Vlogging: A Survey of Videoblogging Technology on the Web

WEN GAO, YONGHONG TIAN, and TIEJUN HUANG

Peking University

and

QIANG YANG

Hong Kong University of Science and Technology

In recent years, blogging has become an exploding passion among Internet communities. By combining the grassroots blogging with the richness of expression available in video, videoblogs (vlogs for short) will be a powerful new media adjunct to our existing televised news sources. Vlogs have gained much attention worldwide, especially with Google's acquisition of YouTube. This article presents a comprehensive survey of videoblogging (vlogging for short) as a new technological trend. We first summarize the technological challenges for vlogging as four key issues that need to be answered. Along with their respective possibilities, we give a review of the currently available techniques and tools supporting vlogging, and envision emerging technological directions for future vlogging. Several multimedia technologies are introduced to empower vlogging technology with better scalability, interactivity, searchability, and accessability, and to potentially reduce the legal, economic, and moral risks of vlogging applications. We also make an in-depth investigation of various vlog mining topics from a research perspective and present several incentive applications such as user-targeted video advertising and collective intelligence gaming. We believe that vlogging and its applications will bring new opportunities and drives to the research in related fields.

Categories and Subject Descriptors: F.1.2 [**Computation by Abstract Devices**]: Modes of Computation—*Online computation*; H.3.4 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; H.4 [**Information Systems Applications**]: Communications Applications—*Computer conferencing, teleconferencing, and videoconferencing*

General Terms: Design, Human Factors, Management

Additional Key Words and Phrases: Survey, vlogs, vlogging, multimedia computing, vlog mining

ACM Reference Format:

Gao, W., Tian, Y., Huang, T., and Yang, Q. 2010. Vlogging: A survey of videoblogging technology on the Web. ACM Comput. Surv. 42, 4, Article 15 (June 2010), 57 pages.
DOI = 10.1145/1749603.1749606, http://doi.acm.org/10.1145/1749603.1749606

The authors are supported by grants from National Basic Research Program of China under contract No. 2009CB320906, Chinese National Science Foundation under contract No. 60605020 and 90820003, and National Hi-Tech R&D Program (863) of China under contract No. 2006AA01Z320 and 2006AA010105. Dr. Qiang is supported by Hong Kong CERG grant 621307.
Authors' addresses: Y. Tian, National Engineering Laboratory for Video Technology, School of EE & CS, Peking University, Beijing, 100871, China; email: yhtian@pku.edu.cn.

DOI 10.1145/1749603.1749606 http://doi.acm.org/10.1145/1749603.1749606

1. INTRODUCTION

When the Web was invented, practically all forms of online media were still very much a one-way street. However, with the rapid development and proliferation of the World Wide Web, particularly with the emergence of Web 2.0 [O'Reilly 2005] and beyond, we now have new means by which to express our opinions and gain access to information instantly. *Weblogs* (*blogs* for short) are playing an increasingly important role in realizing this goal. Like instant messaging, email, cell phones, and Web pages, blogs are a new form of mainstream personal communication, allowing millions of people to publish and exchange knowledge/information, and to establish networks or build relationships in the blog world [Rosenbloom 2004]. Blogs began as a textual genre of personal publishing, but within this genre visual expression, such as photoblogs, and the adaptation of sound and video [Hoem 2005] were developed. By combining the ubiquitous, grassroots blogging with the richness of expression available in video, *videoblogs* (*vlogs* for short) will be an important force in a future world of Web-based journalism and a powerful new media adjunct to our existing televised news sources [Parker and Pfeiffer 2005].

As defined by Wikipedia [2008], *videoblogging*, shortened as *vlogging* in this article, is a form of blogging for which the medium is video. Vlog entries are made regularly and often combine embedded video or a video link with supporting text, images, and other metadata. In recent years, vlogging has gained much attention worldwide, especially with Google's acquisition of YouTube in November 2006 for $1.65 billion. YouTube is a popular vlog site that lets users upload, tag, and share video clips—thereby making them known to the world. Since a video can show a lot more than text, vlogs provide a much more expressive medium for vloggers than text-blogs in which to communicate with the outer world. This has particular appeal to younger audience, who are typically equipped with popular mobile devices such as personal digital assistants (PDAs) and camera phones.

Vlogs have brought about a new revolution in multimedia usage. Nevertheless, a quick look at the current vlogs reveals that not many of the results from multimedia research, such as media content analysis, semantic content classification and annotation, structured multimedia authoring, or digital rights management, found their way into vlogging techniques. For example, a simple shot-segmentation technique could provide YouTube users with small bookmarks for each video and allow them to easily jump into different video scenes, instead of watching it all or using the slider [Boll 2007]. Hence, there are many opportunities for multimedia researchers to provide vloggers with better vlogging techniques for a more powerful experience in designing and using vlogs.

This article outlines the current state of vlogging technology, covers the probable evolution, and highlights what newcomers leading the charge are poised to start. Our article is partly motivated by Parker and Pfeiffer [2005], which for the first time provided an introductory view of vlogging technology before 2005. Since then, vlogging has become a huge success and is a very hot online application with high general interest. We found that it is worthwhile to revisit the current vlogging technology and explore the potential benefits and challenges that the multimedia community offers to vlogging, and vice versa. Our idea is consistent with the panelists' opinions in the panel titled "Multimedia and Web 2.0: Hype, Challenge, Synergy" that was held half a year after the time of this writing (April 2006), at the ACM 2006 Multimedia Conference in Santa Barbara, California.

A possible organization of the various facets of vlogging technology is shown in Figure 1 (see Section 2.3 for a more detailed description). Our article follows a similar structure. For the purpose of completeness and readability, we first review the

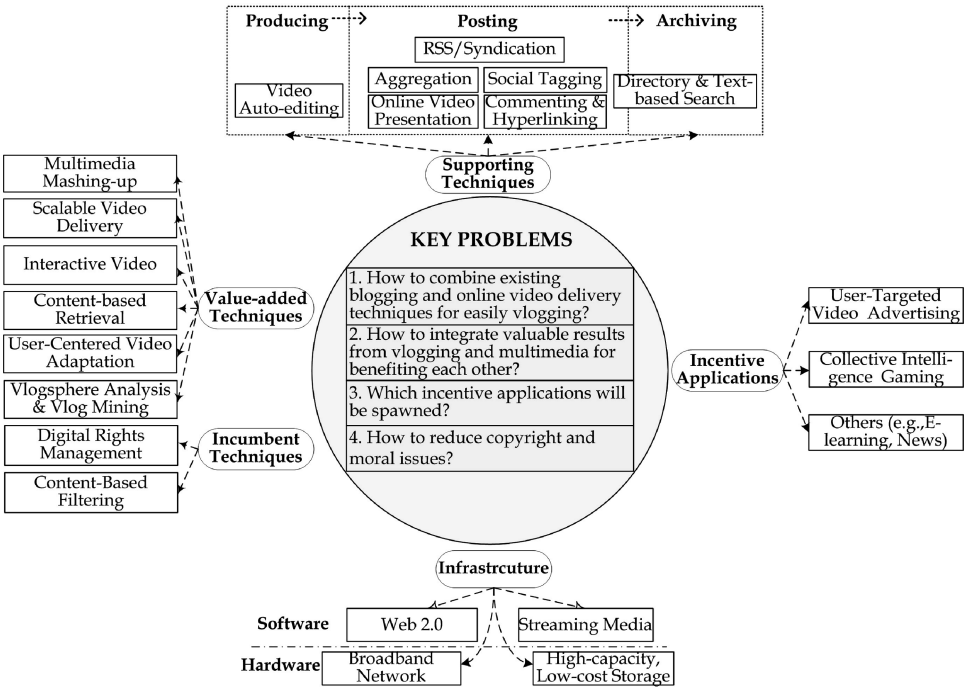

Fig. 1. A view of the different facets of vlogging technology, reflected in the structure of this article.

blog phenomenon of recent years and then present an overview of current vlogs in Section 2. We also summarize in this section the key issues of vlogging technology that need to be answered. The rest of this article is arranged as follows: In Section 3, we give a review of the infrastructure and tools of current vlogging technology. Then, we present a new vision for future vlogging and introduce some techniques from multimedia to help make this vision possible in Section 4. Since the vlogging phenomenon also poses many new computational opportunities for researchers, we pay special attention to these issues in Section 5. The prevalence of vlogging and the further combination of vlogging and multimedia technologies will in turn give birth to attractive applications such as vlog-based advertising and gaming, thus, in Section 6, we discuss some samples of these applications. Finally, we conclude in Section 7.

2. VLOGS AND VLOGGING: A GLOBAL PICTURE

2.1. The Evolution of Blogs: From Text to Audio-Video

In or around 1997, blogging became a quickly spreading passion among Internet literates. A *weblog*, or *blog* is a "frequently updated Web page with dated entries in reverse chronological order, usually containing links with commentary" [Blood 2002]. The term originated from "WeB log," and was promoted further by www.blogger.com as a blog. *Blog* can also be used as a verb, meaning to maintain or add content to a blog. The act itself is often referred to as *blogging*. According to Gill [2004], several primary characteristics of a blog include regular chronological entries; links to related news articles, documents, or blogs (referred to as *blogrolling*); archived entries such that old content remains accessible via static links (referred to as *permalinks*); ease of syndication with RSS (Rich Site Summary, see Section 3.2.2) or XML feed. Blogs contribute to Web

content by linking and filtering evolving content in a structured way. All blogs and their interconnections are often called the *blogosphere* [Rosenbloom 2004]. It is the perception that blogs exist together as a connected community (or as a collection of connected communities) or as a social network.

Many blogs provide commentary or news on a particular subject; others function as more personal online diaries. Five main blogging motivations were identified in Nardi et al. [2004]: documenting one's life; providing commentary and opinions; working out emotional issues; thinking by writing; and promoting conversation and community. Blogs have become an increasingly important way of learning about news and opinions not found in mainstream media, and blogging has become a popular social activity for establishing and maintaining online communities. The number of blogs is growing exponentially—the famous blog search engine, Technorati, reported that there were about 4.2 million blogs worldwide in October 2004 [Rosenbloom 2004], and up to 112.8 million blogs by December 2007. It is estimated that there are about 175,000 new blogs a day, and about 11% (or about 50 million) of Internet users are regular blog readers.

Traditionally, blogging is a textual activity, which is limited because text is only one aspect of the diverse skills needed in order to understand and manage different aspects of modern communication [Hoem 2005]. In recent years, the amount of digital multimedia distributed over the Web has increased tremendously because almost anyone can follow the production line of digital multimedia content. As a result, many different types of blogs have emerged gradually, such as *artlog* (i.e., a form of art-sharing and publishing in the format of a blog); *photoblog* (i.e., a blog containing photos); *sketchblog* (i.e., a blog containing a portfolio of sketches); and in particular *audioblog* (or *podcast*) and *videoblog* (*vlog* for short). They differ not only in the type of content, but also in the way that content is delivered or written.

Aiming to present a richer story for bloggers, audioblogs and vlogs are twins; they exhibit many similarities in terms of their content and production. Simply speaking, an audioblog is a service that provides bloggers with the ability to attach audio to their blogs with a microphone at any time from anywhere. Note that audioblogs are also known as podcasts, since their pioneers are the iPod users; although podcasting is in a stricter sense merely one form of audioblogging. On the other hand, even before audio media became a full-blown component of blogging, video was already being experimented with [King 2003]. Vlogging features many of the same characteristics found in audioblogging, providing vloggers with a more compelling and expressive medium.

Vlogging has experienced tremendous growth over the past several years. The Yahoo! Videoblogging Group saw its membership increase dramatically in 2005. The most popular video-sharing site to date, YouTube, was publicly launched between August and November 2005 and acquired by Google in November 2006. By March 17th, 2008, there were up to 78.3 million videos published on YouTube, with 150,000 to over 200,000 videos published everyday. According to the data collected by Mefeedia, the number of vlogs was only 617 in January 2005 and 8,739 in January 2006, but up to 20,193 in January 2007 [Mefeedia 2007]. Driven by the prevalence of digital cameras and mobile phones and the near-ubiquitous availability of broadband network connections, vlogging has surged to an unprecedented level, and gained much attention worldwide.

2.2. Vlogs: Concept and Taxonomy

According to Wikipedia [2008], a *videoblog*, which is shortened to *vlog* in this article, is a blog that uses video as the primary content. Vlog entries are made regularly, and often combine embedded video or a video link with supporting text, images, and other

metadata. These videos may be embedded and watched on the viewer's Web browser, or downloaded to the viewer's machine or a portable device for later viewing. Like textblogs, vlogs also often take advantage of Web syndication to allow for the distribution of video over the Internet using either the RSS or Atom syndication formats, for automatic aggregation and playback on mobile devices and PCs [Wikipedia 2008]. Apart from the term *vlogs*, videoblogs are also known as v-logs, vid-blogs, movie blogs, vblogs, vidcasts, videocasts, vcasts, v-casts, episodic video, Web shows or online TV, and so on. Vlogs are created by *videobloggers* or *vloggers*, while the act itself is referred to as *videoblogging* or *vlogging*. As a derivative of blogosphere, *vlogosphere* is the collective term encompassing all vlogs. Vlogs exist together as a community or a social network.

Roughly, the *life cycle* of a vlog is defined in this article as consisting of three stages:

—*Producing*: In this stage, the vlogger creates and edits a video, and uploads it to the hosting site and then a new vlog is generated. Sometimes, vloggers can even edit videos online, without a locally installed software.

—*Posting*: The new vlog is distributed online, and starts to get recognized in the vlogosphere and often get linked by other vlogs. In this stage, the vlog can be viewed and commented by other vloggers. Some vlogs even become authorities in a certain field. Rather than always having to remember to visit a site, a vlog reader can configure his or her RSS browser software to automatically subscribe to the timely updates.

—*Archiving*: When the vlog becomes out of date or loses usefulness, it will be archived or even deleted.

To further characterize a vlog, we have to distinguish among several significantly different genres that claim to be vlogs, ranging from simply uploading unedited video files via play-lists to edited sequences, sometimes with complex interactivity [Hoem 2005]:

— *Vogs*: Vogs are made of pre-edited sequences which normally include interactive elements. Typically, they are made with different kinds of software and then posted to individual blog sites. A typical example of vogs is the vlog of B. Obama in the 2008 US presidential election, at http://www.youtube.com/user/BarackObama-dotcom (see Figure 2(a)).

— *Moblogs*: Mobile blog (moblog for short) is a form of blog in which the user publishes blog entries directly to the Web from a mobile phone or other mobile device [Wikipedia 2008]. Entities in moblogs normally contain uploaded pictures, audio-video clips and additional text, not edited sequences. Moblogging is popular among people with camera-enabled cell phones which allow them to either e-mail photos and videos that then appear as entries on a Web site, or use mobile blogging software to directly publish content to a Web server. Here moblogs are used to denote vlogs containing relatively short, autonomous video clips. Examples of moblogs can be found at http://moblog.net/ (see Figure 2(b)), which is a well-known free hosting site for moblogs.

— *Playlists*: Playlists are collections of references to video files on different servers, which may even provide a level of interactivity without manipulating the content in these files. One way to achieve this is by using SMIL (synchronized multimedia integration language). Strictly speaking, playlists are the filter-style vlogs. Examples of playlists can be found at http://www.mefeedia.com/collections/ (see Figure 2(c)), which contains 4,114 playlists, gathering 16,182 videos by April 2008.

It should be noted that the distinction between vogs and moblogs is clear [Hoem 2005]: the videos in vogs are edited, and may offer quite complex interactivity, while moblogs

Fig. 2. Screenshot of exemplary vlogs: (a) vog at http://www.youtube.com/user/BarackObama-dotcom; (b) moblog at http://moblog.net/; (c) playlist at http://www.mefeedia.com/collections/; and (d) diary-style vlog at http://crule.typepad.com/.

are generally easy to use because moblogging is just a matter of simply uploading a video file to a dedicated vlog site.

Vlogs can also be grouped into two categories according to their presentation styles: diary or Web-TV show. The vlogs of the first category take the online video diary style. Such an example is Charlene's vlog at http://crule.typepad.com/ (as shown in Figure 2(d)), where video is included through hyperlinks. Currently, most vlogs take the second category of presentation formats, i.e., the Web-TV show-like style. Such an example is Obama's vlog (as shown in Figure 2(a)), which has inline video files.

Since a video can show a lot more than pure text, pictures, or audio, there are a lot more things that a vlog can cover compared to a typical textblog, photoblog, or audioblog. For example, if one runs a real-estate business and he or she wishes to provide regular virtual tours of the properties for sale, a vlog that features regular Web tours would be very appropriate. In some sense, video is easier to produce, as it only needs to record some realistic scenes and can be informative even with little or no editing. The affordability and portability of most consumer-level video camcorders and some mobile phone handsets mean that anyone can easily express themselves in a documentary or narrative-style film on a daily basis and then post it for viewer consumption [King 2003]. Compared to text- or audioblogs, vlogs can be used in a much wider range of applications such as online education and learning, online gaming, products marketing, and news reporting.

Table I. Differences Among IPTV, Internet Video and Vlogging Services

	IPTV	Internet Video	Vlogging
Content	TV programs, movies	Movies, TV programs, news, partially user-generated video	Mainly user-generated video in talk-radio format
Length	Long clips or live stream	Long clips or live stream, and partially short clips	Short clips (typically 3–5 minutes)
Video Format	MPEG 2/4, H.264 (with high-definition)	Windows Media, Real, Quick Time, Flash and others	Mostly Flash (often with low-definition)
Video Quality	"Broadcast" TV Quality, Controlled QoS	Best effort quality, QoS not guaranteed	Often low quality, no QoS
Content Organization	EPGs by timeline or topic category order	Often by topic category, by view times/upload date etc.	As dated entry in reverse chronological order
Users	Known customers with known IP and locations	Any users (generally unknown)	Any users (generally unknown)
Footprint	Local (limited operator coverage)	Potentially supranational or worldwide	Potentially supranational or worldwide
Receiver Device	Set-box with a TV display	PC and mobile devices	Mobile devices and PC
Producing	Professional filmmaking tools for filmmakers	Advanced editing tools for service providers	Simple-to-use editing tools for common users
Transmission	Broadband network with CDN support	Internet access, often using P2P transmission	Internet access, possibly using P2P transmission
Delivery Form	Video-on-demand, live broadcast, delayed live	Video-on-demand, partially live broadcast	Video-on-demand
Interactivity	Program selection, few interaction	(Sometimes) voting or ranking, with a few interaction	Commenting or hyperlinking, social interaction
Searchability	Metadata/EPG search	Metadata/caption-based, partially content-based search	Metadata search, XML-based aggregation
RSS Support	No	Partially yes	Yes
Reliability	Stable	Subject to connection	Subject to connection
Security	Users are authenticated and protected	Unsafe	Unsafe
Copyright	Protected video	Often unprotected	Mostly unprotected
Other Services	EPGs, onsite support	Low-quality on-demand services, generally no support	Generally no support

*Some comparison items between IPTV and Internet video are quoted from Martinsson [2006].
**CDN (content delivery network) is a system of computers networked together across the Internet that cooperate transparently to deliver content (especially large media content) to end users.
***EPG means electronic program guide, an on-screen guide to scheduled programs, typically allowing a viewer to navigate, select, and discover content by time, title, channel, genre, and so on.

2.3. Vlogging: Key Technological Challenges

Vlogging is a form of blogging for which the medium is video. However, online delivery of video content is not a new activity, either by direct downloading or via streaming. It is useful to make a comparison among three closely related services of online video delivery (see Table I): IPTV (Internet Protocol Television), Internet video, and vlogging. From the content perspective, the indexed video content in three services have some differences. Generally speaking, IPTV often provides high-quality TV programs and movies with guaranteed quality of services (QoS); a general online video service often supports a wide range of videos, from low-quality videos generated by nonprofessional users, to high-quality TV clips or movie files, either in live broadcast or via downloaded file; while videos in vlogs are often generated by nonprofessional users in a talk-radio format and with fairly short length, encoded with relatively low resolution and archived in reverse-chronological order. Different video types and dissimilar

service targets then characterize their different supporting technologies, as shown in Table I.

Like text-only blogging, we need supporting techniques to facilitate effective vlogging. However, it is clear that a simple combination of textblogging tools and online video delivery does not meet the ever-growing requirements of vloggers [Parker and Pfeiffer 2005]. Thus, with the explosive growth of vlogs worldwide, several challenges are posed for vlogging technology, which are summarized as follows:

— *Basic supporting issue*: The first challenge mainly addresses the basic supporting infrastructure and techniques for vlogging. Network bandwidth and media storage can be considered as two major *hardware* infrastructures for vlogging, while Web 2.0 and streaming media technologies are two *software* infrastructures. To facilitate vloggers designing and using vlogs, possible solutions range from simply integrating textblogging tools with online delivery of video, to flexibly combining different multimedia systems and services on all stages of vlogging. Even in the simplest situation, the supporting platform also needs some blogging softwares and systems (e.g., RSS/Syndication, content management systems, hosting and directory services), and necessary video production and distribution tools (e.g., video creation and coding software, streaming servers). Given these matured systems, tools and services, vlogging technology will go further by "mashing-up" them into more powerful supporting platforms.

— *Value-added issue*: Vlogging is a typical application of Web 2.0. As pointed out by Boll [2007], Web 2.0 and multimedia can benefit each other, and we should integrate each other's valuable results and best practices. Thus the second challenge is mostly about what multimedia technology could give to vlogging (not to the much broader field, Web 2.0) and vice versa. In this sense, we refer to this challenge as a *value-added issue*, compared with the "basic supporting issue".

— *Incumbent issue*: The ease with which anyone can view and upload videos to YouTube and similar vlog hosting sites also poses potential copyright, moral, and legal issues [Meisel 2008]. For example, Viacom, a large corporation whose holdings include MTV and Comedy Central, sued YouTube in March 2007 for over $1 billion for "massive intentional copyright infringement" [Peets and Patchen 2007]. The "incumbent techniques" for distributing content can react aggressively to this new source of competition and pursue technological strategies to combat the abuse of unauthorized content by vloggers.

— *Incentive application issue*: In the development process of vlogging, YouTube's success is undoubtedly one of the major milestones, mainly because it created an incentive application platform that can provide a better user experience around sharing video clips and publishing vlogs online. This shows us that the spawning of incentive applications may be one important issue for future vlogging technology. Some promising examples are vlog-based advertising and gaming. These incentive applications may drive the further development of vlogging.

These four challenges cover most of the important problems and aspects related to the current and future vlogging technology. We believe that embarking on the research and development of vlogging technology should consider and tackle these challenges. Along with their respective possibilities, we can comprehensively survey, analyze, and quantify the current progress and future prospects of vlogging technology. Towards this end, different facets will be addressed one by one in the following sections, as illustrated in Figure 1.

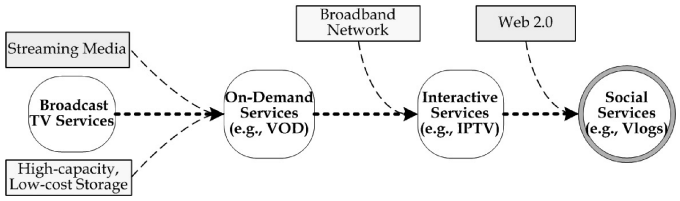

Fig. 3. An overview of sketchy roadmaps on video services.

3. VLOGGING TECHNOLOGY IN THE REAL WORLD

By the nature of its task, current vlogging technology boils down to two intrinsic problems: (a) from a *system perspective*, what software, hardware, and network infrastructures are need to support online vlogging applications and services; and (b) from a *user perspective*, how to provide vloggers with technological supports for all stages of vlogging. Thus this section is dedicated to understanding the current status of vlogging infrastructures, techniques, and tools in the real world from the two perspectives above.

3.1. Vlogging Infrastructure

As far as technological advances are concerned, growth in video services has unquestionably been rapid. Figure 3 shows an overview of sketchy roadmaps on video services, which are roughly divided into four phases: broadcast TV services (e.g., traditional TV); on-demand services (e.g., VOD); interactive services (e.g., IPTV); and social services (e.g., vlogs). We can see that this evolving progress is advanced by several key technologies. For example, high-capacity, low-cost storage and streaming media technologies enable on-demand video services, and IPTV is often provided in conjunction with VOD and broadband network technologies. Similarly, vlogging services emerge when these technologies and the new generation of the Web, Web 2.0, meet. Compared with interactive video services such as IPTV, what makes vlogs new is their social participation and embeddedness. From this perspective, we refer to these kinds of new video services as *socially-interactive video services*, or simply *social services*. It should also be noted that for social services such as vlogs, these technologies can serve as the enabling infrastructures. Among them, broadband network and media storage can be considered two major *hardware* infrastructures for vlogging technology. Meanwhile, vlogging is closely related to Web 2.0 and streaming media technologies (referred to as the *software* infrastructures).

3.1.1. Broadband Network and Media Storage. The advances in computer networking combined with high-capacity, low-cost storage systems made online delivery of video practical, either by direct downloading or via streaming. Compared to text-, photo- or even audioblogs, vlogs obviously place many more requirements on network bandwidth and media storage capacity. Take YouTube as an example. In August 2006, the Wall Street Journal published an article revealing that YouTube was hosting about 6.1 million videos and had more than 65,000 videos uploaded everyday, consequently requiring about 45 terabytes of storage space and as much as 5 to 6 million dollars per month in running costs (specifically the required bandwidth). These figures were up to 78.3 million videos hosted totally and 150,000 videos uploaded daily in March 2008. We can speculate that the required storage space and corresponding running costs are exponential in growth.

From a user perspective, the limited network bandwidth also poses a significant challenge for efficient video transmission in the current network environment where sufficiently high bandwidth is not widely available. To reduce network congestion, peer-to-peer (P2P) transmission technology is widely used. The P2P network refers to a network that primarily relies on participants in the network, rather than on a few dedicated servers for its service. This is a fundamentally different paradigm compared with the client-server architecture, and brings a number of unique advantages such as scalability, resilience, and effectiveness to cope with dynamics and heterogeneity [Liu et al. 2008]. BitTorrent is a protocol designed for transferring files in a P2P network, which has been exploited by many vlogs and related tools, such as FireANT, to improve the transmission efficiency.

Discussion. From the perspective of vlogging service providers, video storage and transmission costs are still significant; from the perspective of vloggers, however, despite the network, bandwidth is still a concern for the vloggers who do not have sufficiently high bandwidth available; however, the problem will ventually solve itself with the advent of near-ubiquitous broadband/wireless Internet access. Therefore, more detailed analysis of broadband network and media storage is outside the scope of this article.

3.1.2. Web 2.0. Web 2.0 refers to an updated and improved version of WWW that allows the users to communicate, collaborate, and share information online in completely new ways. Web 2.0 has numerous definitions. As pointed out by O'Reilly [2005], there are huge numbers of disagreements on what Web 2.0 means. However, the central point of Web 2.0 is the user's stronger involvement and participation in the Web, forming social networks and virtual communities in a global network. Generally speaking, Web 1.0 was marked by isolated information silos and static Web pages where the user used search engines or surfed from one Web site to another; Web 2.0, on the other hand, is characterized by a more living, dynamic, and interactive Web that is based on social networks, user-generated content, and that is an architecture of participation [Best 2006; Jensen 2007].

One of the key drives in the development of Web 2.0 is the emergence of a new generation of Web-related technologies and standards [Anderson 2007]. Technologically, Web 2.0 is mainly considered [Boll 2007] to represent: (1) the combination of matured implementation techniques for dynamic Web applications; (2) the mashing of existing Web services into value-added applications; and (3) an emphasis on community and collaboration, as well as the establishment of direct distribution platforms for user-generated content. It should be noted that the technological infrastructure of Web 2.0 is still evolving.

One important feature of Web 2.0 is that it puts less emphasis on the software and far more on the application that provides a service. Key Web 2.0 applications include blogging (including audioblogging and vlogging), wiki, tagging and social bookmarking, multimedia sharing, and content syndication [Anderson 2007]. Some well-known commercial examples that count as Web 2.0 applications are Flickr, YouTube, Upcoming, Wikipedia, Facebook, and Google Earch.

Discussion. Web 2.0 provides simple, but interesting, working services and is better at involving and integrating users so that they can participate more. Given the strong focus on media in Web 2.0, however, not many of the valuable results from multimedia found their way into Web 2.0 [Boll 2007]. It can be argued that vlogging is exactly one of the best fields in which they could meet.

3.1.3. Streaming Media. From a user perspective, streaming media are viewable immediately after downloading starts [Lawton 2000]. The real-time content is streamed through the user datagram protocol (UDP). In terms of transport protocols, a streaming-media server uses internet protocol (IP) multicast, the real-time streaming protocol (RTSP) incorporating UDP, or the transmission control protocol (TCP), or TCP alone—a hybrid strategy with real-time streaming via HTTP when a firewall cannot pass UDP data [Pieper et al. 2001]. Currently, most vlogs can support the streaming media preview mode. This means that the user can check out an audio or video before deciding to download it.

During the past two decades, there have been tremendous efforts and many technical innovations made in support of real-time video streaming. IP multicast represents an earlier attempt to tackle this problem [Deering and Cheriton 1990]. It forwards IP datagrams to a group of interested receivers by delivering the messages over each link of the network only once, and creating copies only when the links to the destinations split. As a loosely coupled model that reflects the basic design principles of the Internet, IP multicast retains the IP interface and requires routers to maintain a per-group state. Unfortunately, today's IP multicast deployment remains largely limited to reach and scope due to concerns regarding scalability, deployment, and support for higher level functionality [Liu et al. 2008]. To address many of the problems associated with IP multicast, several researchers have advocated moving multicast functionality away from routers towards end systems [Chu et al. 2000]. As such a representative nonrouter-based architecture, P2P-based broadcast has emerged as a promising technique that supports real-time video streaming, but is also cost effective and easy to deploy [Li and Yin 2007; Liu et al. 2008]. Thus, many P2P video streaming solutions and systems are proposed.

Discussion. With more and more vlogs taking the presentation form of online Web-TV, streaming media technology is expected to be more cost-effective, more scalable, and easier to deploy. However, although P2P applications such as file download and voice-over-IP have gained tremendous popularity, P2P-based video delivery is still in its early stage, and its full potential remains to be seen. We will explore this issue further in Section 4.1.1.

3.2. Existing Supporting Techniques for Vlogging

Essentially, existing vlogging technology is a simple mashing of textblogging techniques with online video delivery. Figure 4 shows a paradigm of existing supporting techniques for vlogging; specifically, they focus mainly on the following issues:

—How to support efficient online video presentation and delivery;
—How to publish the machine-readable items containing metadata (such as the author's name and publication date) and content for videos;
—How to add comments to vlogs, or how to interact with vlogs through hyperlinks;
—How to automatically aggregate updates from widely distributed vlogs;
—How to collaboratively create and manage tags for annotating and categorizing video content;
—How to effectively index and search the regularly updated vlogs (and videos).

Below, we review each of these issues.

3.2.1. Online Video Presentation and Delivery. The broadband-powered ability to view and download online video content is just one part of vlogging; the other part is appropriate video presentation and delivery interfaces that allow vloggers to quickly and

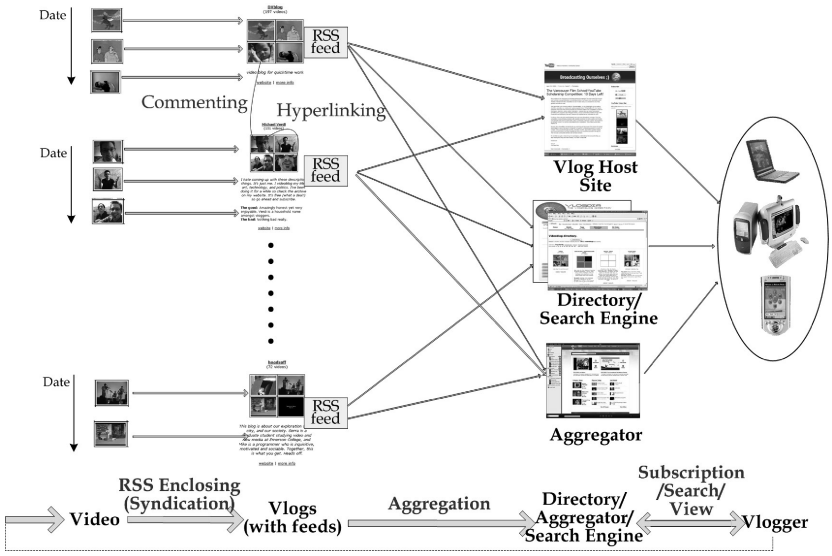

Fig. 4. A paradigm of existing supporting techniques for vlogging.

intuitively access vlogs in a Web browser or even use a mobile handset. The continuous improvement in video delivery solutions, including video compression/coding, online video player, and streaming video delivery, has played an important role in advancing the prospects of online vlogging services.

Video compression and coding for streaming media delivery on the Internet has been studied for more than ten years [Conklin et al. 2001], and is still an active research field. Generally speaking, video compression is used to (often lossy) reduce the size of the video data, followed by codecs (i.e., coding and decoding devices or softwares) to compactly but accurately represent the video content. The present generation of codecs, such as Windows Media Video (WMV), VC-1, AVC, and DivX, are now capable of compressing rather large files to much smaller while maintaining picture resolution. For the video stream to be useful in stored or transmitted forms, it must be encapsulated together in a container format. A container format is a computer file format that can contain various types of data, compressed by means of standardized audio/video codecs. The most popular containers for streaming video are AVI, ASF, QuickTime, RealMedia, DivX, and MP4. Finally, coded videos can be played in video players such as Apple's QuickTime, Microsoft's Media Player, and RealNetworks' RealPlayer.

However, compared with RealMedia, QuickTime or Windows Media, Flash video (FLV) has much greater presence in the real-world applications of online video delivery. FLV videos can be delivered in the following ways: (1) using embedded video within SWF files; (2) using progressive download FLV files; and (3) streaming video from own Flash Media Server or from a hosted server using Flash Video Streaming Services. Table II summarizes the characteristics of FLV video delivery techniques (quoted from Adobe [2009] with necessary simplification).

The current crop of online video and vlog Web sites such as MySpace, YouTube, and Google Video use FLV codec as the codec of choice, because it is highly efficient, playable and viewable on most operating systems via the widely available Adobe Flash Player and Web browser plug-in [Amel and Cryan 2007]. In contrast to video in other formats, often played in a dedicated external player, Flash player is directly embedded into a Web page without requiring the user to launch a separate application. FLV

Table II. Summary of Video Delivery Techniques in FLV Video

	Embedded Video	Progressive Download	Streaming Delivery
Encoding	VP6 video codec for FLV Player 8^+, or be encoded by other codecs and then packaged by Flash	Plug-in or stand-alone FLV Encoder, also support MPEG-4 formats encoded using H.264 codec	Same as progressive download, additionally with bandwidth detection capabilities in streaming
File Size	AV streams and Flash interface contained in a single large file	SWF and FLV files are separate, resulting in a smaller SWF file size	Same as progressive download
Timeline Access	Individual keyframes are visible on Flash Timeline	Be played back at runtime, keyframes are not visible	Same as progressive download
Publishing	Need to republish the entire video file	Not require referencing the video file directly	Same as progressive download
Frame Rate	Video's and SWF's frame rates are the same	Video's and SWF's frame rates may be different	Same as progressive download
ActionScript Access	Control SWF content's playback on Flash Timeline for video playback	Use the netStream object for video playback	Same as progressive download, and can also use server-side ActionScript
Components	No video-specific components	Media, FLVPlayback components	Media, FLVPlayback, sever comm. components
Web Delivery	SWF file is progressively downloaded	Video files are progressively downloaded, cached for play	Video files are streamed from server, played locally
Performance	AV sync. is limited after 120s of video	Bigger and longer video, with reliable AV sync. and best image quality	Best performance, with optimal bitrate and limited image quality
Usage	Short video clips ($<$1m, $<320\times240$ and $<$12fps)	Long video clips ($>720\times480$ and $<$30fps)	Very long video clips, live and multiway streaming

*This table is taken from Table 2 in Adobe [2009] with necessary simplification.

is also popular due to its small file size (exactly for this reason it is often called a *microvideo*), interactive capabilities, and progressive downloading (i.e., video begins playing without having fully downloaded the clip) [Owens and Andjelic 2007]. To support different connection speeds, FLV codecs provide a wide range of encoding profiles, from a total bitrate of 44 Kbps to 800 Kbps. Recent versions of Flash have added accessibility functions [Reinhardt 2008]. For example, additional content, such as subtitles or transcripts, can be associated with a Flash movie, consequently making FLV content visible to search engines; Flash Lite is available for some cell phones and other devices.

Discussion. Due to high streaming efficiency, FLV codec is quickly becoming the mainstream codec of online video. However, the quality of FLV videos suffers greatly. By comparison, typical DVD bitrates are around 8 Mbps, and DV video has a bitrate of 25 Mbps. With more high-bandwidth networks being rolled out, higher-quality video delivery techniques will be adopted to provide vloggers with better experiences.

3.2.2. Media Syndication. For a number of blogs, tracking updates everyday is a cumbersome and tedious task. Thus, we need an XML-based format for publishing headlines of the latest updates posted to a blog or Web site for use by other sites and direct retrieval by end users. The format, known as a *feed*, includes a headline, a short description, and a link to the new article. This *syndication* process enables Web sites or blogs that share a common interest to expand their content. By subscribing to feeds, bloggers can quickly review the latest updates on those blogs from a consolidated index rather than browsing from blog to blog.

Currently, the major blogging syndication formats are *RSS* and *Atom*. RSS stands for rich site summary, RDF (resource description framework) site summary, or really simple syndication, and is actually an XML-based metadata description and

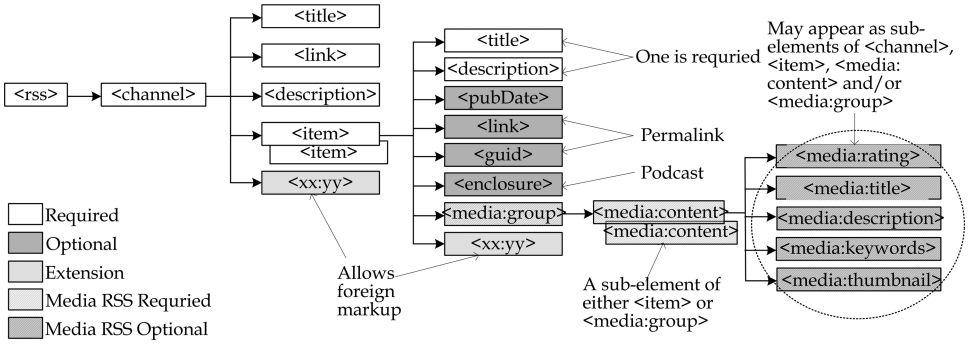

Fig. 5. An element tree for RSS 2.* and Media RSS.

syndication standard for publishing regular updates to Web-based content. RDF is a family of W3C specifications originally designed as a metadata model, but which has come to be used as a general method of modeling information through a variety of syntax formats. Currently, RSS 1.* and RSS 2.* are two major lineages of RSS. RSS 1.* uses namespaces and RDF to provide extensibility, and thus is suitable for the applications that need advanced RDF-specific modules; while RSS 2.* uses modules to provide extensibility, and thereby is suitable for general-purpose, metadata-rich syndication. Figure 5 shows a simple element tree for RSS 2.*. At the top level, an RSS document is a <rss> element, with a mandatory attribute-called version that specifies the version of RSS that the document conforms to. Subordinate to the <rss> element is a single <channel> element, which contains information about the channel (metadata) and its contents. (For more details about RSS, see http://www.rss-specifications.com/rss-specifications.htm).

To further enhance the enclosure capabilities of RSS 2.*, a group from Yahoo! and the Media RSS community worked on a new RSS module in 2004, called Media RSS. It extends enclosures to handle other media types, such as short films or TV, as well as provides additional metadata with the media, thereby enabling content publishers and bloggers to syndicate multimedia content. (For more details about Media RSS, see http://search.yahoo.com/mrss).

By incorporating lessons learned from RSS, the Internet Engineering Task Force (IETF) is working on another syndication format—Atom. Atom defines a feed format for representing and a protocol for editing the Web resources such as blogs and similar content. The feed format is analogous to RSS but more comprehensive, and the editing protocol is novel [Parker and Pfeiffer 2005]. (For more details about Atom, see http://www.ietf.org/html.charters/atom-pubcharter.html).

It should be noted that some multimedia syntax standards such as MPEG-21 can also be used to syndicate video in vlogs. MPEG-21 is the newest of a series of MPEG standards. With the aim to enable the use of multimedia resources across a wide range of networks and devices, MPEG-21 provides an open standards-based framework for multimedia delivery and consumption [Burnett et al. 2003].

Discussion. Since overlaps exist between a blogging language such as RSS or Atom and a multimedia metadata markup language such as MPEG-21, a combination might be the best solution [Parker and Pfeiffer 2005]. Therefore, media syndication of vlogs is more complex than that of textblogs.

3.2.3. Vlog Aggregation. With media syndication, we can use a program or a Web site, known as an *aggregator*, to regularly check the updates of various vlogs, and then

present results in summary forms for vloggers. A more exciting use is to automatically aggregate vlogs into topic-based repositories (i.e., vlog directories). Thus, the *aggregation* process lets updates from widely distributed Web sites be viewed as an integrated whole, consequently allowing the automated creation of continuous feeds for a particular topic from distributed sources.

Media aggregators are sometimes referred to as "podcatchers" due to the popularity of the term "podcast" used to refer to a feed containing audio or video. They can be used to automatically download media, playback the media within the application interface, or synchronize media content with a portable media player. Current media aggregators essentially employ the textblog aggregation technique. Once feeds of various vlogs are available, a vlog aggregator just retrieves the headlines, textual descriptions, thumbnails, and links of new video items, without performing more complex, but important, video content analysis and summarization. Despite this, there are still some advanced features that differentiate vlog aggregators from textblog aggregators, like the following:

— Bad internet archive warning: Even if it only contains a video file of 1 to 3 minutes, a vlog entry is much larger in size than a textblog entry. To reduce the storage cost, many vloggers choose to periodically remove their out-of-date video items from their vlog archives but retain the corresponding feeds in the archives. Thus, detecting whether the archived videos are valid is indeed an important feature of vlog aggregators such as FireANT.

— Filtering and alerting adult feeds: With more image/video content being placed online, the chance that individuals will encounter inappropriate or adult-oriented content increases. Hence detecting adult feeds is also a necessary function of vlog aggregators. Towards this end, a simple method is to exploit the <media:adult> or <media:rating> element in Media RSS, or classify feeds based on their text content. Some more advanced methods such as pornographic content-detection based on images [Rowley et al. 2006] or classification of texts and images [Hu et al. 2007] can also be used.

— Support for a streaming media preview mode: Due to the large size of video files in vlogs, some existing vlog aggregators (e.g., FireANT) and video surrogates [Song and Marchionini 2007] provide support for a streaming media preview mode. This means that the user can check video content before deciding to download it. This is one of the special features of vlog aggregators compared with textblog aggregators.

— Full BitTorrent support and scheduling for downloading media files: Again due to the large size of video files in vlogs, some existing vlog aggregators (e.g., FireANT) also support downloading media files in BitTorrent to improve the transmission efficiency. Users can even schedule the downloading time for further reducing network congestion. This is featured by desktop vlog aggregators.

Discussion. Video aggregation is one of the key innovations in the online video ecosystem. As users are increasingly content-source agnostic and enjoy both professional and user-generated content equally, the role of aggregators is growing [Owens and Andjelic 2007]. One possible research direction for vlog aggregation is to integrate the existing video content analysis and summarization technologies into video aggregators. We will further explore these topics later.

3.2.4. Commenting and Hyperlinking in Vlogs. Typically, bloggers may post a response from others, allowing a critique of ideas, suggestions, and the like, which some argue is a defining characteristic of a blog. Meanwhile, each blog entry contains one, and often several hyperlinks to other Web sites, blogs, and stories. Usually, there is

a standing list of links, known as *blogrolls*, to the author's favorite bookmarks. Some current blogging systems even support a *linkback* functionality. A linkback is a method for Web authors to obtain notifications when other authors link to one of their documents, enabling authors to keep track of who is linking to or referring to their articles. There are three linkback methods, that is, Refback, Trackback, and Pingback, but they differ in how they accomplish this task [Wikipedia 2008]. In some sense, linkbacks are similar to incoming hyperlinks in Web pages.

Similarly, vloggers can still post their text comments on other vlogs, and put hyperlinks or blogrolls directly in the supporting text of videos. A more powerful technique is to embed outgoing and incoming hyperlinks in the video clips themselves. Embedded outgoing hyperlinks enable a video to link "out" to a Web page or another video, so as to offer some supplementary materials when the viewer watches the commentary under discussion; while incoming hyperlinks let the portion of the video under discussion be directly referenced by a specifically constructed URL pointed from another video, Web page, or email document [Parker and Pfeiffer 2005]. Currently, most popular media players such as QuickTime, RealPlayer, or Media Player can crudely support embedded hyperlinks.

Discussion. In current vlogs, video is mainly used as a powerful narrative media, and readers can only add comments to vlog entries in text format. However, if the discussions among vloggers were in video format, they would be much more lively. Researchers in the multimedia community have conducted some initial explorations on this topic. For example, Yan et al. [2004] proposed a tool called "conversant media," which is a video-based learning application that supports asynchronous, text-based discussion over a network. Comments and replies are attached to the specific locations on the time-based media that inspired them. Via this process, participants can learn from both the video content and the resulting peer-to-peer commentaries.

3.2.5. Social Tagging. Social tagging (a.k.a. folksonomy, collaborative tagging, social classification, social indexing) is the practice and method of collaboratively annotating and categorizing digital content by using freely chosen keywords (i.e., *tags*) [Jakob 2007]. Hence social tagging can be viewed as a user-driven and user-generated classification of content—both created by and used by the community of users. Users are able to trace who has created a given tag, and thereby can also discover the tag sets of another user who tends to interpret and tag content in a way that makes sense to them. The result is often an immediate and rewarding gain in the user's capacity to find related content (a practice known as "pivot browsing") [Wikipedia 2008]. Moreover, social tagging also makes it possible to establish a form of multiple, overlapping associations that resemble the brain's way of functioning instead of the rigid categories of a given taxonomy [Jensen 2007].

Social tagging is very useful for the indexing and search of online videos in vlogs. When there is not enough text description provided by a vlogger, tags created by the community of users seem to be the only searchable metadata that is simple to implement and currently available. So in the vlog hosting sites or aggregators such as Youtube, most of the search operations are based on the textual tagging of the visual content.

Discussion. Tags are subjective labels that might be misleading in semantics, and limited in representing the content of a video. Some recent studies are attempting to integrate automatic annotation and social tagging to better understand media content. ALIPR, an automatic image annotation system at http://www.alipr.com, has recently been made public for people to try and have their pictures annotated [Datta et al. 2008]. As part of the ALIPR search engine, an effort to automatically

validate computer-generated annotation with human labeled tags is being made to build a large collection of searchable images. In this way, the manifold results of automatic and semiautomatic annotation and social tagging might make the image and video data truly "the next Intel inside" (as O'Reilly [2005] calls it).

3.2.6. Vlog Search. As vlogs increase in population, how to effectively index these vlogs and make them more easily searchable is becoming a new challenge. The aim of vlog search is to make vlogs as easily searched by Web search engines as normal Web pages. Searching video clips or streams in vlogs is, however, much harder than searching textblogs. There are two main approaches. The first, and most widely used, is to employ the traditional text-based retrieval techniques, which rely on the description text or user-generated tags in RSS feeds of vlogs. Tags may describe the genre of a clip, the actors that appear in it, and so on. The tagged video can then be easily searched. The second approach uses software to "listen" to the video's soundtrack, as used by Blinkx, Podzinger, and Truveo video search engines [Wales et al. 2005]. Turning spoken dialogue into text requires fancy algorithms, and is not always reliable, as anyone who has ever used a speech-recognition system can testify. But the resulting text is then simple to search.

In addition, vlog search services may provide some advanced functions. For example, VlogMap (http://community.vlogmap.org) combines the geographic data about the location with vlog entries to provide a world map of searched vlogs.

Discussion. Ideally, search engines should index the video content automatically by scanning for embedded transcripts and timed metadata or even by performing automated analysis of the video content [Parker and Pfeiffer 2005]. However, currently few vlog search engines can support content-based video search from vlogs. Given the (still) poor capabilities of video analysis technology, it is not surprising that a video search on the Web remains rather hit-or-miss. This is indeed ongoing work for future vlog search engines, which will be discussed in detail in Section 4.1.3.

3.3. Current Vlogging Tools

Nowadays, most vlogs are powered by vlogging hosting services or stand-alone tools. These systems make it easier to edit a self-made video, to set up a vlog, to update, distribute, and archive its content, and to search vlogs and videos from widely distributed sites. Some systems are free and open sources, others are commercial products. In this installment, current vlogging tools and services are surveyed from several different, while related, perspectives.

3.3.1. Online Video Editing and Publishing Tools. Video editing is the process of rearranging or modifying segments of video to form another piece of video. The goals of video editing are the same as for film editing—the removal of unwanted footage, the isolation of desired footage, and the arrangement of footage in time to synthesize a new piece of footage [Wikipedia 2008]. With video editing softwares, vloggers are able to cut and paste video sequences and integrate audio (background music, special effects, and so forth) so as to create more attractive vlogs.

In the home video domain, there are a number of inexpensive yet powerful video editing tools such as iMovie, Abode Premiere. Ordinary users often want to create more professional videos using a set of videos that they captured. Towards this end, there are many recent studies that aim to automate home-video editing, ensuring that the edited video is of satisfactory quality. For example, Hua et al. [2004] proposed an optimization-based automated home-video editing system that could automatically select suitable or desirable highlight segments from a set of raw home videos and then

align them with a given piece of incidental music, to create an edited video segment of the desired length. Similar MTV-style video generation system can also been found in Lee et al. [2005].

The increased popularity of vlogging has resulted in a large increase in online video editing activity. Currently, the Web is populated by tools and sites that allow users to edit videos online within their browsers, without the need to install a specific software on their local computers. These video editing tools and sites, such as International Remix, Jumpcut, Videoegg, Eyespot, Motionbox, Photobucket, or One True Media, and so on [Bordwell and Thompson 2006], are generally characterized by their great ease of use, which makes it a breeze to do both basic and advanced editing even for nonexperts. Collaboratively developed by Yahoo! Research Berkeley and the San Francisco Film Society, Remix is a platform for Web-based video editing that provides a simple video authoring experience for novice users [Schmitz et al. 2006]. This system was featured at the 2006 San Francisco International Film Festival Web site for one month. Another well-known online video editing tool is the Jumpcut, a 2007 Webby Award nominee and Yahoo acquisition. Jumpcut offers very strong editing and enhancment tools, with a clean, clear interface and community features.

Online video editing and vlog publishing are often integrated into one system. Given a user-generated video, the next thing to do to set up a vlog is to publish the video over the Internet. Vlog It, developed by Serious Magic, Inc., and Broadcast Machine, developed by the Participatory Culture Foundation (PCF), are two typical examples of such tools. Vlog It makes it easy to share a video online, since all technical details like video compression, file transfer, and thumbnail creation are handled automatically. Similarly, Broadcast Machine can also be used to publish internet TV channels, including vblogs and podcasts, by using BitTorrent. For a large video, network bandwidth is always a key concern for the vloggers who do not have sufficient high bandwidth available. Keeping this in mind, Brightcove has developed a platform to allow vloggers to distribute video content of all sizes over the Internet.

Discussion. A pioneering digital filmmaker, Kent Bye, presented an excellent idea for creating a distributed, collaborative video-editing system, in which multiple individuals could actively participate in the putting together of a full-feature movie. We believe that with the advances in interactive multimedia research, more powerful techniques or tools might be developed to allow vloggers to collaboratively build more expressive vlogs in the near future.

3.3.2. Vlog Aggregators. Currently, vlog aggregators can choose from two conventional means of displaying content: Web aggregators that make the view available in a Web page, and desktop aggregators that deliver feeds through stand-alone, dedicated software on a user's desktop.

FireANT is a desktop vlog aggregator. By harnessing the power of RSS, FireANT can automate the download and display the freshest video content in vlogs through easy point-and-click interfaces, streaming media preview mode, and BitTorrent clients. Moreover, FireANT integrates Yahoo! video search, and allows vloggers to tag their audios or videos for easy categorization, search, and retrieval. Another example of desktop vlog aggregators is Democracy Player (a.k.a. Miro) developed by PCF. Aiming to make online video "as easy as watching TV," Democracy Player combines a media player and library, content guide, video search engine, as well as podcast and BitTorrent clients.

Founded in December 2004, Mefeedia is the first Web-based vlog aggregator, and is also the first and most complete vlog directory on the Web. Through Mefeedia, vloggers can easily find and watch videos in vlogs, and fill their video iPod or Sony PSP with

videos. In Mefeedia, a user can add tags to videos to help him or her as well as other people to find videos more easily. Moreover, Mefeedia also provides some advanced functions such as vertical video communities (i.e., categorizing and featuring popular videos by topic, event or location); vlogger social networks (i.e., making vloggers aware of which friends are watching); and personalized recommendations (i.e., recommending new videos, shows, and channels based on vloggers' personal interests and what they have watched). By combining satellite imagery, e-maps, and Google Earth, VlogMap is a fun and interesting vlog aggregator that shows where participating vloggers are located around the world, along with links to key information about their vlogs. So far, YouTube is the most popular online vlog aggregator. It fueled the growth of other video aggregation destinations, like MySpace Videos, Revver, and DailyMotion. Today, these sites represent general destinations that function primarily as diverse user-generated and remixed content.

Discussion. In recent years, YouTube and similar vlog aggregators celebrated huge commercial success. However, they are not necessarily navigation-friendly user environments. For this reason, a number of topic-specific and meta-aggregators have emerged in the online video space. By adding a layer of filtering over the mass of content, some *niche* or *vertical* aggregators (e.g., brand aggregators such as ON Networks and SuTree) can be developed to focus on a specific topic or an area of interest [Kelley 2008]. On the other hand, with the sheer bulk of video that is becoming available, meta-aggregators can be developed by mixing "push" and "pull" models. Give the preselected types of video content, the meta-aggregators can bring the most relevant videos to vloggers by using content-based video filtering [Owens and Andjelic 2007]. This is also called *prospective search* [Irmak et al. 2006].

3.3.3. Vlog Search Tools. Search engines are the most popular tools on the Web to help users find the information they want quickly. Similarly, vloggers can utilize search engines to easily find required vlogs and videos.

In general, most vlog directories or hosting sites are equipped with some kind of search tools or services, which are either customized according to particular needs or integrated with well-known search engines such as Google or Yahoo!. As mentioned before, most vlog search techniques are rather simple, that is, they employ the traditional text-based retrieval techniques on RSS feeds of vlogs. Similarly, most of the current video search engines (e.g., Yahoo! video search, Singingfish, Mefeedia, MSN Video search) also search the text descriptions (i.e., metadata) of all videos in vlogs for relevant results. Furthermore, some commercial search engines can extract and index close-captions for video search. For example, Google Video searches the close-captioning content in television programs and returns still photos and a text excerpt at the point where the search phrase is spoken; Blinkx builds its video search engine by combining speech recognition with context clustering to understand the actual content (spoken words) of video, thereby facilitating more semantic searches of video files [Blinkx 2007]. Table III shows a simple comparison among five popular video search engines from several aspects.

Discussion. Generally speaking, the video search engine is still in its infancy. Thus we should investigate how to exploit the existing video content analysis and retrieval technologies to facilitate video searching in the online vlogosphere. We address this issue in the next section.

Table III. Features of Existing Main Video Search Engines on the Web

Feature	Google Video	Yahoo! Video	MSN Video	Blinkx	Mefeedia
Navigation by Categories	Hot videos	Hot videos 20 genres	Hot videos, 9 categories	12 genres	Topic, event, location
Metadata-based Search	√	√	√	√	√
Close-caption based Search	√	×	×	√	×
Content-based Search	×	×	×	Speech recognition, phoneme mapping	×
Constraints of Search Results	By for-sale/ free or length	By format, size length, domain	By sources: MSN or Web	By source channels	×
Result Ranking	By relevance, popularity, rating or date	By relevance, most viewed, newest	By date	By relevance or date	By date
Related Video of Results	√	×	√	√	×
Result Views	List, grid, TV	List, grid, TV	List, grid, TV	List, TV Wall	List
Support of Video Upload	√	√	√	×	√
RSS Support	Media RSS, openSearch RSS	Media RSS	RSS 2.0	RSS 2.0	RSS 2.0
Subscription Support	×	√	×	√	×
Multi-lingual Support	√	√	×	√	×
Adult Video Filtering	√	√	×	√	Do not show thumbnails
Personalized Settings	√	√	×	×	×
Search Assist	×	√	×	×	×

4. VLOGGING 2.0: ADDRESSING THE VALUE-ADDED AND INCUMBENT ISSUES

With better user experience than textblogs and more social networking features than the traditional online video services, vlogs provide a new and rapidly evolving application in the era of broadband and mobile multimedia. However, current vlogging technology is essentially an understandable consequence of textblogging technology. It is clear that with the evolution of the vlogosphere, vlogging technology will grow synchronously.

To present a possible futurescape of vlogging technology, we would like to look at the swarms of emerging user behaviors in online video and vlogging. According to the study by Owens and Andjelic [2007], there are seven key behaviors relating to online video and vlogging: creating personal taste, seeking peer pressure, gravitating toward quality, searching for Web videos, plucking and aggregating, exploring social scenes, and DIY (directing it yourself). It was also found that for shaping online video behavior, immediate access, convenient navigation, and intelligent content discovery appear to be central; vlogging, social interaction, and creative involvement make up the other budding consumer trends. Keeping this analysis in mind, we argue that the next-generation vlogging technology (called *vlogging 2.0*) should be characterized by at least the following salient features:

—*scalability*: to support scalable online video and vlogging services so as to meet the different demands of vloggers (e.g., various video quality requirements and access bandwidths);

—*interactivity*: to support feasible user interaction in the form of interactive video, such as deep linking [Lynch 2007], interactive navigation of video content [Hammoud 2006];

—*searchability*: to enable more sophisticated indexing and search of user-generated video based on metadata, video content, and relevant deep links;

—*accessibility*: to support *any* vlogger to access *any* format of online video in vlogs using *any* device *anywhere*.

Moreover, vlogging 2.0 technology should also provide support for vloggers and vlog-hosting providers to reduce potential copyright, moral, and legal risks.

Once realized, vlogging 2.0 technology will be immensely smarter and more relevant to users, regardless of their devices or personal interests. Due to the efforts made by researchers in multimedia, we have accumulated many valuable results, ranging from content creation and understanding to user-centered delivery of and interaction with videos. Some of these results can be introduced to help make this vision possible. Similar discussions can also be found in Boll [2007], which covers what multimedia research and Web 2.0 have in common, where the two meet, and how they can benefit each other. Instead of a cursory overview of the convergence of the two broad fields, we focus mainly on two key challenges that vlogging 2.0 should tackle with the help of multimedia technology, that is, the "value-added issue" and the "incumbent issue" (see Section 2.3).

4.1. Value-Added Techniques

We will describe several techniques that can potentially be used to support the *scalability*, *interactivity*, *searchability*, and *accessibility* of vlogging 2.0, including scalable video delivery, interactive video, content-based retrieval, and user-centered content adaptation. Finally, we will discuss how to flexibly combine these techniques and systems at all stages of vlogging via the mashing-up approach.

4.1.1. Scalable Video Delivery. As mentioned in Section 3.2.1, FLV video has reached near ubiquity in the current crop of online video and vlog sites such as MySpace, YouTube, and Google Video. However, the great streaming efficiency of FLV videos is obtained at the cost of greatly degraded quality. In other words, the service providers can effectively reduce the running cost of network bandwidth by using low-quality and low bitrate FLV videos; on the other hand, users must passively accept watching low-quality videos, even some user-generated videos are of higher-quality when they were initially created and uploaded. To attract more users to vlogging, we must nevertheless improve the quality of online video while retaining reasonable streaming efficiency. This trend can also be evidenced by more and more high-definition videos recently placed online.

With the rapid growth of network broadband and that the increasingly fierce competition for broadband market share, prices continue to decline while speed is on an ever upward trajectory. On the other hand, online video delivery technologies and solutions are continuously improve via optimizing scalable video codecs and developing centralized and/or peer-assisted delivery solutions, enabling more efficient video transmission over the Internet. In short, it is now possible to distribute high-quality video efficiently and economically.

Scalable video coding. Bandwidth and QoS characteristics of the varying broadband networks suggest that online videos in vlogs will require variable, or scalable bitrate encoding schemes that leverage network intelligence to dynamically tune bitrates up or down. To effectively support scalable delivery of online video, there are three

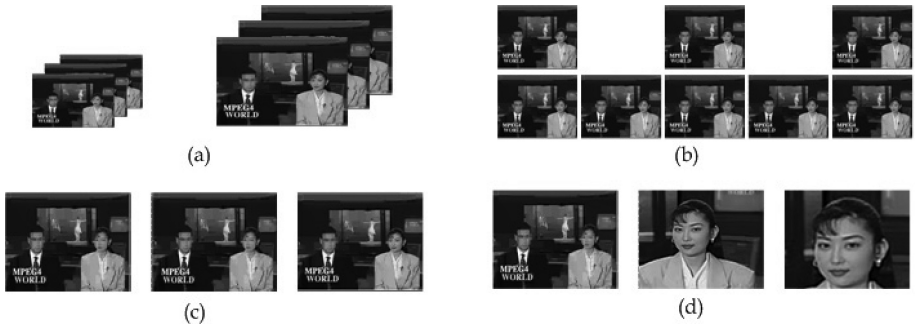

Fig. 6. An illustration of various scalable coding schemes: (a) spatial scalability; (b) temporal scalability; (c) quality scalability; and (d) attention scalability.

requirements the video codecs must meet: (1) *Be scalable*: online video should be able to scale from the smallest encoding all the way to full high-definition quality, so as to provide more flexibility in meeting different streaming demands (e.g., different access bandwidths and latency requirements). (2) *Real-time decoding*: since streaming video applications running on mobile devices or embedded in Web browsers must be simple, real-time decoding with low decoding complexity is desirable. (3) *Embedded playable*: video can be placed within HTML in today's browsers by using some simple JavaScript to load the video and control playback and interactivity using any HTML user interface [Lynch 2007].

To address the scalability issue, various video coding techniques have been proposed, including scalable video coding and multiple descriptive coding. In general, a scalable coder generates a high-quality video bitstream consisting of multiple layers, achieved through scaling frame rate, size, or quality (see Figure 6(a)–(c)); while a receiver, depending on its capability, can subscribe to the base layer only with the basic playback quality, or subscribe to additional layers that progressively refine the reconstruction quality [Liu et al. 2003; Ohm 2005; Schwarz et al. 2007]. The early scheme was called *layered coding*, used in MPEG-2 [Ohm 2005], in which the stand-alone availability of enhancement information (without the base layer) is useless, because differential encoding is performed with reference to the base layer. By data partitioning, the bitstream is separated into different layers, according to the importance of the underlying elements for the quality of the reconstructed signal. To support even more flexible scalability, a new form of scalability, known as *fine granular scalability* (FGS), was developed and adopted by MPEG-4 [Li 2001]. In contrast to the conventional scalable coding schemes, FGS allows for a much finer scaling of bits in the enhancement layer. Slightly different from scalable video coding, a multiple descriptive coder generates multiple streams (referred to as descriptions) for the source video, and any subset of the descriptions, including each single one, can be used to reconstruct the video [Castro et al. 2003; Padmanabhan et al. 2002]. The scalable and multiple descriptive coding schemes provide two flexible solutions for transmission over heterogeneous networks, additionally providing adaptability for bandwidth variations and error conditions. However, scalable coding has efficiency concerns due to the iterative motion estimation and transformation of all the layers, and often requires great computational power on the receiver's side in order to assemble and decode multiple layers [Ohm 2005].

To address the real-time decoding issue, low decoding complexity is desirable. For example, Lin et al. [2001] employed a least-cost scheme to reduce decoding complexity. However, the solution to this issue is more relevant to the implementation of decoders

by optimizing the decoding process or using decoding pipelining [Chen et al. 2005]. Besides the lightweight implementation of real-time decoding, the embedded playability can often be realized using Web-related programming languages such as JAVA.

Standardization is the key to interoperability and, by defining open interfaces, we open the doors to seamless convergence of online video. At present there are four mainstream video coding standards: MPEG-2, MPEG-4, H.264 (a.k.a. JVT, MPEG-4 AVC or AVC), and AVS. The first three standards are specified by MPEG, while the fourth is a Chinese independent formulation. Technologically, MPEG-2 is the first generation of information source standards, and other three are the second generation standards. MPEG-4 offers some FGS solutions [Diepold and Mőritz 2004], where the changing network capabilities for video delivery can be taken into account in real-time by appending enhancement layers to the video stream in case of more bandwidth becoming available. H.264 is another up-to-date video coding standard, which employs variable block-size motion compensation for intercoding and directional spatial prediction for intracoding to achieve high coding efficiency [Wiegand et al. 2003]. AVS is the abbreviation for "Advanced audio and Video coding Standard" (see http://www.avs.org.cn). Specified by a standards workgroup led by our lab, AVS is a set of integrated standards that contains system, video, audio, digital rights management, and so forth. In terms of coding efficiency, MPEG-4 is 1.4 times of MPEG-2, and AVS and H.264 are more than twice of MPEG-2. Currently, MPEG-4 and H.264 have been widely adopted in mobile phones, set-top boxes, game consoles, camcorders, and in Web video playback via clients such as QuickTime. As a newcomer, AVS has also been applied in many commercial applications such as IPTV, internet live video, and mobile TV. These coding standards are likely to become the optional video encoding formats for video on the Web.

P2P video streaming. There are more and more vlogs that employ the Web-TV-like presentation. Video embedded in Web browsers enables almost all PCs to view video reliably and consistently in a completely transparent way [Lynch 2007]. Technologically, this will pose a significant challenge to supporting platforms when large-scale high-quality videos are allowed to broadcast simultaneously. Due to its scalability, peer-to-peer (P2P) technology is an appealing paradigm, providing video streaming over the Internet for vlog service providers.

As mentioned above, P2P file-sharing technology such as BitTorrent has been applied in some vlogging tools such as FireANT to effectively enhance the file transmission performance. However, live video streaming faces several challenges that are not encountered in other P2P applications [Liu et al. 2008; Locher et al. 2007]:

—*large scale*: corresponding to tens of thousands of users simultaneously participating in the broadcast;

—*performance-demanding*: involving bandwidth requirements of hundreds of kilobytes per second;

—*real-time constraints*: requiring timely and continuously streaming delivery (although minor delays can be tolerated through buffering; nevertheless it is critical to get uninterrupted video);

—*gracefully degradable quality*: enabling adaptive and flexible delivery that accommodates bandwidth heterogeneity and dynamics.

The early P2P streaming proposal was built upon a similar notion of IP multicast, referred to as application-level multicast or end-system multicast [Chu et al. 2000]. In recent years, a large number of proposals have emerged for P2P video streaming, which can be broadly divided into two categories [Liu et al. 2008]: *tree-based* and *data-driven randomized* approaches. In tree-based approaches, peers are organized into structures

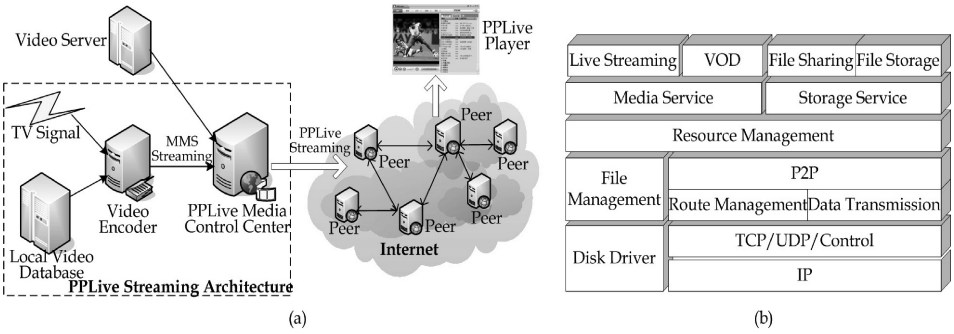

Fig. 7. An example of the P2P video streaming system, PPLive: (a) the system diagram; and (b) the peer node architecture.

(typically trees) for delivering data, with each data packet being disseminated using the same structure. When a node receives a data packet, it also forwards copies of the packet to each of its children. Data-driven approaches contrast sharply, with tree-based designs in that they use the availability of data to guide the data (e.g., gossip algorithms) instead of constructing and maintaining an explicit structure for delivering data. In a typical gossip algorithm [Eugster et al. 2004], a node sends a newly generated message to a set of randomly selected nodes; these nodes do similarly in the next round, and so do other nodes until the message is spread to all. More insights on P2P streaming technology can be found in two comprehensive surveys [Li and Yin 2007; Liu et al. 2008].

P2P video streaming is not only an active field of research, but there are already practical systems and commercial products emerging, for example, Coolstreaming [Zhang et al. 2005]; JumpTV (http://www.jumptv.com); PPLive (http://www.pplive.com); SopCast (http://www.sopcast.org); and AVStreamer [Huo 2006]. Built in March 2004, Coolstreaming has been recognized as one of the earliest large-scale P2P streaming systems. It is based on a data-centric design, in which every peer node periodically exchanges its data availability information with a set of partners and retrieves unavailable data from one or more partners, while also supplying available data to other partners [Zhang et al. 2005]. At its peak, it supported 80,000 concurrent viewers with an average bitrate at 400 Kbps with two PC servers. Since then, PPLive became one of the largest P2P streaming systems, with more than 100 channels and an average of 400,000 daily viewers [Hei et al. 2006]. Figure 7 shows the system diagram and peer node architecture of PPLive [Huang 2007]. JumpTV and SopCast are similar commercial products that have attracted tens of millions of downloads and supported several hundreds of thousands of daily viewers. Developed by our lab, AVStreamer is the first AVS-based P2P streaming media broadcast system, and currently has been applied in China Netcom and Unioncast.tv.

Discussion. Video coding and streaming are possibly the two most active fields of research and development currently. The success of Youtube confirms the mass market interest in Internet video sharing, where scalable video coding and P2P streaming serve as two underlying vehicles. For video coding, a promising research topic is to perform scalable coding by assigning more bits for regions of interest (ROIs) which might attract more user attention and less bits for other regions (as shown by Figure 6(d)) [Lu et al. 2005]; while for P2P video streaming, probably the biggest problem is that it is still unclear what the appropriate revenue model for P2P broadcast should be, with users scattered over the global Internet [Liu et al. 2008]. Nevertheless, several popular P2P streaming services, such as PPlive and SopCast, are gradually

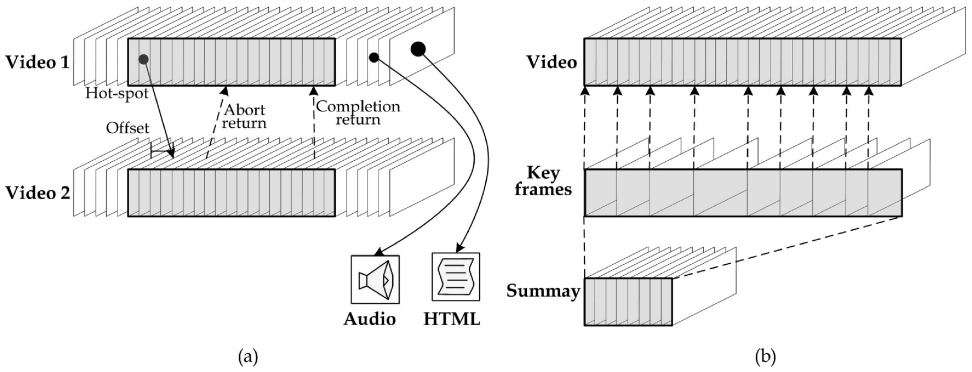

Fig. 8. Two interactive video diagrams: (a) interactive video via hyperlinks; and (b) interactive video via key frames and summary.

moving into mainstream business, suggesting that the online video business will finally live up to its hype.

4.1.2. Interactive Video. Interactive video is becoming the future standard for the most attractive video formats, which will offer users nonconventional interactive features, powerful knowledge-acquisition tools, as well as nonlinear ways of navigating and searching [Hammoud 2006]. At this point, vlogging seems to be one of best applications for emerging interactive video technology. Here we adopt from Hammoud [2006] the definition of interactive video, defined as a digitally enriched form of the original raw video sequence, providing viewers with attractive and powerful forms of interactivity and navigational possibilities. In this context, a vlog can be viewed as an interactive video document that contains not only the raw video but also several kinds of data in a time-synchronized fashion for enhancing the video content and making the video presentation self-contained. This section will address the following two major issues that concern different types of interactivity in vlogging (as illustrated in Figure 8): (1) interactive video that enriches videos by hyperlinking; and (2) video structuring and summarization for interactively navigating video content.

Video hyperlinking. *Hypervideo* (a.k.a., *hyperlinked video*) is a displayed video stream that contains embedded, user-clickable anchors, allowing navigation between video and other hypermedia elements [Smith and Stotts 2002]. For example, hypervideo might involve creating a link from an object in a video that is visible for only a certain duration, and making a deep linking that points not only to a video but to particular time positions or markers in video streams (as shown in Figure 8(a)). The initial concept regarding hypervideo can be traced back to HyperCafe [Sawhney et al. 1996], a popular experimental prototype of hypervideo that places users in a virtual cafe where the user dynamically interacts with the video to follow different conversations. Since then, many interesting experiments have followed, and some hypervideo products were developed to support interactivity in video content, such as HyperSoap [Agamanolis and Bove 1997]; Advene [Aubert and Prié 2005]; HotStream [Hjelsvold et al. 2001]; HyperFilm [Pollone 2001]; VideoClix [VideoClix 2008]; MediaLoom [Tolva 2006]; and HotVideo [HotVideo 2008]. Nevertheless, the extensive application of hypervideo has not happened yet, and the authoring tools are, at the moment, available from only a small number of providers. This situation is rapidly changing, perhaps with the unprecedented prevalence of online video sharing and vlogging. Hypervideo will introduce a new set of interactivity options that give vloggers advanced interaction and

navigational possibilities with the video content, for example, directly commenting or providing supplementary videos. Direct searching of videos could also be greatly facilitated by hypervideo models. Perhaps the most significant consequence of hypervideo will result from commercial advertising [Owens and Andjelic 2007]. Hypervideo offers the possibility of creating video clips where objects link to advertising or e-commerce sites, or providing more information about particular products. This type of advertising is better targeted and likely to be more effective (for further discussion, see Section 6.1).

Compared to hypertext, hypervideo is challenging, due to the difficulty of video segmentation, object detection, and tracking [Gelgon and Hammoud 2006]. Basically, viewers can click on some user-clickable items of interest (called *hot-spots*) in hypervideo to watch a related clip, which is often logically attached to objects or regions within the video [Smith and Stotts 2002]. In order to automatically build hyperlinks in a video, it is necessary to segment the video into meaningful pieces and to provide a context, both in space and time, to extract meaningful elements from the video sequence. Unlike the nonsemantic segmentation that aims at extracting some uniform and homogeneous regions with respect to visual properties, semantic video segmentation is defined as a process that typically partitions the video images into meaningful objects according to some specified semantics. There are many video segmentation methods in the current literature that exploit different kinds of information (i.e., spatial, temporal, spatio-temporal, and attention). These methods can be summarized simply as the implementation of four phases: object-mask generation, postprocessing, object tracking, and object-mask update [Li and Ngan 2007]. For more details, see Li and Ngan [2007] and Zhang [2006].

Once the required nodes have been segmented and combined with the associated linking information, the metadata must be incorporated into the original video for playback. The metadata is placed conceptually in layers or tracks in a video stream. Several existing standardization efforts in multimedia already allow annotations and hyperlinks to be attached to media resources. The first is the W3C's Synchronized Multimedia Interaction Language (SMIL) 2.0 [SMIL 2001], an XML-based language that allows authors to write interactive multimedia presentations. SMIL has outgoing hyperlinks and elements that can be addressed inside by using XPath and XPointer. In the strict sense, however, a SMIL document is not a single, temporally addressable time-continuous data stream [Pfeiffer et al. 2005]. The second effort is the Continuous Media Markup Language (CMML), which is an XML-based markup language for authoring annotation tracks for time-continuous data. With timed metadata and outgoing hyperlinks, it can be used to create both free-text annotations and structured metadata to describe a video's content, and can be easily serialized into time-continuous frames. For synchronized delivery of the markup and the time-continuous data over the Web, a streamable container format, called Annodex, was developed to encapsulate CMML together with the time-continuous resource [Pfeiffer et al. 2003]. Recently, MPEG has specified a new standard, Lightweight Application Scene Representation (LASeR) [LASeR 2007], as a scene description format for lightweight embedded devices such as mobile phones. The LASeR specification defines a LASeR engine which has rich-media composition capabilities based on a scalable vector graph (SVG) Tiny 1.1 and enhanced with key features for mobile services (e.g., a binary encoding, dynamic updates, fonts, and so on). To multiplex the scene stream and the synchronized media, a simple aggregation format (SAF) is provided as a streaming-ready format for packaging scenes and media together and streaming them onto such protocols as HTTP, TCP, MPEG-2, and so on [LASeR 2007].

With the hypervideo representation, there are some hypervideo authoring works (e.g., Gelgon and Hammoud [2006] and Smith and Stotts [2002]) that focus on the

automated markup of hyperlinks in a real-time video stream as well as stored video. By utilizing the hypervideo model, the MUMMY project developed a mobile interactive video tool that could be used to exchange and retrieve knowledge among members of a community in a mobile scenario [Finke 2004]. Some hypervideo authoring tools and systems were also developed to semi- or automatically create hypervideo content, including OvalTine [Smith and Stotts 2002]; Hyper-Hitchcock [Girgensohn et al. 2003]; VideoClix [VideoClix 2008]; Asterpix (http://www.asterpix.com); Adivi (http://adivi.meticube.com); AdPoint (http://www.adpointonline.com); and so on.

Video structuring and summarization. Another type of interactive video technology is to use video structure analysis and summarization to provide a representative short summary of the video prior to downloading or watching it, or to present a list of visual entries (e.g., key frames) that serve as meaningful access points to desired video content, as opposed to accessing the video from the beginning to the end [Hammoud 2006]. The two interactive video forms are useful for vloggers to preview and navigate video content in a very efficient nonlinear fashion.

For efficient access of video data, parsing video structure at different granularities is the first important step. Generally speaking, video can be hierarchically represented by five levels: key frames, shots, scenes, story units, and video. Hence video structure analysis typically involves four steps: shot boundary detection, key frame extraction, scene analysis, and story unit segmentation. As the fundamental element above the frame level, the video shot is often defined as a series of interrelated consecutive frames taken contiguously by a single camera and representing a continuous action in time and space [Rui et al. 1999]. Therefore, to build up the semantic table-of-contents structure for the video, the first step is to locate every shot by finding its beginning and end, which is often called the *shot boundary detection* task. In general, the video creators may use a variety of editing types to transfer from one shot to another, most of which can fall into one of the following four categories: hard cut, fade, dissolve, and wipe [Cotsaces et al. 2006]. Given these categories, shot boundary detection algorithms work by first extracting one or more features from a video frame or a ROI, and then using different methods (e.g., static thresholding [Cernekova et al. 2003]; adaptive thresholding [Boccignone et al. 2005; Yu and Srinath 2001]; probabilistic detection [Hanjalic 2002; Lelescu and Schonfeld 2003]; trained classifier [Cooper et al. 2007; Lienhart 2001]; and graph partition models [Yuan et al. 2007]) to detect shot changes from these features. After shot segmentation, some representative frames are often chosen as key frames to represent the scene more concisely, and are more adequate for video indexing, browsing, and retrieval. Generally speaking, key frame extraction algorithms can be briefly categorized into four classes: the simplest being sampling-based, shot-based, segment-based, and others [Li et al. 2001]. As the oft-based key frame extraction approach, shot-based methods often work by dividing one video shot into several clusters and then selecting cluster centroids as key frames. Different clustering methods can be used in this process, including discriminative [Girgensohn and Boreczky 2000] and generative clustering models [Hammoud and Mohr 2000; Liu and Fan 2005]. As shown in Figure 8(b), this process lets us organize video data according to its temporal structures and relations, hereby building a bookmark of video data.

To parse video content at a higher semantic level, many research efforts in recent years have been made on the automatic detection and recognition of scenes, highlights, and story units, especially for news, sports, and movie videos. Scene analysis and story unit segmentation play an important role in video analysis, since scene and story units usually contain intact events in the video that users are mostly interested in. In general, a story unit is defined as a segment of a video with a coherent focus that contains at least two independent, declarative clauses [Besacier et al. 2004]. Thus, story unit

segmentation is to segment the stream of data from a source into topically cohesive story units. Automatic segmentation of continuous video into constituent story units is challenging due to the diversity of story types and the complex composition of attributes in various types of stories [Hua et al. 2004]. For similar reasons, story segmentation has been one of the tasks of the TREC Video Retrieval Evaluation (TRECVID) since 2003. Many effective algorithms [Browne et al. 2003; Chaisorn et al. 2003; Hoashi et al. 2004; Hsu et al. 2004] were presented at the TRECVID workshops. Aside from this, other research efforts have been made on story unit segmentation such as those of O'Hare et al. [2004]; and Zhai et al. [2005]. In all these algorithms, multimodal processing algorithms that involve the processing not only of the video frames but also the text, audio, and speech components have proven effective. Nevertheless, in order to further extend the application scenarios, we need to develop a generic story unit segmentation algorithm applicable to videos in the nonspecific domain. Although there are some initial attempts (e.g., Lin and Zhang [2001]), developing such a general method for a story unit segment is a very difficult task.

Instead of segmenting video into a list of visual entries, video summarization aims at providing a short summary of the video prior to downloading or watching it. Similar to the extraction of keywords or summaries in text document processing [Dimitrova et al. 2002], video summarization is the process of extracting an abstract representation that compresses the essence of the video in a meaningful manner. As a valuable tool, video summarization has been receiving more and more attention in recent years. Basically, techniques in automatic video summarization can be categorized into two major approaches: static storyboard summary and dynamic video skimming. A static video summary is often composed of a set of key frames extracted or synthesized from the original video. This may be realized in a systematic manner (i.e., by taking the first frame of each shot), or in a competitive process for shot frames as described above [Borth et al. 2008; Orriols and Binefaa 2003; Zhang et al. 1997]. On the other hand, dynamic video skimming is a shorter version of the original video, made up of a series of short video clips and their corresponding audio segments extracted from the original sequence. Compared with static storyboard summary, dynamic video skimming preserves the time-evolving nature of a video by linearly and continuously browsing certain portions of video content, depending on a given time length. The simplest way to do dynamic video skimming is to generate an overview for a given video, [Lu 2004; Mahmood and Ponceleon 2001; Sundaram and Chang 2003]. Automatic summarization of video content in terms of extracting video highlights is an even more challenging research topic, since it requires more high-level content analysis [Dimitrova et al. 2002]. Highlights normally involve detection of important events in the video. A successful approach is to utilize the information from multiple sources, including sound, speech, transcript, and image analysis of video. For video highlight generation, most research has tried to find a solution for domain-specific videos where some special features can be used, like sports videos [Chang et al. 2002; Li et al. 2003; Rui et al. 2000], and news [Lie and Lai 2004; Pan et al. 2004]. However, most of these video summarization approaches are based on low-level video features. Thus they may not be able to guarantee that the generated video summary contains the most semantically important content. Video summary can also be extracted by exploiting semantic annotation [Zhu et al. 2003] or by modeling and exploiting user attention models without full semantic understanding of video content [Ma et al. 2005; Ngo et al. 2005].

Discussion. Interactivity is indeed one of the key characteristics of vlogs, in contrast to other online videos. In this section we describe two techniques from multimedia research that may be used to enhance the interactivity of vlogs. However, more research efforts must be made for further improving these techniques. For example, more

accurate object detection and segmentation may be useful to automatically build hyperlinks in a video; better story unit segmentation and highlight detection may be used to improve video summarization or to build up the semantic table-of-content for video.

Vlogging also poses some new challenges for video hyperlinking, structure analysis, and summarization techniques. First, the analysis of "unstructured" video generated by a nonprofessional user (i.e., home video) is still a challenging problem due to the relative low quality, unrestricted nature, and lack of storyline of this type of video [Abdollahian and Delp 2007]. For example, each home video clip can often be considered as a one-shot video sequence determined by the camera start and stop operations. Thus, in home video analysis, shot segmentation is not of interest. A pioneering work of unstructured video analysis was proposed by Abdollahian and Delp [2007], who made use of camera motion information for unstructured video classification and summarization. Second, a vlog entry can essentially be viewed as a document with multiple videos if some comments are also in video form. However, the optimal construction of summaries becomes much more difficult in the case of multivideos. If we try to build the summaries progressively by adding key frames one by one, it may happen that a key frame would work well for the performance of a summary, but would create many ambiguities with other previously selected key frames [Huet and Meri 2006]. The problem of multivideo summarization has not been fully investigated yet, with very few exceptions such as Huet and Meri [2006].

4.1.3. Content-Based Retrieval. As a reasonable solution borrowed from existing Web search technology (see Sections 3.2.6 and 3.3.3), current video search systems used in vlogging mostly depend on metadata such as user-provided tags, an editorially written title or summary, a transcript of the speech in the video or text extracted from Web pages. By doing so, however, they limit themselves to never actually understanding the actual videos. Therefore, some video search engines have recently begun experimenting with content-based methods such as speech recognition, visual analysis, and recognition [Blinkx 2007]. Most of these content-based retrieval systems share a common high-level architecture [Gibbon 2006] (see Figure 9), including crawler or RSS handler, video processing and analysis, indexing and summarization, search models, and query processing. Moreover, a video content adaptation component is also needed to support universal access, particularly when vloggers use mobile devices for vlogging. In this section, we mainly focus on two major modules of the system—content-based video analysis and search models; the video content adaptation module will be discussed in the next section.

Online Video Content Analysis. In general, video may be classified by its origin: user-generated, enterprise (semi-pro), and professional. Most existing video content analysis work often uses professional (or enterprise) content such as movies, news or sports videos for research. Compared with professional or enterprise content, however, user-generated videos have several special characteristics, including:

—*topic diversity*: user-generated videos may cover a wide range of topics (e.g., videos on YouTube cover many topics, ranging, in descending order from music, entertainment, people, comedy to travel);

—*short length*: the average length of user-generated videos is very short (e.g., 2 minutes 46.17 seconds, from YouTube statistics), each video, in most cases, contains only one episode;

—*microvideo*: the quality of user-generated videos is generally low, especially with the wide adoption of FLV videos;

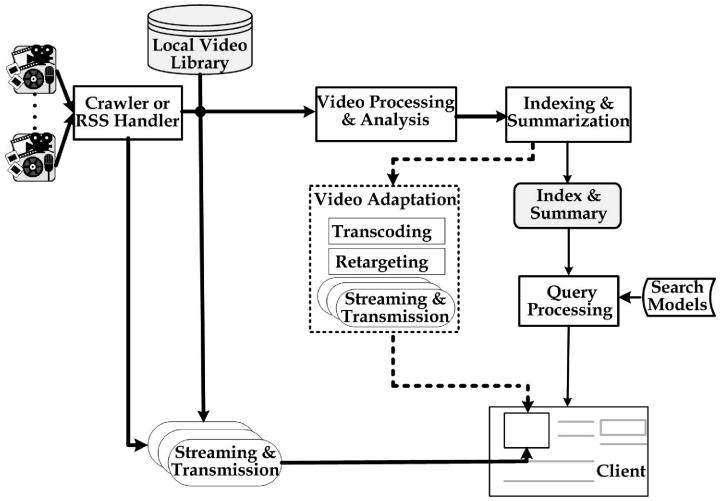

Fig. 9. Typical content-based video search architecture in vlogging 2.0; dashed lines indicate optional components.

—*Hyperlinking*: through hyperlinks, videos in related vlogs become a fully integrated, topic-specific "Web of video" [Parker and Pfeiffer 2005], which makes them different from other online videos.

Due to these characteristics, user-generated videos are often much harder to analyze and to extract meaning from than professionally produced videos.

Generally speaking, the main challenge for video content analysis is to understand media by bridging the *semantic gap* between the bit stream on the one hand and the interpretation by humans of the visual content on the other [Chang et al. 2007]. Considering the characteristics of user-generated videos in vlogs mentioned before, two additional challenges should be addressed: (1) *robustness*: video analysis models and algorithms should effectively process the low-quality microvideos, often of short length and on diverse topics; (2) *online processing*: since the size of video often makes caching from remote servers costly, and caching may also violate copyrights, it may be preferable to do the video analysis online [Gibbon 2006]. Hence, our focus here is on semantic concept detection and its application to user-generated video search. Specifically, we will discuss popular feature extraction, scene classification and object detection, event detection, and real-time analysis and indexing models.

Feature extraction is definitely the first step towards content-based video analysis and retrieval. Generally speaking, this often includes feature extraction, description, dimension reduction, and indexing. After the video is segmented and key frames are chosen, low-level visual features such as color, texture, edge and shapes can be extracted from the key frame set in video and represented as feature descriptors. There are two categories of visual features: global features that are extracted from a whole image, and local or regional features that describe the chosen patches of a given image. For most tasks, local features demonstrate better performance, in contrast to global features, since they often have better correspondence to semantic concepts or objects [Bosch et al. 2007]. Recently, some novel features such as scale-invariant feature transform (SIFT) [Lowe 2004] were proposed in order to represent visual content in a scale, rotation, illumination, and noise invariant way. Based on histograms of local orientation, the SIFT descriptor is computed on an affine covariant region around the interest point. For spatio-temporal data in video sequences, 3D-SIFT can be calculated around

the spatio-temporal (ST) interest points [Scovanner et al. 2007]. Due to the aforementioned advantages, features using local interest points such as SIFTs and 3D-SIFTs are well-suited for the analysis of microvideos. After postprocessing feature descriptors such as dimension reduction, they can be stored in the database using indexing models for future image-based queries.

To facilitate concept-based video retrieval, two effective approaches are scene classification and object detection. While a shot is the building block of a video, a scene conveys the semantic meaning of a video to the viewers. Scene classification attempts to assign semantic labels (e.g., outdoors vs. indoors, or news vs. a commercial) to video segments based on low-level features. The general approach to scene classification relies on the definition of one or more image templates for each content class [Xiong et al. 2006]. To categorize a generic scene that consists of one or more semantically-correlated consecutive shots, key frames are extracted and classified against the image template of every content class by using various classifiers such as on HMM (hidden Markov model)-based classifier [Huang et al. 2005], or SVM (support vector machine) [Zhu and Ming 2008]. To further improve scene classification and extract semantic clues for indexing, object detection and recognition is developed for some domains (e.g., faces). For example, Cotsaces et al. [2006] used the pre-extracted output of face detection and recognition to perform fast semantic query-by-example retrieval of video segments; Liu and Wang [2007] proposed an approach for automatically generating a list of major casts in a video sequence based on multiple modalities, specifically, speaker information and face information. The two approaches can be used directly for concept detection in vlogs, since a large proportion of user-generated videos are related to special scenes (e.g., entertainment, travel) or objects (e.g., people). Another notable direction for semantic labeling of visual content is to explore the relations among image content and user-provided tags or description [Chang et al. 2007]. In addition, automatic video annotation (e.g., Qi et al. [2007]) can be used to automatically assign metadata in the form of captions or keywords to a video.

As a higher-level semantic concept, events can be defined as real-world things/activities that unfold over space and time. As a kind of social activity, vlogs are usually created by vloggers as a record of events in their lives. Thus event detection plays an important role in concept-based retrieval of videos in vlogs. A general characterization of multimedia events was proposed in Xie et al. [2008], motivated by the maxim of five "W"s and one "H" for reporting real-world events in journalism: *when*, *where*, *who*, *what*, *why*, and *how*. In this framework, an event detection system can be grouped into three categories based on its problem setup and detection target: detecting known events and activities, unsupervised event detection and pattern detection, and recovering event semantics from photos and videos. In Haering et al. [2000], a three-level video event detection methodology was proposed and applied to animal-hunt detection in wildlife documentaries, which includes a feature level, neural network-based classifier level, and shot descriptor level. Video event detection can potentially profit from a combined analysis of visual, auditory, and textual information sources [Snoek and Worring 2005]. Much work has been done in event analysis with multimodal analysis, for example, in sports video [Babaguchi et al. 2002; Duan et al. 2005; Sundaram and Chang 2003] or news video [Chaisorn et al. 2003; Hsu et al. 2004]. However, semantic event detection is still a challenging problem due to the so-called semantic gap issue and the difficulty of modeling temporal and multimodal characteristics of video streams. For a better understanding of event detection, we refer the readers to a comprehensive survey in Xie et al. [2008].

By training, a statistical detector for each concept (e.g., scene category, object, event) in the visual lexicon, can create a pool of semantic concept detectors to generate multidimensional descriptors in the semantic space [Chang et al. 2007]. Then each video

segment can be represented by a semantic concept vector [Naphade et al. 2002], in which elements denote the confidence scores or likelihood values for detecting different concepts. Similar to the term frequency vector used for indexing text documents, such a representation can be used to support concept-based video search.

It should be noted that much research work on content-based video analysis and retrieval addresses indexing archived video, in which the archiving system performs the indexing as an essentially offline operation and then uses streaming technology to present the indexing results on the Web [Pieper et al. 2001]. This methodology imposes no requirements on real-time indexing performance, which in turn will play an important role in future vlogs, since live video delivery is likely to be adopted by more vloggers. In general, live streaming video retrieval offers the challenge of analyzing and indexing the live video in a single pass and in real-time because the stream must be treated as effectively infinite in length, thus precluding offline processing [Pieper et al. 2001]. Moreover, we could not simply copy the video to a local machine for preprocessing and indexing due to the proprietary issue [Wales et al. 2005]. Although there are already some works that address the real-time analyzing and indexing of broadcast news (e.g., Hilley and Ramachandran [2007]; O'Connor et al. [2001]; and Pieper et al. [2001]) or sports videos (e.g., Xu et al. [2006]), more research effort should be devoted to this issue, especially for the user-generated videos

Search models. Recent advances in video content analysis and information retrieval technology will lead to new search models in future vlog systems. Some significant characteristics of these search models include the ability to fuse different search modalities (e.g., image-based similarity retrieval, and text- or concept-based search), and context awareness, and support for persistent search, and so on.

With the help of user annotation and automatic video analysis, video search cues are available from a rich set of modalities, including user-provided metadata and tags, textual speech transcripts, low-level audio/visual features, and high-level semantic concept detectors. Thus, a *query-adaptive multimodal search* strategy can be employed to provide the most useful retrieval results [Kennedy et al. 2008]. Some multimedia video search models have been proposed recently, differing in the fusion strategies of search modalities. In Benitez et al. [1998], a meta-search framework for image and video retrieval was proposed by first disseminating incoming queries to different search engines and then combining the results with equal weights or different relative strengths that were evaluated from previous queries via a relevance feedback mechanism. A *query-independent strategy* was successfully used in TRECVID by independently applying different core search tools such as text search, image matching, and concept-based search, and combining the results through a weighted summation of either scores or ranks [Hauptmann and Christel 2004]. The weighting assigned to each of the available search methods is selected over some validation set to maximize the average performance of the system over a broad range of queries. Instead, a *query-class-dependent approach* was proposed by Chua et al. [2004], in which optimal fusion strategies are determined for each query class, either by learning weights over a training set or by hand-tuning weights. Moreover, the query classes can be dynamically created when incoming queries are received [Xie et al. 2007]. Zhang et al. [2006] described an application of query class dependency for video search in Web video collections by classifying each video in the search set and incoming queries as belonging to one or more predefined classes. Compared with query independent- and query-class-based combination methods, Yan and Hauptmann [2006] proposed a series of approaches, called probabilistic latent query analysis (pLQA), which is able to discover latent query classes automatically without using prior human knowledge, to assign one query to a mixture of query classes, and to determine the number of query

classes under a model-selection principle. Although there has been a great deal of improvement, there are still many remaining challenges, such as the acquisition of a large number of query classes and training data [Kennedy et al. 2008].

Another search model that is widely investigated and can potentially be applied in vlogging 2.0 is *contextual search*. Generally speaking, context information such as a vlogger's current time, location, state, personal history, preferences, and social context can be used to clarify his or her query and improve retrieval results. Contextual search is one of the major long-term challenges in information retrieval [Allan and Croft 2002]. Many applications have made significant strides in this direction, differing in which type of context information is exploited in the search process. By treating each interaction with the search engine as a request for further information related to the current documents, the current state of the user is inferred from the document (or video) that he or she is browsing at the time of issuing the search [Finkelstein et al. 2002; Kraft et al. 2005, 2006]. Different strategies for infusing user queries with additional context can then be used to improve search results, for example, by rewriting the user's query or reordering/filtering the initial results with a few key terms extracted from the current document context. Using this approach, Yahoo! launched its Y!Q contextual search tool in February 2005 (http://yq.search.yahoo.com) in order to use that context as a way to improve the relevance of search queries [Kraft et al. 2005]. Geographic context is another kind of context information that can be very useful for vlog search. Location context of users can be identified by allowing users to drag and drop digital content onto a map such as Yahoo! Maps, or Google Map, or even employing cell location techniques when a user takes a photograph or video with his camera-phone. In the ZoneTag system (http://zonetag.research.yahoo.com/), geographic and social context are used to assist users with labeling and sharing their personal photographs, by suggesting a set of geographically relevant tags and giving greater weight to tags used by their friends. Location context can also be used to expand and/or refine mobile search, giving preference to search results that are geographically close to the user's current location [Kennedy et al. 2008]. Google and Yahoo! have developed customized search interfaces for cell phone users based on their huge internet Web page databases. It should be noted that Wen et al. [2004] proposed a general probabilistic model for contextual retrieval to tackle incompatible context, noisy context, and incomplete query issues.

The third search model is the so-called *persistent search*. Persistent search is one of the most powerful applications of RSS, which supports subscribing to the RSS feed of a particular search engine's results for given search terms. This allows the feed reader's software to check for new search results every time a vlogger logs in, deliver them to his/her in-box as soon as they are available, or otherwise sit quietly waiting until new results become available [Rhind et al. 2005]. It is a great way to stay up-to-the-moment about issues of interest, without performing searches manually every time. Besides the RSS feeds, persistent search for vlogging 2.0 should also support subscribing to video summaries of a particular search engine's results for given queries, with the help of video segmentation and summarization tools.

Discussion. Despite significant progress, there are still a large number of open issues for content-based video analysis and retrieval. For a detailed discussion of these challenges, we refer the readers to existing surveys [Chang et al. 2007; Kennedy et al. 2008; Lew et al. 2006; Xiong et al. 2006]. From a perspective of user search behaviors, blog searches have different intentions than general Web searches [Mishne and de Rijke 2006]. However, the development of the specialized retrieval technology aimed at the distinct features of vlogs is still in its early stages. How to effectively analyze the user-generated microvideos and exploit hyperlinks between videos, and also

efficiently perform online processing for indexing, still remains challenging. We also need to tackle the still-open issue of the semantic understanding of audio and video by using contextual information that might include time and location as well as the social context or situational media usage [Boll 2007].

4.1.4. User-Centered Video Adaptation. It is expected that, in the near future, *any* vlogger can access *any* format of online video by using *any* device *anywhere*. To realize the so-called *universal multimedia access*, video adaption technology has gained much attention as an emerging field [Chang and Vetro 2005; Pereira et al. 2005]. It transforms the input video to an output in video or augmented multimedia form by utilizing manipulations at multiple levels (signal, structural, or semantic) in order to meet diverse resource constraints and user preferences while optimizing the overall utility of the video. There has been a vast amount of activity in research and standard development in this area [Chang and Vetro 2005], such as video format transcoding (i.e., transforming video from one format to another); video retargeting (i.e., cropping video for a small size display); content replacement (i.e., replacement of selected elements in a video entity); and video synthesis (i.e. synthesis of new content presentations for better navigation). Here we focus on video transcoding and retargeting that most probably will be applied in vlogging 2.0.

At the moment there are no uniformly supported video formats, but overall we want video to be accessible across platforms. Thus, transcoding technology can be used to transcode video from one format in vlogs to another in order to make the video compatible with the new usage environment. One basic requirement in transcoding is that the resulting video quality of the new bitstream be as high as possible, or as close as possible, to the bitstream created by coding the original source video at the reduced rate [Ahmad et al. 2005]. To realize format transcoding, one straightforward implementation is to concatenate the decoder of one format with the encoder of the new format. However, such implementation may not be feasible at times, due to the potential excessive computational complexity or quality degradation. Alternate solutions for reducing complexity and optimizing quality can be achieved by using transcoding optimization methods such as requantization, rate control, and mode decision [Xin et al. 2005]. In applications that involve real-time transcoding of live videos for multiple users, design of the video transcoding system requires novel architectural- and algorithm-level solutions in order to reduce the hardware complexity and improve video quality [Chang and Vetro 2005]. Although there are many commercial transcoders between different coding standards (e.g., MPEG-4, H.264), video transcoding is still an active research topic due to its high practical value for a wide range of network-based video applications. For more details about video transcoding, we refer the readers to the surveys by Vetro et al. [2003]; Ahmad et al. [2005]; and Xin et al. [2005].

Another technique tightly related to content adaptation is so-called *video retargeting* [Liu and Gleicher 2006; Wolf et al. 2007]. With the prevalence of handheld mobile devices used for vlogging, the natural gap between the high resolution of videos and the relative small display seriously disrupts the user's viewing experience. Thus, when a video is displayed on a small screen, it must be "retargeted" by cropping or scaling to better suit the target display and preserve the most of the original information (take Figure 10 as an example). Currently, video retargeting is solved in industry by several simple methods, which include cropping the surrounding part of the video and keeping the middle part untouched; resizing the whole video and adding lettering-boxes back into the upper and bottom frames; and nonuniform sampling. Although these methods can be implemented efficiently, various problems may occur. The first one may break the composition of the frame; the second one usually leads to too small frames to

Fig. 10. Example of video retargeting: (a) original 624 × 352 video; (b) a 120 × 120 video by simply resizing; and (c) a 120 × 120 video by retargeting with visual saliency.

view; and the last one may introduce distortions or wrapping affects. Thus, a smarter approach should find a better way to chop the less salient regions while preserving the most informative part for clear browsing on small displays. More recently, several video retargeting approaches were introduced to adapt video frames to small displays by effectively measuring visual saliency [Ei-Alfy et al. 2007; Fan et al. 2003; Liu and Gleicher 2006; Wen et al. 2004]. Video retargeting is a very interesting but hot research topic, but further progress is required.

Discussion. Video adaptation is an emerging field that encompasses a wide variety of useful techniques and tools for responding to the need for transmitting and consuming video content in diverse types of usage environments and contexts. However, despite the burgeoning activity and many advances, this field is in need of an analytical foundation and solutions to many challenging open issues such as automatic learning of user visual attention patterns, semantic event-based adaptation, and so forth [Chang and Vetro 2005].

4.1.5. Multimedia Mashing-up. The effective design of vlogging 2.0 solutions requires consideration of the flexible combination of multimedia systems and services at all stages of multimedia processing, from video content analysis and understanding, storage, usage, and sharing to delivery and consumption [Boll 2007]. A general response to the above issue is to develop an integrated system with modular designs of subsystems and to provide well-defined abstractions of the requirements and performance for each subsystem [Chang and Vetro 2005]. Recently, a new breed of Web-based data integration applications, that is, *mashing-up*, is growing across the entire Internet. According to Wikipedia [2008], a mash-up is a Web site or application that combines content from more than one source into an integrated experience. The data used in a mash-up is usually accessed via an application programming interface (API) or RSS/Atom feeds. The goal of a mash-up is to bring together in a single place data sources that tend live in their own data silos. There are several popular genres of mash-up applications, including mapping, video and photo, search and shopping, and news [Wikipedia 2008]. A well-known example is the use of cartographic data from Google Maps to add location information to real-estate data, thereby creating a new and distinct Web service that was not originally provided by either source.

Like Web portals, mash-ups can also be viewed as content aggregation technologies. The technologies that facilitate the development of mash-ups include XML-based syndication formats such as RSS and Atom, Ajax (asynchronous JavaScript and XML) for creating interactive Web applications, platform-neutral protocols for communicating with remote services such as SOAP (services-oriented access protocol) and REST (representational state transfer), screen scraping for data acquisition, and so on [Merrill 2006]. A mash-up application is architecturally comprised of three participants that are logically and physically disjoint [Merrill 2006] as follows: the API/content providers that provide the content being mashed; the site where the mash-up application is hosted; and the client's Web browser where the application is rendered graphically and where user interaction takes place.

As mash-up applications become more feature- and functionality-rich, mash-up development is replete with technical challenges that need to be addressed Merrill [2006]), such as semantic meaning and data quality, the Ajax model of Web development, the protection of intellectual property and consumer privacy, and especially the development of "multimedia mash-ups." Initial attempts have been made to reuse, combine, and merge different multimedia services into a new and attractive application. In Spoerri [2007], the searchCrystal tool set is proposed to visualize Web, image, video, news, blog, and tagging search results in a single integrated display when geographical meta-data is not available. In Tuulos et al. [2007], a story mash-up is designed to combine the textual stories written in the Web by some people with matching photos taken by other people to form complete stories. By using the Hop programming language, a photograph gallery and a podcast receiver are implemented by Serrano [2007] with a few lines of code. Youtube and VlogMap can also be viewed as two mash-ups for vlogging, hosting Google Maps (or Google Earth) to show the geographic location where the video was shot.

Discussion. As an emerging technology of Web 2.0, mashing-up has recently attracted much attention. However, the mashing-up applications are still in their infancy, most of them currently emphasize only how to combine and superimpose diverse data sources on a map. The question remains how multimedia mashing-up can be done more effectively.

4.2. Incumbent Techniques

As vlogging becomes more popular, one of the biggest social issues facing vloggers and vlog hosting sites is the tradeoff between the protection of copyright and consumer privacy versus the fair-use and free flow of information, mostly because the copyright issue has been widely used as a potential competitive weapon against vlog-hosting providers by existing disseminators [Meisel 2008]. Meanwhile, according to local laws, regulations, and policies, certain videos should also be filtered out, for example, to protect kids from pornographic videos. This installment will describe two optional "incumbent techniques" that can potentially be used for vlogging 2.0 to reduce such legal, economic, and moral risks.

4.2.1. Digital Rights Management. In early 2006, a short video titled "The Bloody Case That Started From A Steamed Bun" was red hot on the Internet in China. This video was remade by a vlogger from "The Promise," a popular Chinese commercial movie, thereby raising a wide-ranging debate on online video copyright in China. The vlogs give us the opportunity to document our lives, provide commentary and opinions, express emotions, and freely articulate ideas through vlogging in an expressive video content; but this does not mean that vloggers could freely distribute anything that might be related to private or commercial rights, with no authorization. Another

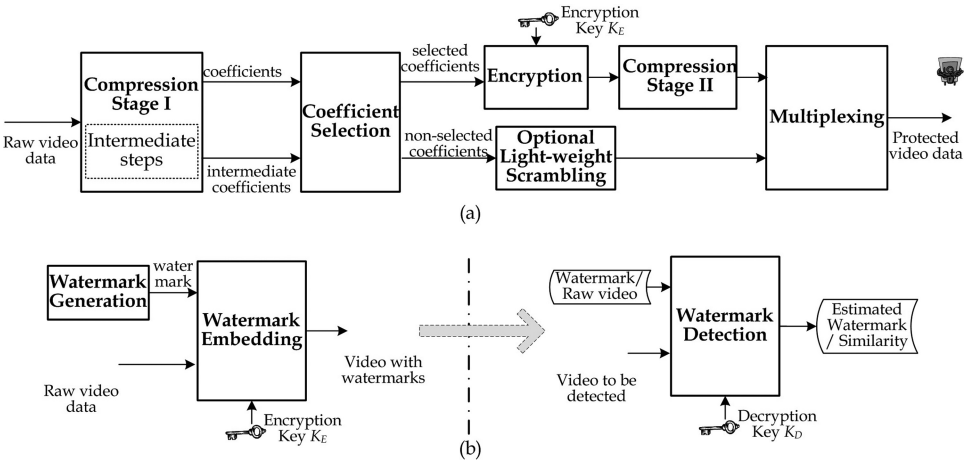

Fig. 11. Overview of two typical DRM technologies: (a) selective video encryption; and (b) video watermarking.

well-known example is Viacom's $1 billion lawsuit against YouTube in March 2007 [Peets and Patchen 2007]. Effective technical means must be used to prevent the emergence of such negative phenomena, including *digital rights management* (DRM) technology [Lin et al. 2005].

A DRM system protects and enforces the rights associated with the use of digital content. The primary objective of DRM is to ensure that access to protected content (such as video) is possible only under the conditions specified by the content owner [Lin et al. 2005]. To fulfill such an objective, a common approach is to use encryption technology to protect video files or streaming videos from unlicensed distribution. Encryption is the process of controlling access to confidential data, known as *plaintext*, by scrambling the data into an unintelligible form (i.e., *ciphertext*) via an *encryption key*. The inverse process, known as *decryption*, is very easy to perform with knowledge of the decryption key, yet very difficult to perform without it. Generally speaking, if data to be protected is encrypted, then only licensed devices can access it. Different from traditional data encryption, the characteristics of a video stream, such as bit-sensitivity, determine that several issues should be taken into account in video encryption [Zeng et al. 2006], such as complexity, content leakage (or perceptibility), compression efficiency overhead, error resilience, adaptability and scalability, multilevel encryption, syntax compliance, and content-agnostic discuss. In particular, an encryption approach that may be beneficial for vlogs is to use selective encryption of video data. As shown in Figure 11(a), the goal of selective encryption is to encrypt or scramble only a portion of the video data, such that the displayed video quality will be unacceptable if decryption is not performed [Liu and Eskicioglu 2003]. Besides reducing the computational cost for video encryption, selective encryption may also allow a video decoder to identify structures in the compressed video data (such as headers, timing, and synchronization information) without decryption, so that efficient access to specific portions of the video is desired, and decryption is necessary only when these specific portions are displayed [Lin et al. 2005].

Encryption techniques do not offer any protection once the encrypted data is decrypted. As a complementary solution, *digital watermarking* is proposed for copyright violation problems (see Figure 11(b)). The watermark is a signature embedded in the video content, which, in addition to being invisible or inaudible to humans, should

also be statistically undetectable and resistant to any attempts to remove it [Doërr and Dugelay 2003]. The embedded watermark may be detected by a watermark detector, which uses signal processing to obtain the watermark from the watermarked video. A DRM system can use watermarking in a variety of ways such as copyright or owner identification, copy protection, access control, content-tracking, fingerprinting, or traitor-tracing [Lin et al. 2005]. As a derivative of the digital watermarking technique, *digital fingerprinting* was proposed for tracking rates by embedding the relevant information about the copy consumer rather than the signature of the content owner [Lee and Yoo 2006].

Discussion. As two active security mechanisms, both encryption and watermarking are only feasible in a controllable, constrained scope of applications. However, they fail to solve the copyright protection issues in the open, uncontrollable Internet environment. More recently, the *video signature*, sometimes called *video DNA*, has attracted much attention in MPEG [2007]. Video signature is a discriminating feature which can be used to uniquely represent the visual content of a video clip. Generally, different video clips must have distinguishable signatures, while videos generated from the same source must have the same, invariant signature. Thus, the video signature can be extracted to identify the video content and then discriminate whether different videos were generated from the same source.

4.2.2. Content-Based Filtering. According to local laws, regulations, and policies, some portions of a video should also be filtered out automatically. For example, a child is not allowed to watch violent/adult sections of a video. In this sense, video filtering should also be an *optional* incumbent technique for vlogging 2.0. Many content-filtering functionalities can be done via textual annotations attached to the content by authors. Some video sites also use image-based adult content detection (e.g., Rowley et al. [2006]) on thumbnail images to indicate the pornographic content attribute of videos in vlogs. However, a thumbnail image is far from presenting all the content of a video. Hence video filtering systems should be built upon more advanced video analysis functions.

Video content filtering is an application of video content analysis. After a video is segmented into different key frames or shots, individual segments can be classified into different categories based on audio-visual or close-caption information, and such a classification can be used to filter the content efficiently [Krishnamachari et al. 2000]. In this process, various image-based filters can be used, such as skin-colored filters [Fonseca and Pereira 2004]; skin-dependent and independent filters [Rowley et al. 2006]; shape-based filters [Zhang et al. 2006; Zheng et al. 2006]; blank frame detection-based filters [Dimitrova et al. 2002]; and multiple features-based filters [Zeng et al. 2005]. These filters can also be integrated into content filtering mechanisms (e.g., to inhibit the exhibition of scenes with nudity) or more advanced vlog recommendation systems. Moreover, if the automatic video summarization system creates an MPEG-7 description of the summary, based on the MPEG-7 description of the content and on criteria specified by a summary producer or by the consumer himself, then the filters can act directly on the MPEG-7 description, rather than directly analyzing the video content in question [Fonseca and Pereira 2004; Krishnamachari et al. 2000].

Video content filtering can also be used for video recommendation [Angelides 2003]. With more and more vlogs being placed online, it is impossible for users to manually flip through all the vlog channels to find videos of interest. Therefore, it is necessary to automatically filter the video content that is not in accordance with the users' interests, and at the same time recommend relevant videos or vlogs according to the users' current viewing experiences or preferences. Many existing vlog or video-sharing sites,

such as YouTube, MySpace, Yahoo!, and Google Video, already provide video recommendation services. However, all these existing recommendation techniques largely depend on collections of user profiles and the available surrounding text descriptions (e.g., title, tags, and comments). Instead, video recommendation was formulated in Yang et al. [2007] as finding a list of the most relevant videos in terms of multimodal relevance, and relevance feedback was adopted to automatically adjust intraweights within each modality and among different modalities via users' click-though data. An interactive approach was also developed in Fan et al. [2008] to enable personalized news video recommendation by seamlessly integrating the topic network and hyperbolic visualization to achieve interactive navigation and exploration of large-scale collections of news videos at the topic level. Although a more efficient recommendation system is still a challenging research problem, future vlog aggregators would be able to automatically recommend videos or sections of videos to users.

Discussion. The challenges of controlling access to specific videos have grown as quickly as vlogging sites such as YouTube have increased in popularity. Currently, many vlog hosting sites and video search engines have taken some SafeSearch approaches to screen offensive content. For example, Yahoo SafeSearch is designed to filter out the adult-oriented text, video, and images. However, even though SafeSearch is activated during a video search by using the word beaver, Yahoo nevertheless return several pornographic results. This indicates that video content filtering should be improved further. Meanwhile, we can depend on the community of vloggers to monitor the quality of the posted videos. By exploiting so-called *group influence*, collaborative filtering (CF) approaches are used to filter out information or patterns by using techniques involving collaboration among multiple agents [Cho et al. 2007]. The combination of content-based and collaborative filtering techniques is expected to achieve a better performance.

4.3. Summary

In this section we present some possible technical solutions to support future vlogging, and give a simple review of the related major techniques. In particular, we argue that scalable video delivery, interactive video, content-based video retrieval, and user-centered video adaptation techniques are potentially helpful in providing vlogging 2.0 with better scalability, interactivity, searchability, and accessability, and the incumbent techniques such as DRMs and content-based filtering can potentially be used in vlogging 2.0 to reduce the legal, economic, and moral risks. Here we do not mean that all these techniques will be used practically in future vlog systems. However, it is highly possible that by adopting these techniques, future vlogging 2.0 will become more powerful and compelling.

On the other hand, vlogs and vlogging also introduce many new challenges for multimedia research, some of which are described in this section. These challenges will in turn bring new driving forces to multimedia research. In short, vlogging and multimedia can benefit each other, and it is possible to integrate their valuable results and best practices.

5. VLOG MINING: TOWARDS HARNESSING COLLECTIVE INTELLIGENCE

With a large number of vlogs placed online, the community surrounding vlogs and vloggers (a.k.a. the *vlogoshpere*) becomes as important as the videos themselves. In the highly self-referential vlogging community, vloggers pay attention to other vloggers and thus magnify their visibility and power. To harness such collective intelligence (called the *wisdom of crowds* by O'Reilly [2005]), we should devote our research efforts

to analyzing, understanding, and further utilizing the vlogosphere (i.e., vlog/vlogger community). Some challenging questions are how to discover the potentially useful patterns from the temporal *Web of video* that is formed by hyperlinked videos in vlogs; what is the structure of the vlogosphere from the perspectives of both blogger relationships and information diffusion; how does the vlogosphere evolve; and so on. The in-depth investigation to these problems may hint at future killer applications and experiences.

5.1. Mining on Blogs and Vlogs

The explosive growth of vlogs (and blogs) is generating large-scale data sets that we cannot afford to ignore. In recent years, many algorithms [Chang and Tsai 2007; Shen et al. 2006; Tsai et al. 2007; Wu et al. 2008] have been explored for text mining, Web mining, and statistical analysis can also be adopted to blog mining. However, the blogosphere, and particularly the vlogoshpere, differ from the general Web in that they consist of highly dynamic and dated content that could benefit more from the specialized mining techniques. For example, videos in vlogs may contain incoming and outgoing hyperlinks, allowing the video to become a fully integrated part of the Web [Parker and Pfeiffer 2005]. To the best of our knowledge, this special characteristic of vlogs has not been fully investigated by researchers in text mining, Web mining, and multimedia retrieval.

5.1.1. Blog mining. Blog mining tasks, such as blog popularity or authority-ranking, important blogger discovery, blog spam detection, and automatic trend discovery have attracted the attention of researchers. Although built upon blogs, some approaches can also be easily adopted to vlog mining.

Blog ranking. For the blog-ranking problem, a direct solution is to measure the influence, popularity, or authority by counting hits or citations equally. There are a number of Web services that track blog interconnectivity in the blogosphere and provide a popularity ranking based on the counting approach. For example, BlogStreet developed a blog influence quotient (BIQ) to rank a blog's importance or influence, such that if a blog with a high BIQ blogrolls site x, then the BIQ of site x goes up (and vice versa); VlogMap uses the circulation to measure feed readership and popularity, and uses hits to provide a detailed look at the readership of vlogs, where circulation is recorded over the life of a feed and hits are a raw measure of requested traffic for a feed. Due to its simplicity, the counting approach cannot capture the amount of "buzz" a particular post or blog creates, and also cannot reflect the evolutionary trends of bloggers' interests. Therefore, better blog-ranking approaches should reflect the dynamic nature of the blogosphere. In Adar et al. [2004], a PageRank-like blog-ranking algorithm, *iRank*, was proposed to measure how important blogs are for propagating information. Similarly, a HITS-like blog-ranking algorithm, *EigenRumor*, was proposed in Fujimura et al. [2005] to score each blog entry by weighting the hub and authority scores of the bloggers based on eigenvector calculations. Recently, some blogging behavior features were introduced into blog-ranking approaches, such as blog update frequency and the comment numbers [Wetzker et al. 2007], and conversation mass [Mcglohon et al. 2007].

A related blog mining task is the blogger's rank. As mentioned previously, EigenRumor can effectively identify the good or important bloggers [Fujimura et al. 2005]. An algorithm was proposed in Nakajima et al. [2005] for analyzing blog threads (each thread represents a set of blog entries comprising a conversation) and identifying the important bloggers, including agitators who stimulate discussion, and summarizers who provide summaries of the discussion. In Shen et al. [2006], three approaches, namely cosine similarity-based, topic-based, and two-level similarity-based methods,

were designed to discover the latent friends of bloggers based on the content of their blog entries.

Blog trend discovery. Trend discovery may be one of most important tasks in blog mining, since the basic feature of blogs is that they consist of highly dynamic and time-stamped entries on different topics such as personal lives, politics, culture, and the like. Although the blogosphere is biased towards a particular cross-section of society, identification of trends within the blogging world is a way to take the pulse of its contributors and of a segment of society [Glance et al. 2004].

As the first trend-discovery algorithm in blogs, BlogPulse executes a set of data-mining algorithms, including phrase-finding, people-finding and identification of key paragraphs to discover aggregate trends characterizing the past days' publications to the blogosphere [Glance et al. 2004]. The site BlogPulse.com also implemented a trend search that iterates the same search query for all dates within a specified date range, bins the counts into time buckets, and plots the result. Similarly, Anjewierden et al. [2005] also implemented the identification of knowledge flows in a tool called BlogTrace. BlogTrace uses the semantic Web technology to represent blogs and linkage structures; a term-extraction procedure to determine concepts in natural language entries; and a statistical algorithm to compute a relevance metric for extracted concepts and their relations to other concepts for any particular blog author. The *Kanshin* system [Fujimura et al. 2005] was developed for understanding the concerns of people in blogs, by identifying five concern patterns such as periodic pattern, gradual increase pattern, sensitive pattern, trailing pattern, and others. There are also many works that apply text topic-detection methods to discover topic trends in blogs. For example, probabilistic latent semantic analysis (pLSA) was used in Tsai et al. [2007] to detect keywords from various corporate blogs with respect to certain topics; a probabilistic model was proposed in Li et al. [2005] to incorporate both content and time information in a unified framework for retrospective news-event detection; the *eigentrend* concept was introduced in Chi et al. [2006] to represent the temporal trend in a group of blogs with common interests using singular value decomposition (SVD) and higher-order SVD.

Blog spam detection. With the massive increase in the number of blogs, comment spams have proliferated. A comment spam is essentially a link spam originating from comments and responses that support dynamic editing by users [Mishne et al. 2005]. Comment spams can easily be made into blogs. For example, a spammer writes a simple agent that randomly visits blogs and posts comments that link back to the spammer's page. Comment spam, and link spam in general, poses a major challenge to search engines, as it severely threatens the quality of their ranking, while traditional email spam-detection techniques by themselves are insufficient for the blogosphere [Mishne et al. 2005].

Usually, the spammers create links between sites that have no semantic relation, for example, a personal blog and an adult site. This divergence in the language models can be exploited to effectively classify comments as spam or nonspam. Such a language-modeling approach was used in Mishne et al. [2005] to detect link spams in blogs. Similarly, a blog spam-detection algorithm was proposed in Kolari et al. [2006] by using SVMs. In the experiments, the detection accuracy of blog spams achieved 88%. By exploiting the temporal regularity of spam blogs in both content and linking patterns, a spam blog-detection approach was proposed in Lin et al. [2007] with a 90% accuracy. We can further examine different statistical spam-detection algorithms (e.g., Zhang et al. [2004]) to more effectively identify spams in blogs.

Discussion. There are also some other mining tasks associated with blogs and blogospheres such as blogger behavior analysis [Furukawa et al. 2007], opinion-mining

[Attardi and Simi 2006], and blog summarization [Chang and Tsai 2007]. Theoretically, most of these mining algorithms can be applied to vlog mining by replacing text feature extraction with the fusion processing of text and video data. For example, an opinion-mining algorithm was proposed in Kitayama and Sumiya [2006] by exactly integrating blogs and news video streams. However, vlog-mining tasks are in general much more complex than those in textual blogs, since vlogs take video as the primary content, often with just few accompanying texts and additional metadata to provide context. As discussed in Section 4.1, this requires robust video content analysis algorithms (e.g., video-event detection).

5.1.2. Hypervideo Link Analysis. Motivated by the successful data-mining algorithms and the tremendous appeal of efficient video database management, various video-mining approaches have been proposed on video databases, including special pattern detection [Fan et al. 2002; Matsuo et al. 2003]; video clustering and classification [Pan and Faloutsos 2002]; and video association mining [Zhu et al. 2005]. However, most of these approaches have not addressed the special characteristics of online videos in vlogs such as hyperlinking. Analyzing the hyperlink patterns will be beneficial to topic detection, summarization, and retrieval of videos. To the best of our knowledge, there is no existing work on hypervideo link analysis. Thus we review below some closely related work that exploits link-attached video for video retrieval or video summarization.

The temporal feature of blogs makes them ideal data sets for temporal link analysis, that is, utilizing the time dimension in the context of link analysis [Amitay et al. 2004]. A new term, *dated-inlink*, was introduced by Amitay et al. [2004] to denote an ordered pair (u, t), where u is a URL that was last modified at time t and which links to page p. The benefits of dated-inlinks are demonstrated in several applications such as the activity measure within a topical community, topic trend detection, and link-based ranking schemes that capture timely authorities. However, a formal model of such timely authority was not given in Amitay et al. [2004]. Instead, temporal aspects (e.g., freshness, rate of change) were introduced by Berberich et al. [2005] into the well-known PageRank algorithm, leading to two variations of temporal PageRanks, *T-Rank Light* and *T-Rank*. These works have built a starting point for temporal link analysis on vlogs.

Without considering the embedded hyperlinks in videos, there are two kinds of links attached to multimedia objects: hyperlinks that point from pages to multimedia files, and embedded relations that indicate a multimedia object (e.g., an image, audio, or video) embedded into a page. By making use of the two kinds of hyperlinks, a modified multimedia PageRank algorithm was proposed for multimedia retrieval on the Web [Yang and Chan 2005]. Based on a "shots and links" view of the continuous video content, a measure of the interestingness or importance of each video shot, *ShotRank*, was proposed in Yu et al. [2003] to organize the presentation of video shots and generate video skims. This hyperlink analysis algorithm may be easily applied to vlogs, in which the "shots and links" view can be directly realized by embedded hyperlinks in videos.

Discussion. Work in hypervideo link analysis has been very limited, even though it is of practical importance. One reason for this could be that the goals and possible applications of hypervideo link analysis are not defined as clearly as hyperlink analysis for Web search. On the other hand, what makes hypervideo link analysis different from the traditional hyperlink analysis is that the analysis must be closely integrated with online video analysis such as event detection in user-generated videos, which is still recognized as an extremely difficult question by the multimedia community.

5.2. Community Analysis of the Vlogosphere

Blogging in general and vlogging in particular have been turned into powerful tools for establishing and maintaining online communities. It is the perception that bloggers exist together as a connected community (or as a collection of connected communities) or as a social network [Wikipedia 2008]. The blogosphere provides an interesting opportunity to study social interactions, including the spread of information, opinion formation, and influence. Below, we discuss two major research topics regarding the vlogosphere (and the blogosphere in general): the structure and evolution models, and information diffusion models.

5.2.1. The Structure and Evolution Models. The blogosphere is extremely buzzy and dynamic, in which new topics emerge and blogs constantly rise and fall in popularity. Thinking about the blogosphere as a dynamic social network would help us gain a deeper understanding of how it works. In this sense, the structure and evolution models may be a good starting point for blogosphere analysis.

In Kumar et al. [2004], the blogosphere was modeled as three layers: the individual bloggers, a friendship web between pairs of bloggers with shared locations and/or interests, and the evolution of blog communities. Keeping these in mind, some structure characteristics of the blogosphere were analyzed in Kumar et al. [2004] based on the profiles of 1.3 million bloggers, such as the geographic distribution structure, the age and interest structure and the burstiness of blogs, and there were some fascinating insights into blogosphere evolution and blogger behavior. In Chau and Xu [2007], three network analysis methods were used to mine communities and their relationships in blogs, including topological analysis for ensuring that the network among bloggers is not random; centrality analysis for identifying the key nodes in a network; and community analysis for identifying social groups in a network. In Kale [2007] and Joshi et al. [2007], a polar social network graph was generated in the context of blogs, where a vertex represents a blogger and the weight of an edge represents the sentiment (e.g., bias/trust/distrust) between its connecting vertices (the source and destination bloggers). This sentiment was called *link polarity*, and was calculated by using the sentiment of the text surrounding the link. In Chen et al. [2007], the blogging behavior over the blogosphere was modeled from multiple dimensions: temporal, content, and social, assuming that these dimensions are dependent on each other.

It is widely assumed that most social networks show a "community structure," that is, groups of vertices that have a high density of edges within them, with a lower density of edges between groups [Newman 2003]. Researchers studying the theory of social networks calculate the "clustering coefficient" of a network of friends, defined as the chance that two of one person's friends are themselves friends. It was shown in Kumar et al. [2004] that the clustering coefficient is 0.2 for the analyzed network of bloggers, meaning that at a remarkable 20% of the time, two friends of the same blogger are themselves friends. This *small-world effect* has obvious implications for the dynamics of information propagation processes taking place in the blogosphere.

Discussion. Without using content analysis to model the community structure, the social network methods mentioned above can be used directly for the vlogoshpere. However, the problem will become much more complex if we consider the interrelations among the content, social, and temporal dimensions of vlogs. For example, the identification and characterization of sentiment in videos remain challenging problems. We also lack further analysis and large-scale experiments to show whether the well-known *six-degrees of separation principle* is still valid in the vlogosphere, since the change of media forms (e.g., from text-only to video) may shorten the information

propagation chain, and consequently the friendship network. This should be a topic for future research.

5.2.2. Information Diffusion Models. Blogs link together in a complex network through which new ideas and discourse can flow. Through such a network, bloggers influence each other and their audience by posting original content and commentary on topics of current interest. In this sense, the blogosphere can be viewed as an information network [Newman 2003] or an influence network [Tian et al. 2003]. Information diffusion models focus on how the information diffuses in a blog network, not the content itself. Such models are very important for determining the popularity of blogs in the blogosphere and exploring potential applications for blogs. For example, blogs (and particularly vlogs) offer an excellent, inexpensive, and nearly real-time tool for online advertising (e.g., evaluating the effectiveness and health of a company's image and image-affecting activities [Gruhl et al. 2004]).

The core idea of information diffusion models in the blogosphere is that the spread of a piece of information through a blog network can be viewed as the spread of disease or the propagation of an innovation through the network [Adar et al. 2004; Gruhl et al. 2004]. In disease-propagation models, a disease spreads from one seed node only if the transmission probability is larger than an epidemic threshold, and an epidemic spreading throughout networks typically follows a power law in which the probability that the degree of a node is k is proportional to $k^{-\alpha}$, for a constant α with $2 \leq \alpha \leq 3$ [Moore and Newman 2000]. Many real-world networks have this property, including the network defined by blog-to-blog links [Kumar et al. 2004].

In social network analysis, the diffusion of innovation has been studied extensively, typically including the threshold models and the cascade models. In the threshold models [Granovetter 1978], a node u is influenced by each neighbor v according to a nonnegative weight such that $\sum_{v \in Neighbor(u)} \omega_{u,v} \leq 1$, and u adopts if and only if $\sum_{\text{adoptors } v \in Neighbor(u)} \omega_{u,v} \leq \theta_u$, where θ_u is a threshold for u uniformly at random from the interval $[0,\ 1]$. In the cascade models [Goldenberg et al. 2001], a node u adopts with some probability $p_{v,u}$ whenever a social contact $v \in Neighbor(u)$ adopts. Based on these models, a microscopic characterization model was developed in Gruhl et al. [2004] to model topic propagation from blogs to blogs in the blogosphere; meanwhile, a macroscopic characterization model was also presented to describe topic propagation through the collected corpus, formalizing the notion of long-running "chatter" topics consisting recursively of "spike" topics generated by outside world events, or, more rarely, by resonances within the community. The results showed that the popularity of chatter topics, remains constant in time while the popularity of spike topics is more volatile. Kumar et al. [2003] analyzed community-level behavior as inferred from blogrolls, that is, permanent links between "friend" blogs. Analysis based on thresholding as well as alternative probabilistic models of node activation was considered in the context of finding the most influential nodes in a network.

Discussion. Although there is a rich literature surrounding propagation through networks, we still need to devote more research effort into information diffusion modeling and analyzing in the vlogosphere. The study of information diffusion in the vlogosphere may give new insights into the commercial applications of vlogs in online advertising and viral marketing. For example, the propagation model of information through a vlog network may give ad agencies suggestions on how to change their strategies in time so as to seize the attention of potential customers. This topic will be explored further in the next section.

5.3. Collaborative Video Recommendation: An Example Application of Vlog Mining

In the final portion of this section, we will simply discuss an example application of vlog mining—*collaborative video recommendation*. As a surrogate to content-based video search and filtering that are currently still in their infancy, collaborative video recommendation is one of the most important tools in finding video content [Adomavicius and Tuzhilin 2005]: 53% of online video searchers discovered online video content through friends [Owens and Andjelic 2007]. Further evidence is that YouTube uses the active sharing function to provide more community and customization features that can potentially facilitate user recommendation. This function enables connecting to users who are watching the same video and also seeing what they have watched most recently.

Traditionally, the recommendation of videos was based on user-viewing patterns or preferences (e.g., similar tags, keywords, or categories), by automatically filtering the video content that was not in accordance with the user's interests and profile. However, this content-based recommendation relies heavily on video content analysis technology and lacks serendipity, that is, content that falls outside the user profile could be relevant to a user but won't be recommended [Angelides 2003]. Instead, collaborative recommendation ignores descriptions and focuses on ratings of items by users of the community. With the "birds-of-a-feather" assumption that people who liked certain things in the past will like similar things in the future, this model incorporates people's tastes without relying on content, which in some cases, such as music or movies, might be hard to compare [McDonald 2003]. Therefore, collaborative video recommendation can be viewed as a direct application of vlog mining and community analysis. For example, with the help of vlog mining, we can construct vlogging-behavior models and vlogging-behavior prediction systems, which can then be used to create a recommender system that can help people find the videos they prefer potential academic collaborators, and so on [Chen et al. 2007].

6. INCENTIVE APPLICATIONS: MOVING BEYOND VLOGS

According to In-Stat, the potential market for online video content worldwide will grow rapidly in the coming years [Owens and Andjelic 2007]. Although the market for online video products is already crowded, it is still in its early stages. Thus the growth of incentive applications, and consequently the extension of vlogging technology, might be one important issue for vlogging 2.0. This may in turn infuse new drives for multimedia research. In this section, we discuss two areas in which significant opportunities may lie: user-targeted video advertising and collective intelligence gaming.

6.1. User-Targeted Online Video Advertising

According to an analysis report by eMarketer [Verna 2007], user-generated content sites attracted 69 million users in 2006 in the USA alone, and generated $1 billion in advertising revenue in 2007; by 2011, they are projected to attract 101 million users in the USA and earn $4.3 billion in ad revenue. Data like this is truly encouraging. As a hybrid of user-generated video and social networking, vlogs present a significant opportunity for advertisers to extend the reach of their campaigns with compelling content. While all advertising on the Web is interactive by nature, vlogs offer a unique and more complex level of engagement with their precise targeting capability [IAB 2008].

In general, there are two ways to advertise in vlog sites: by placing commercial messaging in and around the content or by becoming a part of the content itself [IAB 2008]. To incorporate advertising into user-generated video, one early common method was

"pre-roll" video—a short ad that runs before the video itself. Although pre-roll video ads are still common, some sites, including YouTube, now prefer the "overlay" video ads that pop up every tens of seconds into a video waiting for a user's click, and only cover the bottom of the screen so that the advertising doesn't interrupt, clutter, or delay the user's experience [IAB 2008]. In addition, some other methods are also being used for advertising on online video or vlog sites, such as conversation-targeting, custom communities, dedicated channels, brand profile pages, branding wrappers, and widgets [IAB 2008]. For example, a number of sites now offer conversation-targeting services that allow advertisers to place their Web ads next to relevant conversations, or identify those blogs and Web sites that most frequently discuss the advertisers product or category in the most favorable terms.

For video advertising to remain successful, advertisers must truly engage the audience and create an environment and viewing experience where advertising is welcome. Namely, if advertising is relevant to a user's activity and mindset, this active involvement can carry over from the content to the marketing message; if not, it is likely to be regarded as an annoying interruption. Hypervideo offers an alternative way to enable the precise targeting capability of online video advertising in vlogs, allowing for the possibility of creating video clips where objects link to advertising video or sites, or providing more information about particular products [Wikipedia 2008]. This new model of advertising is less intrusive, only displaying advertising information when the user makes the choice by clicking on an object in a video. Since it is the user who requests the product information, this type of advertising is better targeted and likely to be more effective. Moreover, by modeling and learning the user's visual attention patterns, the hot-spots that correspond to brands can be highlighted further so as to extract more attention from the user. As an initial attempt, a novel advertising system for online video service, called *VideoSense*, was developed to embed more contextually relevant ads in less intrusive positions within the video stream [Mei et al. 2007].

6.2. Collective Intelligence Gaming

Linking and commenting make blogging a kind of social collaboration, where individual bloggers are writing in a context created through a collective, but often unorganized, effort. In particular, vlogs create a comparatively vibrant collaboration and communication among vloggers in all places and time zones wherever interesting topics of discussion arise. Vlogs can facilitate practices that promote media literacy through the creation of collective documentaries [Hoem 2005]. Some vloggers work from assignments or prompts within online communities of similar vloggers such as MedicineFilms.com, for example. Assignment-based vlogging tends to be more collaborative, as every assignment-based vlog is a collaboration task between the assignment's creator and the video's creators. It should be noted that similar *collective intelligence* has also been widely experimented in Web 2.0, such as Wikipedia for collaborative writings and Yahoo! Answers for collaborative questions and answers.

Recently, collective intelligence has been explored in the form of online games, called *Games with a Purpose* [von Ahn 2006]. By playing these games, people contribute to their understanding of entities that make up the Web, and can even collectively solve large-scale computational problems. Four such games were developed at Carnegie Mellon University: *ESP Game*, *Phetch*, *Peekaboom*, and *Verbosity*. For example, *ESP Game* [von Ahn and Dabbish 2004] and *Phetch* [von Ahn et al. 2006] were designed to produce good labels of images on the Web. Similar games were also explored for music annotation at Columbia University [Mandel and Ellis 2007] and University of California at San Diego [Turnbull et al. 2007]. Google Image Labeler, originally

developed from the ESP Game, invites the public to improve its image search engine by working collaboratively to categorize online pictures by agreeing on specific, descriptive tags. Similarly, SFZero, an online role-playing game, describes itself as a "collaborative productive game," relying on its players to generate and to score virtually all of its missions [McGonigal 2007]. Many other large-scale open problems can also be solved using collective intelligence gaming in this unique way, such as monitoring of security cameras, and improving online video search [von Ahn 2006].

There is an increasing trend to harness the wisdom of crowds and the power of the collective. As a comparatively vibrant manner, vlogging may provide a better way to collective intelligence gaming. Some incentive applications and systems should be developed in the near future.

6.3. Other Applications

There is indeed a wide range of applications of vlogging in news, education, entertainment, and marketing. With the growth of mobile devices and video tools, for example, vlogging is likely to grow in popularity among faculty and students, who may use vlogs to record lectures and special events. Several e-learning sites such as www.videolectures.net are drawing an increasing number of students and teachers, allowing them to make online comments and social tagging. This has effectively made everyone a potential lifelong teacher as well as learner. In the future, the potential of vlogs can be exploited further in collaborative learning [Zahn and Finke 2003].

7. CONCLUDING REMARKS

In this article, we have presented a comprehensive survey of vlogging as a new technological trend, highlighted its current status, and foreseen emerging technological directions. We summarize the challenges for vlogging technology as four key issues that need to be answered (i.e., the basic supporting issue, value-added issue, incumbent issue, and incentive application issue). Along with their respective possibilities, we review the currently available techniques and tools that support vlogging, and present a new vision for future vlogging. Several multimedia techniques are introduced to make this vision possible. We also make an in-depth investigation of various vlog mining topics from a research perspective, and present several incentive applications such as user-targeted video advertising and collective intelligence gaming. We believe that with the focus being more on application-oriented, domain-specific vlogs, the field will experience a paradigm shift in the foreseeable future and generate considerable impact in day-to-day life.

Vlogging has initiated interest among different fields of study, such as multimedia computing, machine learning, and data mining, information retrieval, and human-computer interaction. However, the development of specialized video delivery, interaction, retrieval, content adaptation, and rights protection technologies aimed at the distinct features of the vlogosphere is still in its early stages. In the discussions in Sections 4 and 5, we listed some possible research problems that should be conducted in practice with some promising expectations for vlogging applications. From a long-term perspective, further progress should be made to provide vlogging with better scalability, interactivity, searchability and accessability, and to potentially reduce the legal, economic, and moral risks of vlogging applications. Looking to the future, vlogging and its incentive applications such as gaming and advertising will bring new opportunities and driving forces to the research in related fields.

ACKNOWLEDGMENTS

We wish to express our thanks to the anonymous reviewers who provided many helpful comments on the earlier version of this article. We also wish to thank our colleagues in JDL lab for providing the related materials.

REFERENCES

ABDOLLAHIAN, G. AND DELP, E. J. 2007. Analysis of unstructured video based on camera motion. In *Proceedings of SPIE-IS&T Electronic Imaging on Multimedia Content Access: Algorithms and Systems*. SPIE, vol. 6506.

ADAR, E., ZHANG, L., ADAMIC, L., AND LUKOSE, R. 2004. Implicit structure and the dynamics of blogspace. In *Online Proceedings of Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference*. http://www.blogpulse.com/papers/www2004adar.pdf.

ADOBE. 2009. Flash video learning guide. Tech. rep., Adobe Systems, Inc. http://www.adobe.com/devnet/ash/articles/video guide print.html.

ADOMAVICIUS, G. AND TUZHILIN, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Engin. 17*, 6, 734–749.

AGAMANOLIS, S. AND BOVE, V.M., JR. 1997. Multi-level scripting for responsive multimedia. *IEEE Multimedia 4*, 4, 40–50.

AHMAD, I., WEI, X. H., SUN, Y., AND ZHANG, Y.-Q. 2005. Video transcoding: An overview of various techniques and research issues. *IEEE Multimedia 7*, 5, 793–804.

ALLAN, J. AND CROFT, B. 2002. Challenges in information retrieval and language modeling: Report of a workshop held at the center for intelligent information retrieval. Tech. rep. http://www.sigir.org/forum/S2003/ir-challenges2.pdf.

AMEL, A. AND CRYAN, D. 2007. User-generated online video: Competitive review and market outlook. Tech. rep., ScreenDigest. http://www.screendigest.com.

AMITAY, E., CARMEL, D., HERSCOVICI, M., LEMPEL, R., AND SOFFER, A. 2004. Trend detection through temporal link analysis. *J. Am. Soc. Inform. Sci. Technol. 55*, 14, (Special issue on Webometrics) 1279–1281.

ANDERSON, P. 2007. What is web 2.0? Ideas, technologies and implications for education. Tech. rep. TSW0701, JISC Technology and Standards Watch (TechWatch). http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf.

ANGELIDES, M. C. 2003. Multimedia content modeling and personalization. *IEEE Multimedia 10*, 4, 12–15.

ANJEWIERDEN, A., DE HOOG, R., BRUSSEE, R., AND EFIMOVA, L. 2005. Detecting knowledge flows in weblogs. In *Proceedings of the 13th International Conference on Conceptual Structures (ICCS 2005)*, Kassel University Press, 1–12.

ATTARDI, G. AND SIMI, M. 2006. Blog mining through opinionated words. In *Proceedings of the 15th Text Retrieval Conference (TREC 2006)*. SP 500-272. NIST. http://trec.nist.gov/pubs/trec15/papers/upisa.blog.¯nal.pdf.

AUBERT, O. AND PRIÉ, Y. 2005. Advene: Active reading through hypervideo. In *Proceedings of the 16th ACM Conference on Hypertext and Hypermedia*. ACM, New York.

BABAGUCHI, N., KAWAI, Y., AND KITAHASHI, T. 2002. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Multimedia 4*, 1, 68–75.

BENITEZ, A. B., BEIGI, M., AND CHANG, S.-F. 1998. Using relevance feedback in content-based image metasearch. *IEEE Internet Comput. 2*, 4, 59–69.

BERBERICH, K., VAZIRGIANNIS, M., AND WEIKUM, G. 2005. Time-aware authority ranking. *Internet Math. J. 2*, 3, 309–340.

BESACIER, L., QUÉNOT, G., AYACHE, S., AND MORARU, D. 2004. Video story segmentation with multi-modal features: Experiments on trecvid 2003. In *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*. ACM, New York.

BEST, D. 2006. Web 2.0: Next big thing or next big internet bubble? In *Lecture Web Information Systems*, Technische Universiteit Eindhoven.

BLINKX. 2007. Blinkx video seo white paper. Tech. rep., Blinkx. http://www.blinkx.com/seo.pdf.

BLOOD, R. 2002. *The Weblog Handbook: Practical Advice on Creating and Maintaining Your Blog*. Perseus Publishing, Cambridge, MA.

BOCCIGNONE, G., CHIANESE, A., MOSCATO, V., AND PICARIELLO, A. 2005. Foveated shot detection for video segmentation. *IEEE Trans. Circuits Syst.Video Technol.15*, 3, 365–377.

BOLL, S. 2007. Multi tube where multimedia and web 2.0 could meet. *IEEE Multimedia 14*, 1, 9-13.

BORDWELL, D. AND THOMPSON, K. 2006. *Film Art: An Introduction*. McGraw-Hill, New York.

BORTH, D., ULGES, A., SCHULZE, C., AND BREUEL, T. M. 2008. Keyframe extraction for video tagging & summarization. *Informatiktage*, 45–48.

BOSCH, A., MUOZ, X., AND MARTI, R. 2007. A review: Which is the best way to organize/classify images by content? *Image Vision Comput. 25*, 6, 778–791.

BROWNE, P., CZIRJEK, C., GAUGHAN, G., GURRIN, C., JONES, G., LEE, H., MCDONALD, K., MURPHY, N., O'CONNOR, N., O'HARE, N., SMEATON, A., AND YE, J. 2003. Dublin City University video track experiments for trec 2003. In *Online Proceedings of TRECVID Workshop*, NIST. http://wwwnlpir.nist.gov/projects/tvpubs/tvpapers03/dublin.lee.paper.pdf.

BURNETT, I., DE WALLE, R. V., HILL, K., BORMANS, J., AND PEREIRA, F. 2003. Mpeg-21: Goals and achievements. *IEEE Multimedia 10*, 4, 60–70.

CASTRO, M., DRUSCHEL, P., KERMARREC, A.-M., NANDI, A., ROWSTRON, A., AND SINGH, A. 2003. Splitstream: High-bandwidth multicast in cooperative environments. In *Proceedings of the 19th ACM Symposium on Operating Systems Principles*. ACM, New York, 298–313.

CERNEKOVA, Z., KOTROPOULOS, C., AND PITAS, I. 2003. Video shot segmentation using singular value decomposition. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. Vol. 2, 301–304.

CHAISORN, L., CHUA, T.-S., KOH, C.-K., ZHAO, Y., XU, H., FENG, H., AND TIAN, Q. 2003. A two-level multimodal approach for story segmentation of large news video corpus. In *Online Proceedings of TRECVID Workshop*, NIST. http://www-nlpir.nist.gov/projects/tvpubs/tvpapers03/nus.¯nal.paper.pdf.

CHANG, C.-H. AND TSAI, K.-C. 2007. Aspect summarization from blogsphere for social study. In *Proceedings of 7th IEEE International Conference on Data Mining Workshops (ICDMW 2007)*. 9–14.

CHANG, P., HAN, M., AND GONG, Y. H. 2002. Extract highlights from baseball game video with hidden Markov models. In *Proceedings of the IEEE Conference on Image Processing*. 609–612.

CHANG, S.-F., MA, W.-Y., AND SMEULDERS, A. 2007. Recent advances and challenges of semantic image/video search. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IV-1205-IV-1208.

CHANG, S.-F. AND VETRO, A. 2005. Video adaptation: Concepts, technologies, and open issues. *Proc. IEEE 93*, 1,148–158.

CHAU, M. AND XU, J. 2007. Mining communities and their relationships in blogs: A study of online hate groups. *Int. J. Human-Comput. Stud. 65*, 57–70.

CHEN, B., ZHAO, Q. K., SUN, B. J., AND MITRA, P. 2007. Predicting blogging behavior using temporal and social networks. In *Proceedings of the 7th IEEE International Conference on Data Mining*. 439–441.

CHEN, T.-W., HUANG, Y.-W., CHEN, T.-C., CHEN, Y.-H., TSAI, C.-Y., AND CHEN, L.-G. 2005. Architecture design of h.264/avc decoder with hybrid task pipelining for high definition videos. In *Proceedings of the IEEE International Symposium on Circuits and Systems*. Vol. 3, 2931–2934

CHI, Y., TSENG, B. L., AND TATEMURA, J. 2006. Eigen-trend: Trend analysis in the blogosphere based on singular value decompositions. In *Proceedings of the 15th ACM Conference on Information and Knowledge Management (CIKM)*. ACM, New York, 68–77.

CHO, J., KWON, K., AND PARK, Y. 2007. Collaborative filtering using dual information sources. *IEEE Intell. Syst. 22*, 3, 30–38.

CHU, Y., RAO, S. G., AND ZHANG, H. 2000. A case for end system multicast. In *Proceedingsof the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*. ACM, New York, 1–12.

CHUA, T., NEO, S., LI, K., WANG, G., SHI, R., ZHAO, M., AND XU, H. 2004. Trecvid 2004 searchand feature extraction task by Nus Pris. In *Proceedings of TRECVID Workshop*, NIST. http://www-nlpir.nist.gov/projects/tvpubs/tvpapers04/nus.pdf.

CONKLIN, G. J., GREENBAUM, G. S., LIPPMAN, K. O. L., AND REZNIK, Y. A. 2001. Video coding for streaming media delivery on the internet. *IEEE Trans. Circuits Syst. Video Technol. 11*, 3, 269–281.

COOPER, M., LIU, T., AND RIEFFEL, E. 2007. Video segmentation via temporal pattern classification. *IEEE Trans. Multimedia 9*, 3, 610–618.

COTSACES, C., NIKOLAIDIS, N., AND PITAS, I. 2006. Video shot detection and condensed representation: A review. *IEEE Signal Process. Mag. 23*, 2, 28–37.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv. 40*, 2, 1–66.

DEERING, S. AND CHERITON, D. 1990. Multicast routing in datagram internetworks and extended lans. *ACM Trans. Comput. Syst. 8*, 2, 85–110.

DIEPOLD, K. AND MORITZ, S. 2004. *Understanding MPEG-4, Technology and Business Insights*. Focal Press.

DIMITROVA, N., ZHANG, H.-J., SHAHRARAY, B., SEZAN, I., HUANG, T. S., AND ZAKHOR, A. 2002. Applications of video-content analysis and retrieval. *IEEE Multimedia 9*, 3, 42–55.

DOËRR, G. AND DUGELAY, J.-L. 2003. A guide tour of video watermarking. *Signal Process*: *Image Commun. 18*, 4, 263–282.

DUAN, L.-Y., XU, M., TIAN, Q., XU, C.-S., AND JIN, J. S. 2005. A unified framework for semantic shot classification in sports video. *IEEE Trans. Multimedia 7*, 6, 1066–1083.

EI-ALFY, H., JACOBS, D., AND DAVIS, L. 2007. Multi-scale video cropping. In *Proceedings of the ACM International Conference on Multimedia*. ACM, New York, 97–106.

EUGSTER, P., GUERRAOUI, R., KERMARREC, A.-M., AND MASSOULIE, L. 2004. From epidemics to distributed computing. *IEEE Computer 37*, 5, 60–67.

FAN, J. P., LUO, H. Z., ZHOU, A. Y., AND KEIM, D. A. 2008. Personalized news video recommendation via interactive exploration. In *Proceedings of the 4th International Symposium Advances in Visual Computing*. Lecture Notes in Computer Science, vol. 5359. Springer, Berlin, 380–389.

FAN, J. P., ZHU, X., AND LIN, X. 2002. Mining of video database. In *Multimedia Data Mining*. Kluwer, Amsterdam.

FAN, X., XIE, X., ZHOU, H.-Q., AND MA, W.-Y. 2003. Looking into video frames on small displays. In *Proceedings of the ACM International Conference on Multimedia*. ACM, New York, 247–250.

FINKE, M. 2004. Mobile interactive video (d13): 2nd year public report. Tech. rep. IST-2001-37365, MUMMY Project. http://mummy.intranet.gr/includes/docs/MUMMY-D13y2-ZGDV-MobVideo-v02.pdf.

FINKELSTEIN, L., GABRILOVICH, E., MATIAS, Y., RIVLIN, E., SOLAN, Z., WOLFMAN, G., AND RUPPIN, E. 2002. Placing search in context: The concept revisited. *ACM Trans. Inform. Syst. 20*, 1, 116–131.

FONSECA, P. M. AND PEREIRA, F. 2004. Automatic video summarization based on mpeg-7 descriptions. *Signal Process*: *Image Comm. 19*, 685–699.

FUJIMURA, K., INOUE, T., AND SUGISAKI, M. 2005. The Eigen rumor algorithm for ranking blogs. In *Online Proceedings of WWW 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

FURUKAWA, T., ISHIZUKA, M., MATSUO, Y., OHMUKAI, I., AND UCHIYAMA, K. 2007. Analyzing reading behavior by blog mining. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*. 1353–1358.

GELGON, M. AND HAMMOUD, R. I. 2006. Building object-based hyperlinks in videos: Theory and experiments. In *Interactive Video*: *Algorithms and Technologies*. Springer, Berlin.

GIBBON, D. C. 2006. Introduction to video search engines. In *Tutorials at the 15th International World Wide Web Conference*. http://public.research.att.com/~chen/www2006/tutorials/Wed-PM-IntroVideoSearch.pdf.

GILL, K. E. 2004. How can we measure the in˚uence of the blogosphere? In *Online Proceedings of Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference*.

GIRGENSOHN, A. AND BORECZKY, J. 2000. Time-constrained keyframe selection technique. *Multimedia Tools Appl. 11*, 3, 347–358.

GIRGENSOHN, A., SHIPMAN, F., AND WILCOX, L. 2003. Hyper-Hitchcock: Authoring interactive videos and generating interactive summaries. In *Proceedings of the 11th ACM International Conference on Multimedia*. ACM, New York, 92–93.

GLANCE, N., HURST, M., AND TOMOKIYO, T. 2004. Blogpulse: Automated trend discovery for weblogs. In *Proceedings of the Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference*.

GOLDENBERG, J., LIBAI, B., AND MULLER, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Lett. 12*, 3, 211–223.

GRANOVETTER, M. 1978. Threshold models of collective behavior. *Am. J. Sociology 83*, 6, 1420–1438.

GRUHL, D., GUHA, R., LIBEN-NOWELL, D., AND TOMKINS, A. 2004. Information diffusion through blogspace. In *Proceedings of the 13th International World Wide Web Conference*. 491–501.

HAERING, N., QIAN, R. J., AND SEZAN, M. I. 2000. A semantic event detection approach and its application to detecting hunts in wildlife video. *IEEE Trans. Circuits Syst. Video Technol. 10*, 6, 857–868.

HAMMOUD, R. AND MOHR, R. 2000. A probabilistic framework of selecting effective key frames for video browsing and indexing. In *Proceedings of International Workshop on Real-Time Image Sequence Analysis*. 79–88.

HAMMOUD, R. I. 2006. Introduction to interactive video. In *Interactive Video: Algorithms and Technologies*, Springer, Berlin.

HANJALIC, A. 2002. Shot-boundary detection: Unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology 12,* 2, 90–105.

HAUPTMANN, A. G. AND CHRISTEL, M. G. 2004. Successful approaches in the trec video retrieval evaluations. In *Proceedings of the 12th ACM International Conference on Multimedia*. ACM, New York, 668–675.

HEI, X., LIANG, C., LIANG, J., LIU, Y., AND ROSS, K. W. 2006. Insights into pplive: A measurement study of a large-scale p2p iptv system. In *Proceedings of Workshop Internet Protocol TV (IPTV) Services Over World Wide Web in Conjunction (WWW2006)*.

HILLEY, D. AND RAMACHANDRAN, U. 2007. Stampedert: Distributed programming abstractions for live streaming applications. In *Proceedings of the 27th International Conference on Distributed Computing Systems (ICDCS'07)*. 65–74.

HJELSVOLD, R., VDAYGIRI, S., AND LÉAUTÉ, Y. 2001. Web-based personalization and management of interactive video. In *Proceedings of the 10th International World Wide Web Confeence*. ACM, New York, 129–139.

HOASHI, K., SUGANO, M., NAITO, M., MATSUMOTO, K., SUGAYA, F., AND NAKAJIMA, Y. 2004. Shot boundary determination on mpeg compressed domain and story segmentation experiments for trecvid. In *Online Proceedings of TRECVID Workshops*, NIST.
http://wwwnlpir.nist.gov/projects/tvpubs/tvpapers04/kddi.pdf.

HOEM, J. 2005. Videoblogs as "collective documentary". In *Proceedings of Blogtalk 2.0*: *The European Conference on Weblog*. 237–270.

HOTVIDEO. 2008. New initiatives—Hotvideo: The cool way to link. IBM Res.
News. http://www.research.ibm.com/topics/popups/innovate/multimedia/ html/hotvideo.html.

HSU, W., KENNEDY, L., HUANG, C.-W., CHANG, S.-F., LIN, C.-Y., AND IYENGAR, G. 2004. News video story segmentation using fusion of multi-level multi-modal features in trecvid 2003. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*. Vol. 3, 645–648.

HU, W. M., WU, O., CHEN, Z. Y., FU, Z. Y., AND STEPHEN, J. 2007. Maybank: Recognition of pornographic web pages by classifying texts and images. *IEEE Trans. Pattern Anal. Mach. Intell. 29*, 6, 1019–1034.

HUA, X.-S., LU, L., AND ZHANG, H.-J. 2004. Optimization-based automated home video editing system. *IEEE Trans. Circuits Syst. Video Technol. 14*, 5, 572–583.

HUANG, G. 2007. PPLive: A practical P2P live system with huge amount of users. In *Proceedings of the ACM SIGCOMM Workshop on Peer-to-Peer Streaming and IPTV Workshop*.

HUANG, J. C., LIU, Z., AND WANG, Y. 2005. Joint scene classification and segmentation based on a hidden Markov model. *IEEE Trans. Multimedia 7*, 3, 538–550.

HUET, B. AND MERI, B. 2006. Automatic video summarization. In *Interactive Video: Algorithms and Technologies*, Springer, Berlin.

HUO, L. S. 2006. Robust and adaptive media streaming over the internet. Ph.D. dissertation, Institute of Computing Technology, Chinese Academy of Sciences, Beijing.

IAB. 2008. User generated content, social media, and advertising: An overview. Tech. rep., Interactive Advertising Bureau (IAB).

IRMAK, U., MIHAYLOV, S., SUEL, T., GANGULY, S., AND IZMAILOV, R. 2006. Efficient query subscription processing for prospective search engines. In *Proceedings of the Annual Technical Conference USENIX'06*.

JAKOB, V. 2007. Tagging, folksonomy & co - renaissance of manual indexing? In *Proceedings of the International Symposium of Information Science*. 234–254.

JENSEN, J. F. 2007. User generated content: A mega-trend in the new media landscape. In *TICSP Adjunct Proceedings of the 5th European Interactive TV Conference (EuroITV 2007)*.

JOSHI, A., FININ, T., JAVA, A., KALE, A., AND KOLARI, P. 2007. Web 2.0 mining: Analyzing social media. In *Proceedings of the NSF Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation*.

KALE, A. 2007. Modeling trust and influence on blogosphere using link polarity. M.S. thesis, Department of Computer Science and Electrical Engineering, University of Maryland.

KELLEY, M. 2008. Web 2.0 mashups and niche aggregators. Tech. rep., O'Reilly Media, Inc. http://www.oreilly.com/catalog/9780596514006/.

KENNEDY, L., CHANG, S.-F., AND NATSEV, A. 2008. Query-adaptive fusion for multimodal search. *Proc. IEEE 96*, 4, 567–588.

KING, A. 2003. Vlogging: Video weblogs. http://www.webreference.com/new/030306. html#toc1.

KITAYAMA, D. AND SUMIYA, K. 2006. A blog search method using news video scene order. In *Proceedings of the 12th International Multimedia Modeling Conference*. 446–449.

KOLARI, P., FININ, T., AND JOSHI, A. 2006. Svms for the blogosphere: Blog identification and splog detection. In *Proceedings of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*.

KRAFT, R., CHANG, C. C., MAGHOUL, F., AND KUMAR, R. 2006. Searching with context. In *Proceedings of the 15th International World Wide Web Conference*. ACM, New York, 477–486.

KRAFT, R., MAGHOUL, F., AND CHANG, C. C. 2005. Y!q: Contextual search at the point of inspiration. In *Proceedings of the 14th ACM International Conference on Information Knowledge Management*. ACM, New York, 816–823.

KRISHNAMACHARI, S., YAMADA, A., ABDEL-MOTTALEB, M., AND KASUTANI, E. 2000. Multimedia content filtering, browsing, and matching using mpeg-7 compact color descriptors. In *Proceedings of the 4th International Conference on Advances in Visual Information Systems*. Lecture Notes in Computer Science, vol. 1929, Springer, Berlin, 200–211.

KUMAR, R., NOVAK, J., RAGHAVAN, P., AND TOMKINS, A. 2003. On the bursty evolution of blogspace. In *Proceedings of the 12th International World Wide Web Conference*. ACM, New York, 568–576.

KUMAR, R., NOVAK, J., RAGHAVAN, P., AND TOMKINS, A. 2004. Structure and evolution of blogosphere. *Comm. ACM 47*, 12, 35–39.

LASER. 2007. ISO/IEC JTC 1/SC 29/WG 11 WD3.0 of ISO/IEC 14496-20 2nd Ed., information technology coding of audio-visual objects part 20: Lightweight Application Scene Representation (LASeR) and Simple Aggregation Format (SAF).

LAWTON, G. 2000. Video streams into the mainstream. *Computer 33*, 7, 12–17.

LEE, S. AND YOO, C. D. 2006. Video fingerprinting based on centroids of gradient orientations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Vol. 2, II-401-II-404.

LEE, S.-H., YEH, C.-H., AND KUO, C.-C. J. 2005. Home-video content analysis for MTV-style video generation. In *Proceedings of the SPIE*. Vol. 5682. SPIE, 296–307.

LELESCU, D. AND SCHONFELD, D. 2003. Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream. *IEEE Trans. Multimedia 5*, 1, 106–117.

LEW, M. S., SEBE, N., DJERABA, C., AND JAIN, R. 2006. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl. 2*, 1, 1–19.

LI, B. AND YIN, H. 2007. Peer-to-peer live video streaming on the internet: Issues, existing approaches, and challenges. *IEEE Commun. Mag. 45*, 1, 94–99.

LI, B. X., PAN, H., AND SEZAN, I. 2003. Comparison of video shot boundary detection techniques. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 169–172.

LI, H. L. AND NGAN, K. N. 2007. Automatic video segmentation and tracking for content-based applications. *IEEE Commun. Mag. 45*, 1, 27–33.

LI, W. 2001. Overview of ¯ne granularity scalability in mpeg-4 video standard. *IEEE Trans. Circuits Syst. Video Technol. 11*, 3, 301–317.

LI, Y., ZHANG, T., AND TRETTER, D. 2001. An overview of video abstraction techniques. Tech. rep. HPL-2001-191, HP Labs. http://www.hpl.hp.com/techreports/2001/HPL2001191.pdf.

LI, Z. W., WANG, B., LI, M. J., AND MA, W.-Y. 2005. A probabilistic model for retrospective news event detection. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 106–113.

LIE, W.-N. AND LAI, C.-M. 2004. News video summarization based on spatial and motion feature analysis. In *Proceedings of Pacific Rim Conference on Multimedia*. Lecture Notes in Computer Science, vol. 3332. Springer, Berlin, 246–255.

LIENHART, R. 2001. Reliable dissolve detection. In *Proceedings of SPIE Storage and Retrievalfor Media Databases*, vol. 4315, 219–230.

LIN, C.-W., YOUN, J., ZHOU, J., SUN, M.-T., AND SODAGAR, I. 2001. Mpeg video streaming with VCR functionality. *IEEE Trans. Circ. Syst. Video Technol. 11*, 3, 415–425.

LIN, E. T., ESKICIOGLU, A. M., LAGENDIJK, R. L., AND DELP, E. J. 2005. Advances in digital video content protection. *Proc. IEEE 93*, 1, 171–183.

LIN, T. AND ZHANG, H. 2001. Video content representation for shot retrieval and scene extraction. *Int. J. Image Graph. 3*, 1, 507–526.

LIN, Y.-R., SUNDARAM, H., CHI, Y., TATEMURA, J., AND TSENG, B. 2007. Splog detection using content, time and link structures. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. 2030–2033.

LIU, F. AND GLEICHER, M. 2006. Video retargeting: Automating pan and scan. In *Proceedings of the ACM International Conference on Multimedia*. ACM, New York, 241–250.

LIU, J., LI, B., AND ZHANG, Y.-Q. 2003. Adaptive video multicast over the internet. *IEEE Multimedia 10*, 1, 22–31.

LIU, J. C., RAO, S., LI, B., AND ZHANG, H. 2008. Opportunities and challenges of peer-to-peer internet video broadcast. *Proc. IEEE 96*, 1, 11–24.

LIU, L. AND FAN, G. 2005. Combined keyframe extraction and object-based video segmentation. *IEEE Trans. Circuits Syst.Video Technol. 15*, 7, 869–884.

LIU, X. AND ESKICIOGLU, A. M. 2003. Selective encryption of multimedia content in distribution networks: Challenges and new directions. In *Proceedings of the 2nd International Conference on Communications, Internet, and Information Technology*, 527–533.

LIU, Z. AND WANG, Y. 2007. Major cast detection in video using both speaker and face information. *IEEE Trans. Multimedia 9*, 1, 89–101.

LOCHER, T., MEIER, R., SCHMID, S., AND WATTENHOFER, R. 2007. Push-to-pull peer-to-peer live streaming. In *Proceedings of the 21st International Symposium on Distributed Computing (DISC)*. Lecture Notes in Computer Science, vol. 4731, Springer, Berlin, 388–402.

LOWE, D. 2004. Distinctive image features from scale invariant keypoints. *Int. J. Comput. Vision 60*, 2, 91–110.

LU, S. 2004. Content analysis and summarization for video documents. M.S. thesis, Department of Computer Science and Engineering, The Chinese University of Hong Kong.

LU, Z. K., LIN, W., LI, Z. G., LIM, K. P., LIN, X., RAHARDJA, S., ONG, E. P., AND YAO, S. 2005. Perceptual region-of-interest (ROI) based scalable video coding. Tech. rep. Doc. JVT-O056, Joint Video Team, Busan, KR.

LYNCH, K. 2007. Video on the web. In *Online Proceedings of W3C Workshop on Video on the Web*. http://www.w3.org/2007/08/video/positions/adobe.pdf.

MA, Y.-F., HUA, X.-S., LU, L., AND ZHANG, H.-J. 2005. A generic framework of user attention model and its application in video summarization. *IEEE Trans. Multimedia 7*, 5, 907–919.

MAHMOOD, T. S. AND PONCELEON, D. 2001. Learning video browsing behavior and its application in the generation of video previews. In *Proceedings of the 9th ACM International Conference on Multimedia*. ACM, New York, 119–128.

MANDEL, M. AND ELLIS, D. P. W. 2007. A web-based game for collecting music metadata. In *Proceedings of the 8th International Conference on Music Information Retrieval*. 365–366.

MARTINSSON, E. 2006. IPTV the future of television? Report in computer communication and distributed systems. Tech. rep. EDA390, Department of Computer Science and Engineering, Chalmers University of Technology.

MATSUO, Y., SHIRAHAMA, K., AND UEHARA, K. 2003. Video data mining: Extracting cinematic rules from movie. In *Proceedings of the International Workshop on Multimedia Data Management (MDM/KDD 2003)*. 18–27.

MCDONALD, D. W. 2003. Ubiquitous recommendation systems. *Computer 36*, 10, 111–112.

MCGLOHON, M., LESKOVEC, J., FALOUTSOS, C., HURST, M., AND GLANCE, N. 2007. Finding patterns in blog shapes and blog evolution. In *Proceedings of the 1st International Conference on Weblogs and Social Media*. http://www.icwsm.org/papers/2{McGlohon-Leskovec-Faloutsos-Hurst-Glance.pdf.

MCGONIGAL, J. 2007. Why I love bees: A case study in collective intelligence gaming. In *The Ecology of Games: Connecting Youth, Games, and Learning*, 199–227.

MEFEEDIA. 2007. State of the vlogosphere. http://mefeedia.com/blog/2007/03/29/state-of-the-vlogosphere-march-2007/.

MEI, T., HUA, X.-S., YANG, L. J., AND LI, S. P. 2007. Videosense {towards effective online video advertising. In *Proceedings of the ACM International Conference on Multimedia*. ACM, New York, 1075–1084.

MEISEL, J. B. 2008. Entry into the market for online distribution of digital content: Economic and legal ramifications. *SCRIPTed 5*, 1.

MERRILL, D. 2006. Mashups: The new breed of web app. IBM DeveloperWorks library. http://www.ibm.com/developerworks/xml/library/x-mashups.html.

MISHNE, G., CARMEL, D., AND LEMPEL, R. 2005. Blocking blog spam with language model disagreement. In *Online Proceedings of the WWW 1st International Workshop on Adversarial Information Retrieval on the Web*. http://airweb.cse.lehigh.edu/2005/ mishne.pdf.

MISHNE, G. AND DE RIJKE, M. 2006. A study of blog search. In *Proceedings of the 28th European Conference on IR Research (ECIR 2006)*. Lecture Notes in Computer Science, vol. 3936. Springer, Berlin, 289–301.

MOORE, C. AND NEWMAN, M. E. J. 2000. Epidemics and percolation in small-world networks. *Physical Rev. E 61*, 5678–5682.

MPEG. 2007. ISO/IEC JTC1/SC29/WG11 (MPEG). Call for proposals on image and video signature tools. Lausanne, Switzerland.

NAKAJIMA, S., TATEMURA, J., AND HINO, Y. 2005. Discovering important bloggers based on a blog thread analysis. In *Online Proceedings of WWW 2nd Annual Workshop on the We-blogging Ecosystem: Aggregation, Analysis and Dynamics*. http://www-idl.hpl.hp.com/blog-workshop2005/nakajima.pdf.

NAPHADE, M. R., BASU, S., SMITH, J. R., LIN, C.-Y., AND TSENG, B. 2002. Modeling semantic concepts to support query by keywords in video. In *Proceedings of the IEEE International Conference on Image Processing*. 145–148.

NARDI, B. A., SCHIANO, D. J., GUMBRECHT, M., AND SWARTZ, L. 2004. The blogosphere: Why we blog. *Commun. ACM 47*, 12, 41–46.

NEWMAN, M. 2003. The structure and function of complex networks. *SIAM Rev. 45*, 2, 167–256.

NGO, C.-W., MA, Y.-F., AND ZHANG, H.-J. 2005. Video summarization and scene detection by graph modeling. *IEEE Trans. Circuits Syst. Video Technol. 15*, 2, 296–305.

O'CONNOR, N., MARLOW, S., MURPHY, N., WOLENETZ, M., ESSA, I., HUTTO, P., STARNER, T., AND RAMACHANDRAN, U. 2001. Fischlar: an on-line system for indexing and browsing broadcast television content. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. Vol. 3, 1633–1636.

O'HARE, N., SMEATON, A., CZIRJEK, C., O'CONNOR, N., AND MURPHY, N. 2004. A generic news story segmentation system and its evaluation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 3, 1028–1031.

OHM, J.-R. 2005. Advances in scalable video coding. *Proc. IEEE 93*, 1, 42–56.

O'REILLY, T. 2005. What is web 2.0? Design patterns and business models for the next generation of software. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

ORRIOLS, X. AND BINEFAA, X. 2003. Online Bayesian video summarization and linking. In *Proceedings of the International Conference on Image and Video Retrieval*. Lecture Notes Computer Science, vol. 2383, Springer, Berlin, 338–348.

OWENS, G. AND ANDJELIC, A. 2007. Online video roadmap. Tech. rep., Avenue A*j*Razorfish.

PADMANABHAN, V. N., WANG, H. J., CHOU, P. A., AND SRIPANIDKULCHAI, K. 2002. Distributing streaming media content using cooperative networking. In *Proceedings of the 12th International Workshop on Network and Operating Systems Support for Digital Audio and Video*. ACM, New York, 177–186.

PAN, J.-Y. AND FALOUTSOS, C. 2002. Videocube: A novel tool for video mining and classification. In *Proceedings of the International Conference on Asian Digital Libraries*. 194–205.

PAN, J.-Y., YANG, H., AND FALOUTSOS, C. 2004. MMSS: Multi-modal story-oriented video summarization. In *Proceedings of the 4th IEEE International Conference on Data Mining*. 491–494.

PARKER, C. AND PFEIFFER, S. 2005. Video blogging: Content to the max. *IEEE Multimedia 12*, 2, 4–8.

PEETS, L. AND PATCHEN, J. 2007. Viacom v YouTube. *IP Newsl. 30*, 4.

PEREIRA, F., VAN BEEK, P., KOT, A. C., AND OSTERMANN, J. EDS. 2005. special issue on analysis and understanding for video adaptation. *IEEE Trans. Circuits Syst. Video Technol. 15*, 10, 1197–1199.

PFEIFFER, S., PARKER, C., AND PANG, A. 2005. The continuous media web: A distributed multimedia information retrieval architecture extending the world wide web. *Multimedia Syst. J. 10*, 6, 544–558.

PFEIFFER, S., SCHREMMER, C., AND PARKER, C. 2003. Annodex: A simple architecture to enable hyperlinking, search and retrieval of time-continuous data on the web. In *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*. ACM, New York, 87–93.

PIEPER, J., SRINIVASAN, S., AND DOM, B. 2001. Streaming-media knowledge discovery. *Computer 34*, 9, 68–74.

POLLONE, M. 2001. Hyperfilm: Video hypermedia production. Tech. rep., Hyperfilm project. http://www.hyperfilm.it/.

QI, G.-J., HUA, X.-S., RUI, Y., TANG, J. H., MEI, T., AND ZHANG, H.-J. 2007. Correlative multi-label video annotation. In *Proceedings of the ACM International Conference on Multimedia*. ACM, New York, 17–26.

REINHARDT, R. 2008. Delivering a reliable flash video experience. In *Adobe Flash CS3 Professional Video Studio Techniques*, Pearson Education, Inc. (Chapter 12).

RHIND, A., BAMFORD, M., AND GRIFFIN, J. 2005. Blogging. Blogging. In *MC Insight*. http://salsadigital.typepad.com/salsadigital/files/mcinsight july.pdf.

ROSENBLOOM, A. 2004. The blogosphere. *Comm. ACM 47*, 12, 31–33.

ROWLEY, H. A., JING, Y. S., AND BALUJA, S. 2006. Large scale image-based adult-content filtering. In *Proceedings of the 1st International Conference on Computer Vision Theory and Applications*. 290–296.

RUI, Y., GUPTA, A., AND ACERO, A. 2000. Automatically extracting highlights for TV basketball programs. In *Proceedings of the 8th ACM International Conference on Multimedia*. ACM, New York, 105–115.

RUI, Y., HUANG, T. S., AND MEHROTRA, S. 1999. Constructing table-of-content for videos. *Multimedia Syst. 7*, 5, 359–368.

SAWHNEY, N., BALCOM, D., AND SMITH, I. 1996. Hypercafe: Narrative and aesthetic properties of hypervideo. In *Proceedings of the ACM International Conference on Hypertext*. 1–10.

SCHMITZ, P., SHAFTON, P., SHAW, R., TRIPODI, S., WILLIAMS, B., AND YANG, J. 2006. Remix: Video editing for the web. In *Proceedings of the 14th ACM International Conference on Multimedia*. ACM, New York, 797–798.

SCHWARZ, H., MARPE, D., AND WIEGAND, T. 2007. Overview of the scalable video coding extension of the h.264/avc standard. *IEEE Trans. Circuits Syst. Video Technol. 17*, 9, 1103–1120.

SCOVANNER, P., ALI, S., AND SHAH, M. 2007. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM International Conference on Multimedia*. ACM, New York, 357–360.

SERRANO, M. 2007. Programming web multimedia applications with hop. In *Proceedings of the 15th ACM International Conference on Multimedia*. ACM, New York, 1001–1004.

SHEN, D., SUN, J.-T., YANG, Q., AND CHEN, Z. 2006. Latent friend mining from blog. In *Proceedings of the 5th International Conference on Data Mining*. 552–561.

SMIL. 2001. Synchronized Multimedia Integration Language (SMIL 2.0). World Wide Web Consortium (W3C). http://www.w3.org/TR/smil20/.

SMITH, J. AND STOTTS, D. 2002. An extensible object tracking architecture for hyperlinking in real-time and stored video streams. Tech. rep. TR02-017, Department of Computer Science, University of North Caroline at Chapel Hill.

SNOEK, C. G. M. AND WORRING, M. 2005. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools Appl. 25*, 1, 5–35.

SONG, Y. AND MARCHIONINI, G. 2007. Effects of audio and visual surrogates for making sense of digital video. In *Proceedings of the ACM International Conference on Computer-Human Interaction*. ACM, New York, 867–876.

SPOERRI, A. 2007. Visual mashup of text and media search results. In *Proceedings of the 11th International Conference on Information Visualization*. 216–221.

SUNDARAM, H. AND CHANG, S.-F. 2003. Video analysis and summarization at structural and semantic levels. In *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*, Springer, Berlin.

TIAN, Y. H., HUANG, T. J., AND GAO, W. 2003. The influence model of online social interactions and its learning algorithms. *Chinese J. Computers 28*, 7, 848–858.

TOLVA, J. 2006. Medialoom: An interactive authoring tool for hypervideo. Project report. http://www.mindspring.com/"jntolva/medialoom/paper.html.

TSAI, F. S., CHEN, Y., AND CHAN, K. L. 2007. Probabilistic techniques for corporate blog mining. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining Workshops*. Lecture Notes on Artificial Intelligence, vol. 4819. Springer, Berlin, 35–44.

TURNBULL, D., LIU, R., BARRINGTON, L., AND LANCKRIET, G. 2007. A game-based approach for collecting semantic annotations of music. In *Proceedings of the 8th International Conference on Music Information Retrieval*. 535–538.

TUULOS, V., SCHEIBLE, J., AND NYHOLM, H. 2007. Combining web, mobile phones and public displays in large-scale: Manhattan story mashup. In *Proceedings of the 5th International Conference on Pervasive Computing*. Ledcture Notes in Computer Science, vol. 4480, Springer, Berlin, 37–54.

VERNA, P. 2007. User-generated content: Will web 2.0 pay its way? Tech. rep., eMarketer. http://www.emarketer.com/.

VETRO, A., CHRISTOPOULOS, C., AND SUN, H. 2003. Video transcoding architectures and techniques: An overview. *IEEE Signal Process. Mag. 20*, 2, 18–29.

VIDEOCLIX. 2008. Videoclix, where entertainment meets commerce. http://www.videoclix.com/docs/Brochure.pdf.

VON AHN, L. 2006. Games with a purpose. *Computer 39*, 6, 96–98.

VON AHN, L. AND DABBISH, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 319–326.

VON AHN, L., GINOSAR, S., KEDIA, M., LIU, R., AND BLUM, M. 2006. Improving accessibility of the web with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 79–82.

WALES, C., KIM, S., LEUENBERGER, D., WATTS, W., AND WEINOTH, O. 2005. IPTV - The revolution is here. http://www.eecs.berkeley.edu/~binetude/course/eng298a_2/IPTV.pdf.

WEN, J.-R., LAO, N., AND MA, W.-Y. 2004. Probabilistic model for contextual retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, 57–63.

WETZKER, R., ALPCAN, T., BAUCKHAGE, C., UMBRATH, W., AND ALBAYRAK, S. 2007. B2Rank: An algorithm for ranking blogs based on behavioral features. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. 104–107.

WIEGAND, T., SULLIVAN, G. J., BJNTEGAARD, G., AND LUTHRA, A. 2003. Overview of the h.264/avc video coding standard. *IEEE Trans. Circuits Syst. Video Technology 13*, 7, 560–576.

WIKIPEDIA. 2008. http://en.wikipedia.org/wiki/.

WOLF, L., GUTTMANN, M., AND COHEN-OR, D. 2007. Non-homogeneous content-driven video-retargeting. In *Proceedings of the IEEE 11th International Conference on Computer Vision*. 1–6.

WU, Y. P., CHEN, J., AND LI, Q. 2008. Extracting loosely structured data records through mining strict patterns. In *Proceedings of the IEEE 24th International Conference on Data Engineering*. 1322–1324.

XIE, L. X., NATSEV, A., AND TESIC, J. 2007. Dynamic multimodal fusion in video search. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. 1499–1502.

XIE, L. X., SUNDARAM, H., AND CAMPBELL, M. 2008. Event mining in multimedia streams. *Proc. IEEE 96*, 4, 623–646.

XIN, J., LIN, C.-W., AND SUN, M.-T. 2005. Digital video transcoding. *Proc. IEEE 93*, 1, 84–97.

XIONG, Z. Y., ZHOU, X. S., TIAN, Q., RUI, Y., AND HUANG, T. S. 2006. Semantic retrieval of video: Review of research on video retrieval in meetings, movies and broadcast news, and sports. *IEEE Signal Process. Mag. 18*, 3, 18–27.

XU, C. S., WANG, J. J., WAN, K. W., LI, Y. Q., AND DUAN, L. Y. 2006. Live sports event detection based on broadcast video and web-casting text. In *Proceedings of the 14th ACM International Conference on Multimedia*. ACM, New York, 221–230.

YAN, R. AND HAUPTMANN, A. 2006. Probabilistic latent query analysis for combining multiple retrieval sources. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, 324–331.

YAN, X., WAN, K. W., TIAN, Q., LEONG, M. K., AND XIAO, P. 2004. Two dimensional timeline and its application to conversant media system. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. 507–510.

YANG, B., MEI, T., HUA, X.-S., YANG, L. J., YANG, S.-Q., AND LI, M. J. 2007. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*. ACM, New York, 73–80.

YANG, C. C. AND CHAN, K. Y. 2005. Retrieving multimedia web objects based on a pagerank algorithm. In *Proceedings of the 14th International World Wide Web Conference*. ACM, New York, 906–907.

YU, B., MA, W.-Y., NAHRSTEDT, K., AND ZHANG, H.-J. 2003. Video summarization based on user log enhanced link analysis. In *Proceedings of ACM International Conference on Multimedia*. ACM, New York, 382–391.

YU, J. AND SRINATH, M. D. 2001. An efficient method for scene cut detection. *Pattern Recogn. Lett. 22*, 13, 1379–1391.

YUAN, J. H., WANG, H. Y., XIAO, L., ZHENG, W. J., LI, J. M., LIN, F. Z., AND ZHANG, B. 2007. A formal study of shot boundary detection. *IEEE Trans. Circuits Syst. Video Technol. 17*, 2, 168–186.

ZAHN, C. AND FINKE, M. 2003. Collaborative knowledge building based on hyperlinked video. In *Proceedings of the Computer Support for Collaborative Learning (CSCL'03)*. 173–175.

ZENG, W., ZHENG, Q.-F., AND ZHAO, D. B. 2005. Image guard: An automatic adult image recognition system. *Chinese High Technol. Lett. 15*, 3, 11–16.

ZENG, W. J., YU, H., AND LIN, C. 2006. *Multimedia Security Technologies for Digital Rights Management*. Elsevier, Amsterdam.

ZHAI, Y., YILMAZ, A., AND SHAH, M. 2005. Story segmentation in news videos using visual and text cues. In *Proceedings of the 4th International Conference on Image and Video Retrieval*. Lecture Notes in Computer Science, vol. 3568, Springer, Berlin, 92–102.

ZHANG, H. J., ZHONG, D., AND SMOLIAR, S. W. 1997. An integrated system for content-based video retrieval and browsing. *Pattern Recogn. 30*, 4, 643–658.

ZHANG, L., ZHU, J., AND YAO, T. 2004. An evaluation of statistical spam filtering techniques. *ACM Trans. Asian Lang. Inform. Process. 3*, 4, 243–269.

ZHANG, R., SARUKKAI, R., CHOW, J.-H., DAI, W., AND ZHANG, Z. 2006. Joint categorization of queries and clips for web-based video search. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. ACM, New York, 193–202.

ZHANG, X., LIU, J. C., LI, B., AND YUM, P. 2005. DONet/CoolStreaming: A data-driven overlay network for live media streaming. In *Proceedings of the IEEE Conference on Computer Communications*. 2102–2111.

ZHANG, Y.-J. 2006. A*dvances in Image and Video Segmentation*. IRM Press, Idea Group Inc.

ZHENG, Q.-F., ZENG, W., GAO, W., AND WANG, W.-Q. 2006. Shape-based adult images detection. *Int. J. Image Graph. 6*, 1, 115–124.

ZHU, X. Q., FAN, J. P., ELMAGARMID, A. K., AND WU, X. D. 2003. Hierarchical video content description and summarization using unified semantic and visual similarity. *Multimedia Syst. 9*, 1, 31–53.

ZHU, X. Q., WU, X. D., ELMAGARMID, A. K., AND AAND L. D. WU, Z. F. 2005. Video data mining: Semantic indexing and event detection from the association perspective. *IEEE Trans. Knowl. Data Engin. 17*, 5, 665–676.

ZHU, Y. Y. AND MING, Z. 2008. Svm-based video scene classification and segmentation. In *Proceedings of the International Conference on Multimedia and Ubiquitous Engineering*. 407–412.

Received April 2006; revised May 2008; accepted February 2009