in a database to determine whether the query video is a copy of any source video in the database. Unlike general video retrieval, copied videos undergo various geometric and image transformations based on the source video, causing varying degrees of visual changes known as copy attacks [**?**]. Common copy attacks include encoding format conversion, frame size changes, aspect ratio changes, and border addition.

Based on the features used, existing content-based copy detection methods can be categorized into three major types [**?**]: digital signature-based methods, keyframe-based methods, and trajectory-based methods.

Digital signature-based methods typically represent entire video content as a global feature value for rapid video-level matching. For example, they average color histograms [**?**] or ordinal features [**?**] from all frames in a video. Literature [**?**] has demonstrated that such methods are only effective against minor copy attacks. Moreover, because they ignore temporal information in videos, they can only detect copies of entire videos and cannot identify partial segment copies.

Keyframe-based methods sample representative frames from videos for matching. The core algorithm involves how to match two unequal-length keyframe sequences. For instance, C.Y. Chiu et al. [**?**] used dynamic programming algorithms to select the longest matching sequence; X. Wu et al. [**?**] employed sliding window methods for keyframe sequence matching; and Hung-Khoon Tan et al. [**?**] represented frame temporal relationships as a directed-edge temporal network, performing visual-temporal consistency verification based on frame-level matching results to accurately detect and locate video copy segments. These methods utilize temporal relationships that offer some robustness to visual changes but are highly sensitive to temporal variations such as segment swapping, frame rate changes, and frame loss.

Trajectory-based methods track interest point changes across video sequences to form spatio-temporal trajectory features. For example, J. Law-To et al. [**?**] used trajectory features to annotate different motion behaviors, while X. Wu et al. [**?**] adopted a bag-of-trajectories approach to address discontinuous temporal pattern problems. Trajectory features simultaneously consider spatial and temporal changes of interest points, demonstrating robustness against complex

copy attacks. However, because extracting interest points and trajectories is extremely time-consuming, these methods have relatively high computational complexity.

3 Concept-Based Semantic Video Retrieval Technology

The concept-based video retrieval framework is illustrated in [Figure 2: see original paper] and comprises three key steps: first, establishing a semantic concept set—determining how many concepts to include and which concepts to select for constructing the semantic space; second, establishing the mapping relationship from low-level features to semantic concepts, i.e., semantic concept detection; and third, establishing the mapping from user queries to semantic concepts, i.e., query analysis. Below we introduce the current research status in each of these three aspects.

3.1 Semantic Concept Set Construction

The first step in concept-based video retrieval is defining an appropriate semantic concept set. Currently, the most widely recognized concept sets include LSCOM (Large-Scale Concept Ontology for Multimedia) [**?**] and Mediamill-101 [**?**]. LSCOM was jointly developed by IBM' s Watson Research Center, Carnegie Mellon University (CMU), and Columbia University, containing definitions for 2,000 semantic concepts and providing manual annotations for 449 concepts on the TRECVID 2005 video dataset, offering an important dataset for multimedia retrieval method research. Building upon LSCOM, researchers further selected 44 concepts to form the LSCOM-lite lexicon, dividing the semantic space into seven mutually orthogonal subspaces: objects, activities, events, scenes/locations, people, graphics, and program, selecting appropriate concepts for each subspace based on concept word usage in queries.

Mediamill-101 was developed by the University of Amsterdam and manually annotated on the TRECVID 2005 video dataset. Based on these semantic concept lexicons, A. Hauptmann et al. from Carnegie Mellon University conducted a series of foundational studies on semantic concept set construction [**?**], reaching an important conclusion: when the concept set scale is around 5,000 and each concept' s detection accuracy is no lower than 10%, concept-based video retrieval can achieve performance comparable to text retrieval (MAP[1]=65%). This conclusion laid the foundation for subsequent concept-based video retrieval technology development.

In 2008, Y. J. Lu et al. from the University of Texas further proposed that different concepts have different semantic gap sizes [**?**]. For example, the concept "Sunset" can be easily described using visual features and has a small semantic gap, while "Europe" is difficult to describe with simple visual features and has a large semantic gap. Concepts with smaller semantic gaps are relatively easier to implement through automatic machine detection based on low-level features

[1]Mean Average Precision, system average accuracy

and are suitable for constructing semantic concept dictionaries. Based on this theory, they first proposed quantifying the semantic gap to automatically select concepts with the smallest semantic gaps for building semantic concept sets. This method requires no manual intervention and possesses strong operability.

The establishment of these large-scale, standardized, annotated semantic concept datasets and the continuous improvement of construction theories are significant for enhancing video retrieval accuracy and standardizing video retrieval evaluations.

3.2 Semantic Concept Detection

For each defined concept, a concept detector must be established through machine learning methods from annotated positive and negative samples. Over the past decade, semantic concept detection for videos/images has been extensively studied [**?**]. Its core modules include: low-level video feature extraction, learning models, and multi-modal feature fusion.

First, effective feature representation is key to successful concept detection. Color (e.g., color histograms, color moments) and texture features (e.g., wavelet textures) are two types of visual features commonly used in computer vision. Compared with these global features that describe overall video/image distribution characteristics, local features demonstrate robustness to geometric and illumination changes in images and have shown outstanding performance in many visual classification tasks in recent years. Local feature extraction includes two parts: local feature point detection and description. Currently, widely adopted feature point detection methods include the Harris corner detection algorithm [**?**] and the Difference of Gaussian (DoG) local feature detection method [**?**], among others. Description methods include the Scale-Invariant Feature Transform (SIFT) local feature descriptor proposed by Lowe [**?**]. Detailed reviews of local feature point detection and description can be found in literature [**?**] and [**?**]. Due to the large number of local feature points extracted from each image, they cannot be directly used to describe visual content. A typical local feature usage method is visual vocabulary (visual vocabulary). First, local feature points are clustered into visual words to generate a visual dictionary. Second, each image' s local feature points are mapped to the visual dictionary to obtain a bag-of-visual-words (BoW) representation [**?**] for each image.

Based on these features, a classifier can be learned for each concept. Currently, the most widely used learning model in concept detection tasks is Support Vector Machine (SVM) [**?**]. Other models include Gaussian Mixture Models [**?**] and Hidden Markov Models [**?**]. The aforementioned methods all build models for individual semantic concepts. However, semantic concepts in videos do not exist independently; different semantic concepts often have contextual constraints or co-occurrence relationships. For example, detecting "sky" and "green field" increases the probability of detecting "landscape" while decreasing the probability of detecting "indoor." Therefore, research is needed to utilize

relationships between different concepts to enhance or exclude certain concepts. In response to this phenomenon, scholars have proposed multi-concept-based video semantic representation methods. Representative approaches include: the method proposed by Hauptmann et al. [?] that uses logistic regression to obtain inter-concept relationships for multi-concept fusion; and the method based on Bayesian Dirichlet Metric and neural networks proposed by Tsinghua University in TRECVID 2007 [?].

Multi-modal fusion methods include early fusion and late fusion. Early fusion refers to combining various features into a long feature vector and training a single concept detector based on this vector. Late fusion refers to training a separate concept detector for each feature and fusing the output results of each detector as the final detection result. Each approach has its pros and cons. The former implicitly considers complementary relationships between different features but faces the challenge of high-dimensional feature processing. The latter is relatively easier to implement and has been adopted by many concept detection systems, but how to weight multiple detectors is key. Snoek et al. [?] conducted a comparative study of these two fusion strategies.

3.3 Concept Mapping

After constructing classifiers for each semantic concept using the aforementioned methods, concept-based video retrieval can be realized by mapping user queries to relevant concept detectors. Based on the features used, this retrieval can be categorized into: text feature-based mapping, visual content-based mapping, and feedback result-based mapping.

Since text is the most direct description of video semantic content, most current systems use text features to map queries to semantic concepts [?][?][?][?]. One approach is ontology-based concept mapping, such as using WordNet [?]. These knowledge bases typically contain relational structures between words, such as hypernymy, hyponymy, and synonymy relationships, as well as semantic similarity measurement algorithms between words, such as RES [?], Lesk [?], WUP [?], JCN [?], and the recently proposed OSS [?] method. Based on these measurement methods, mapping from query keywords to semantic concepts can be achieved. Another approach is data-driven concept mapping. These methods use statistical models, such as Latent Semantic Indexing [?], to analyze co-occurrence patterns among various terms in the database, thereby automatically mining correlations between terms.

In addition to query text, queries are sometimes provided in visual forms such as image examples or video clips. Therefore, mapping from queries to concepts can also be accomplished based on these visual contents. The general process is: using the concept detectors introduced in the previous section to perform corresponding concept detection on query examples, then directly selecting concepts with high posterior probabilities as the concept mapping results for the query [?]. Since this method is highly sensitive to the detection accuracy of concept detec-

tors—once a detector misclassifies a query example, it directly leads to concept mapping errors—researchers have proposed using these noisy concept detection results as features for further statistical analysis [?] or machine learning [?] to obtain more stable concept mapping results.

Compared with statistical analysis based on the entire dataset, researchers believe that statistics on correlations between queries and concepts within a specific subset relevant to the query are more valuable. Typically, this query-relevant specific subset needs to be generated through user annotation, known as Relevance Feedback. This method extracts features from user-annotated sets and maps queries to semantic concepts using the aforementioned text feature-based or visual feature-based methods. To reduce user involvement, some systems simplify the annotation process by assuming the top N results in initial retrieval results as positive samples and the bottom M results as negative samples, known as Pseudo-Relevance Feedback. Since pseudo-relevance feedback methods rely on initial retrieval results, they can degrade retrieval performance when initial results are poor [?].

4 Context-Based Web Video Analysis Technology

Context refers to the conditions and environment on which an object' s existence or occurrence depends [?]. Only by considering contextual information can we correctly understand the semantic content contained in videos and effectively narrow the semantic gap. Simultaneously, contextual information can effectively reduce the search space and improve retrieval performance. For example, a video shot in Australia is unlikely to contain snow scenes. Specifically for web videos, we can categorize context into video-centered context, including video attributes such as length and category; camera parameters such as model and resolution; shooting environment such as location and time; and user-centered context, such as social networks generated by user behaviors like annotation, commenting, and favoriting.

R. Jain et al. [?] used camera EXIF parameters for image classification, achieving better results than content-based methods. X. Wu et al. [?] improved near-duplicate video detection accuracy by considering video duration information. With the popularization of GPS devices, the value of video geographic information has been continuously discovered. Literature [?] and [?] respectively proposed a GeoFolk framework and a Latent Geographical Topic Analysis (LGTA) method to discover region-specific video topics based on video geographic information for comparing topic development across different regions and analyzing hot topics in different areas.

On the other hand, contextual information generated by network user behaviors contains rich statistical knowledge [?]. F. Benevenuto et al. [?] conducted in-depth analysis of YouTube users' video response behaviors, obtaining multiple valuable statistical models for subsequent web video analysis. V. Z. Roelof et al. [?] predicted each user' s favorite photos based on subscription behaviors,

with experiments on a Flickr dataset showing that context-based prediction accuracy (92%) was higher than both text-based (87%) and visual-based (88%) methods. X. Wu et al. performed automatic video categorization by voting based on category information of related videos provided by YouTube. L. Gou et al. [**?**] proposed a Social Network Document Rank (SNDocRank) algorithm that calculates correlations between the query user' s network and the video author' s network to rank video retrieval results, achieving video retrieval more aligned with user interests.

5 Our Work

Over the past three years, our research group has made significant progress in video content analysis research and developed multiple systems. This section focuses on introducing our research advances and related systems in three areas: large-scale web video copy detection, concept-based web video retrieval, and context-based web video topic discovery and retrieval.

5.1 Large-Scale Web Video Copy Detection System

In video copy detection, we have proposed various single-frame-based and video-based copy detection features with the goals of improving detection accuracy and efficiency, and attempted to enhance retrieval efficiency through high-dimensional indexing techniques and GPU acceleration. Our developed video copy detection system achieved third place overall in the 2008 TRECVID video copy detection task and first place in 2009 [**?**].

5.1.1 Robust Visual Feature Mining for Complex Attacks Various complex video copy attacks impose stringent requirements on visual features. We proposed a theory and method for extracting highly robust visual features through sample automatic expansion and stable feature mining [**?**]. This method introduces the concept of full affine space by automatically simulating image affine deformations under different viewpoints, expanding original features into a collection of local features detected under various affine conditions. Second, to identify the most stable representative features from this large amount of expanded information, we employ a global stability-based stable feature mining method to obtain a collection of highly robust local features for each image, using only 5% of the expanded information as the visual information representation under various complex attack modes.

Unlike ordinary image/video retrieval, copied images/videos have undergone copy attack processing. How to measure similarity between such transformed images is the core problem of copy detection. In recent work, we proposed a geometric consistency measurement method based on matching pairs [**?**]. Unlike traditional methods that directly calculate similarity between two matched keypoints, this method measures similarity between geometric transformations of pairwise matching pairs to assess geometric consistency between two images.

The more matching pairs with similar transformations, the more likely the two images are copies. This method' s advantages include its ability to handle both global copy attacks (such as scaling, rotation, and translation) and local transformations, as well as certain degrees of perspective distortion. Additionally, it can simultaneously handle multiple visual pattern transformations existing in a single image.

5.1.2 High-Dimensional Feature Indexing for High-Speed Matching Due to the quantity and dimensionality of local features far exceeding the capacity of traditional matching methods, establishing effective high-dimensional indexing for features is a necessary component for large-scale web video copy detection. We proposed a data distribution-oriented Locality Sensitive Hashing (LSH) high-dimensional indexing method [**?**] that utilizes data distribution information to select projection vectors—obtaining projection vectors through unsupervised learning methods. Simultaneously, to intuitively analyze hash function performance, we introduced the concept of data distribution entropy. By evaluating data distribution entropy, superior hash functions are selected.

The resulting hash functions, while preserving the original data' s neighbor relationships as much as possible, make data indexed in each hash table entry more uniform. Validation on a famous open database demonstrates that under the same precision, our indexing algorithm reduces memory consumption by 30% compared to the original LSH algorithm. Meanwhile, when using the same number of hash tables, both query precision and efficiency are improved.

5.2 Concept-Based Semantic Video Retrieval System

Our research group has been engaged in concept-based general video retrieval research since 2007 and has achieved important results. Our developed latent semantic concept-based video retrieval system achieved second place in the automatic search task at TRECVID 2007 [**?**], first place in 2008 [**?**], and second place in the interactive search task in 2009 [**?**]. Below we introduce two concept selection algorithms and two latent and explicit semantic fusion algorithms.

5.2.1 Multi-Modal Concept Selection Methods Beyond considering semantic similarity between queries and concepts, different concepts play different roles in retrieval. For example, for the query "Find shots of one or more people at a table or desk, with a computer visible," although concepts "Face" and "Person" are highly relevant to the query, they lack discriminative power between positive and negative samples due to similar distributions. On the other hand, concepts "Computer" and "Hand" are relevant to the query and possess strong discriminative power, but their automatic detection accuracy is low, contributing little to retrieval. Based on this analysis, we proposed a Distribution-Based Concept Selection (DBCS) method [**?**] that selects the most valuable concepts for queries by fusing concept detector confidence with the discriminability of concept distributions in relevant and irrelevant sets.

To address incomplete query text descriptions, we further proposed a multi-modal concept association graph-based concept selection model [**?**] that represents relationships between queries and concepts as a network association graph, including similarity relationships between queries and concepts and between concepts themselves, while supporting multi-modal mapping from query examples and query text to semantic concepts. Through manifold ranking algorithms, multi-modal similarity between queries and concepts is propagated throughout the association graph until the network reaches a stable state, thereby selecting the top N concepts with highest similarity for video retrieval. Compared with various star-structure-based concept video retrieval methods, this method demonstrates stronger robustness for queries with sparse text descriptions, improving average precision by nearly 20%.

5.2.2 Explicit Semantic Concept and Implicit Semantic Concept Fusion for Video Retrieval Current concept-based video retrieval requires manually defining a limited concept set (called explicit semantics in this paper). Since this concept set cannot cover the entire query semantic space, zero-probability mapping and non-scalability problems occur during actual retrieval. Second, learning concept detectors requires manually annotating large amounts of training data, which is time-consuming and labor-intensive. Therefore, researchers began exploring new solutions, attempting to extract implicit topics (called implicit semantics in this paper) unsupervised from low-level video features through probabilistic topic models. We proposed a semantic video retrieval framework combining implicit and explicit semantics [**?**] that extracts stable, specific implicit semantics from low-level feature descriptions through Latent Dirichlet Allocation (LDA) models, while mapping user queries to manually defined explicit semantic concept sets based on the aforementioned concept selection algorithm, fusing both types of concepts to achieve video retrieval. The data-driven nature of implicit semantics compensates for zero-probability mapping problems in explicit semantic retrieval, improving retrieval recall, while accurate mapping to a fixed explicit semantic concept set ensures retrieval precision. Building upon this, we further proposed a bipartite graph-based fusion algorithm [**?**] that adaptively weights the two types of concepts according to different queries.

5.3 Context-Based Large-Scale Web Video Analysis

We have achieved important progress in automatic topic discovery and recommendation for large-scale web videos based on context, as well as exploratory results in video retrieval using multiple contextual information sources. Below we introduce these two aspects.

5.3.1 Trajectory-Based Web Video Topic Discovery and Recommendation According to YouTube Report 2009 [**?**], 45% of users log into YouTube without explicit search goals, instead browsing "hot videos" and "hot topics"

proactively recommended by the website, indicating that this query-free automatic video topic discovery and recommendation model is increasingly welcomed by web users. To improve the reliability of web video features, we proposed a global trajectory feature-based web video topic detection method [**?**]. First, each tag is represented as a feature trajectory on a timeline, extracting only salient points (vertices in trajectories) from trajectories for clustering to generate events occurring at that time point. This context-considering trajectory feature effectively filters noise. Second, an event development link graph is constructed by calculating text similarity and visual copy detection similarity between events. By finding optimal paths on this graph, the top N hottest topic trajectories are extracted. This method determines whether these events constitute a topic by considering global link conditions, thus demonstrating strong robustness to local incorrect links. Additionally, this trajectory-based topic discovery method can identify not only content-hot topics but also evolution-hot topics—typically controversial content on the internet that is repeatedly discussed over a period—and potential-hot topics, which are currently only of interest to minority groups but show continuous development trends and may erupt at subsequent points. The latter two topic types cannot be discovered by traditional content-based methods but hold significant importance in network supervision. Integrating these methods, we implemented a trajectory-based web video topic automatic discovery and visualization system [**?**] with excellent user experience.

5.3.2 Social Information-Based Web Video Retrieval Different contextual information sources have certain correlations. For example, different videos uploaded by the same author are more likely to be commented on by the same users. However, existing methods mostly focus on studying a single information source. We proposed a community structure-based video retrieval re-ranking method [**?**] that formalizes various link relationships between users, between videos, and between users and videos as a heterogeneous contextual network. By extracting implicit community structures from this network, stable association patterns among multiple contextual information sources are mined to achieve video retrieval result re-ranking based on community structure. Experimental comparisons on a heterogeneous network containing 82,352 YouTube videos and 39,555 users demonstrate that this method' s retrieval results outperform both pure text-based and pure visual-based methods.

6 Conclusion

Content-based video retrieval technology has experienced nearly a decade of development. Although important progress has been made in certain specific domains such as video copy detection, current technical levels still cannot meet users' needs for content-based retrieval of general videos. The primary technical bottleneck is the semantic gap problem. Consequently, current commercial general video retrieval still mainly relies on text information retrieval technology. However, with the popularization of video websites, web video data with rich contextual information has become the primary object of video retrieval. This

contextual information provides a possibility to bypass complex video content itself and narrow the semantic gap from the perspective of the video's contextual environment, making context-based web video analysis and retrieval a research hotspot in today' s network multimedia era. Simultaneously, we believe that the combination of video retrieval technology and web video data will give rise to richer network multimedia applications.

References

[1] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State-of-the-art and challenges. In ACM Transactions on Multimedia Computing, Communication, and Applications, 2006.

[2] C.W. Ngo, H.J. Zhang, and T.C. Pone, Recent Advances in Content Based Video Analysis, International Journal of Image and Graphics, December 2001.

[3] N. Dimitrova, H.J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, Applications of Video-Content Analysis and Retrieval, IEEE Multimedia, vol. 9, no.3, pp. 42-55, 2002.

[4] M. Flickner, H. S. Sawhney, et al .Query by image and video content: the QBIC system. IEEE computer, 28(9):23-32, 1995.

[5] L. Marco, A. Edoardo, JACOB: Just a content-based query system for video databases. Proc. ICASSP , Atlanta, G A, 1996

[6] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, VideoQ: An Automatic Content-Based Video Search System Using Visual Cues, ACM Multimedia, Seattle, WA, November 1997.

[7] J. R. Smith and S. F. Chang, Visually searching the web for content. IEEE Multimedia , pp.12-20, 1997.

[8] A. W. M. Smeulders, M.Worring, S. Santini, A. Gupta, and R. Jain, Content based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 1349-1380, 2000.

[9] X. Wu, M Takimoto, J Adachi. Scene duplicate detection based on the pattern of discontinuities in feature point trajectories. ACM international conference on Multimedia, pp.51-60, 2008

[10] C. G. Snoek and M. Worring. Concept-based video retrieval. Foundations and Trends in Information Retrieval, 2(4):215-322, 2009.

[11] A. Hauptmann, Y. Rong, W.H. Lin, M. Christel, H.Wactlar, Can High-Level Concepts Fill the Semantic Gap in Video Retrieval? A Case Study With Broadcast News, IEEE Transactions on Multimedia, 9(5): 958-966, 2007

[12] A. Natsev, A. Haubold, J. Te˘si´c, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval: A comparative review and new approaches. In ACM Multimedia (ACM MM), 2007.

[13] A. F. Smeaton, P. Over, and W. Kraaij, Evaluation campaigns and TRECVID, in Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 321–330, 2006.

[14] L. S. Kennedy, Revision of LSCOM event/activity annotations, Technical Report 221-2006-7, Columbia University ADVENT Technical Report, 2006.

[15] C. G. M. Snoek, M. Worring, J. C. v. Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In ACM Multimedia, 2006.

[16] R. Jain and P. Sinha. Content without context is meaningless. In ACM Multimedia (ACM MM), pp.1259-1268, 2010.

[17] X. Jin, A. Gallagher, L. L. Cao, J. B. Luo, and J. W. Han. The wisdom of social multimedia: using flickr for prediction and forecast. In ACM Multimedia (ACM MM ' 10), pp.1235-1244, 2010.

[18] V. Z. Roelof, R. Adam, and G. P. Lluis. Prediction of favourite photos using social, visual, and textual signals. In ACM Multimedia, pp.1015-1018, 2010.

[19] S. Sizov. Geofolk: Latent spatial semantics in web 2.0 social media. In ACM International Conference on Web Search and Data Mining, 2010.

[20] X. Wu, W. L. Zhao, and C. -W. Ngo. 2009. Towards Google challenge: combining contextual and social information for web video categorization. In ACM Multimedia (MM ' 09), pp.1109-1110, 2009.

[21] R. R. Ji, X. Xie, H. X. Yao, W. Y. Ma: Mining city landmarks from blogs by graph modeling. ACM Multimedia, pp.105-114, 2009.

[22] YouTube report 2009, http://youtubereport2009.com/.

[23] L. Gou, H. H. Chen, J. H. Kim, X. Zhang, SNDocRank: a Social Network-Based Video Search Ranking Framework. In ACM MIR ,2010.

[24] X. Wu, C.-W. Ngo, A. G. Hauptmann and H. K. Tan. Real-Time Near-Duplicate Elimination for Web Video Search with Content and Context. IEEE Transactions on Multimedia, 11(2): 196-207, 2009.

[25] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross. Video interactions in online video social networks. ACM Trans. Multimedia Comput. Commun. Appl., 5:30:1-30:25, 2009.

[26] Z. J. Yin, L. L. Cao, J. W. Han, C. X. Zhai, and T. Huang. Geographical topic discovery and comparison. In international conference on World Wide Web, pp. 247-256, 2011.

[27] S. Tang, Y. Zhang, J. Li, J. Cao, H. Luan, Q. He, and X. Zhang, TRECVID 2007 Search Tasks by NUS-ICT, In NIST TRECVID Video Retrieval Workshop, 2007.

[28] Y. J. Lu, L. Zhang, Q. Tian, W. Y. Ma, What Are the High-Level Concepts with Small Semantic Gaps, CVPR ,2008.

[29] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004.

[30] C. Harris, M. Stephens. A combined corner and edge detector, In. Proceeding of 4th Alvey Vision Conference, pp.147-151, 1988.

[31] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool. A comparison of affine region detectors, In IJCV , 65(1/2):43-72, 2005.

[32] K.Mikolajczyk, and C.Schmid. A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(10):1615–1630, 2005.

[33] J. Sivic, A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos, International Conference on Computer Vision, 2003.

[34] J. Tesic, A. Natssev, and J.R. Smith. Cluster-based data modeling for semantic video search. ACM International Conference on Image and Video Retrieval (ACM CIVR), 2007.

[35] A. Amir, W.H., G. Iyengar, C.-Y.Lin, M. Naphade, A. Natsev, C. Neti, H. J. Nock, J. R. Smith, B. L. Tseng, Y. Wu, and D. Zhang. IBM research TRECVID-2003 video retrieval system. in NIST TRECVID, 2003.

[36] B. Pytlik, A.G., D. Karakos, and S. Khudanpur. Trecvid 2005 experiment at johns hopkins university: Using hidden markov models for video retrieval. in NIST TRECVID, 2005.

[37] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In ACM Multimedia (ACM MM). 399-402, 2005.

[38] A. Hauptmann, M.-Y.C. et al., Confounded expectations: Informedia at TRECVID 2004, NIST TRECVID Workshop, 2004.

[39] J. Yuan, Z. Guo, THU and ICRC at TRECVID 2007, NIST TRECVID 2007 Workshop, USA , Nov. 2007.

[40] Y. G. Jiang, J. Yang, C. W. Ngo, and A. G. Hauptmann. Representations of Keypoint-Based semantic concept detection: A comprehensive study, IEEE Transactions on Multimedia, 12(1): 42-53, 2010.

[41] A. Natsev, A. Haubold, J. Te˘si´c, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval: A comparative review and new approaches. In ACM Multimedia (ACM MM), 2007.

[42] M. Campbell, A. Haubold, M. Liu, A. P. Natsev, J. R. Smith, J. Tešić, L. Xie, R. Yan and J. Yang, IBM Research TRECVID-2007 Video Retrieval System, In NIST TRECVID Video Retrieval Workshop, 2007.

[43] T. Mei , X. Hua and et al. MSRA-USTC-SJTU at TRECVID 2007: High-level feature extraction and search, In NIST TRECVID Video Retrieval Workshop. 2007.

[44] X.-Y. Wei, C.-W. Ngo, Y.-G. Jiang, Selection of Concept Detectors for Video Search by Ontology-Enriched Semantic Spaces, IEEE Transaction on Multimedia, Vol. 10, no. 6, 2008.

[45] C. Fellbaum and Ed. WordNet: an electronic lexical database. The MIT Press, 1998.

[46] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy. In IJCAL, pp.448-453, 1995.

[47] Z. Wu and M. Palmer. Verb semantic and lexical selection. In Annual Meeting of the ACL, pp.133–138, 1994.

[48] J. J. Jiang. Semantic similarity based on corpus statistics and lexical taxonomy. In ROCLING, 1997.

[49] T. Hofmann. Probabilistic latent semantic indexing. In Proc. of the 22nd Intl. ACM SIGIR conference, pp. 50–57, 1999.

[50] M. E. Lesk, Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, SIGDOC Conference, Toronto, Ontario, June, 1986.

[51] J. Law-To, L. Chen, A. Joly, et al. Video copy detection: a comparative study. Proceedings of the 6th ACM international conference on Image and video retrieval, pp.371 –378, 2007.

[52] X.S. Hua, X. Chen, H. J. Zhang. Robust video signature based on ordinal measure, International Conference on Image Processing, 2004.

[53] J. Law-To, O. Buisson, V. Gouet-Brunetand, and N. Boujemaa. Robust voting algorithm based on labels of behavior for video copy detection, ACM International Conference on Multimedia, pp.835-844, 2006.

[54] C. Y. Chiu, C. H. Li, H. A. Wang, et al. A time warping based approach for video copy detection[C], Proceedings of International Conference on Pattern Recognition, Hong Kong, pp.228-231, 2006.

[55] H. K. Tan, C. W. Ngo, and T. S. Chua, Efficient mining of multiple partial near-duplicate alignments by temporal network, IEEE Transactions on Circuits and Systems for Video Technology(CSVT), vol. 20, no. 11, pp. 1486–1498, 2010.

[56] H. T. Xie, K. Gao, Y. D. Zhang, S. Tang, J. T. Li, Efficient Feature Detection and Effective Post-Verification for Large Scale Near-Duplicate Image Search，IEEE Transaction on Multimedia, accepted, 2011.

[57] J. Cao, C.-W Ngo, Y. D. Zhang, L. Ma, Trajectory-based Visualization of Web Video Topics, ACM International Conference on Multimedia (ACM MM), Florence, Italy, 2010.

[58] X. Li, D. Wang, J. Li, B. Zhang: Video Search in Concept Subspace: A Text-Like Paradigm, ACM International Conference on Image and Video Retrieval, pp.603-610, Amsterdam, the Netherlands, 2007.

[59] J. Cao, C.-W. Ngo, Y. D. Zhang, J. T. Li, Trajectory-based Visualization of Web Video Topics, IEEE Transactions on Circuits and Systems for Video Technology(CSVT), 2011.

[60] J. Cao, Y. D. Zhang, B. L. Feng, X. F. Hua, L. Bao, X. Zhang and J. T. Li , TRECVID 2008 Search Task by MCG-ICT-CAS, In NIST TRECVID Video Retrieval Workshop. 2008.

[61] J. Cao, H. F. Jing, C.-W. Ngo, Y. D. Zhang, Distribution-based Concept Selection for Concept-based Video Retrieval, ACM International Conference on Multimedia (ACM MM), Beijing, China, Oct. 2009.

[62] B. L. Feng, J. Cao, L. Bao, Y. D. Zhang, S. X. Lin, X. G. Bao, X. C. Yun. Graph-Based Multi-Space Semantic Correlation Diffusion for Video Retrieval. International Journal of Visual Computer, in press, 2011.

[63] L. Bao, J. Cao, Y. D. Zhang, M. Y. Chen, J. T. Li, A. Hauptmann, Explicit and Implicit Concept-based Video Retrieval with Bipartite Graph Propagation Model, ACM Multimedia, Florence, Italy, 2010.

[64] L. Pang, J. Cao, Y. D. Zhang, S. X. Lin, Leveraging Collective Wisdom for Web Video Retrieval through Heterogeneous Community Discovery , ACM MM 11, accepted , 2011.

[65] J. Cao, Y. D. Zhang, B. L. Feng, L. Bao, L. Pang and J. T. Li ,TRECVID 2009 Interactive Search Task by MCG-ICT-CAS, In NIST TRECVID Video Retrieval Workshop. 2009.

[66] K. Gao, X. Wu, H.-T. Xie, W. Zhang, Z.-D. Mao, TRECVID 2009 Copy Detection Task by MCG-ICT-CAS, In NIST TRECVID Video Retrieval Workshop. 2009.

[67] W. Zhang, K. Gao, Y. D. Zhang, J. T. Li, Data-Oriented Locality Sensitive Hashing , ACM International Conference on Multimedia (ACM MM), Florence, Italy, 2010.

[68] K. Gao, Y. D. Zhang, W. Zhang, S. X. Lin, Affine Stable Characteristic Based Sample Expansion for Object Detection, ACM International Conference on Image and Video Retrieval, Xi' an, China, pp.422-429, 2010.

Author Biographies:

Cao Juan: Associate Researcher, Institute of Computing Technology, Chinese Academy of Sciences, caojuan@ict.ac.cn

Zhang Yongdong: Researcher, Institute of Computing Technology, Chinese Academy of Sciences

Li Jintao: Researcher, Institute of Computing Technology, Chinese Academy of Sciences, Director of Forward-Looking Research Laboratory

Note: Figure translations are in progress. See original paper for figures.