Multimedia Information Retrieval at a Crossroad

Qing Li
City University of Hong Kong, China

Yi Zhuang
Zhejiang University, China

Jun Yang
Carnegie Mellon University, USA

Yueting Zhuang
Zhejiang University, China

INTRODUCTION

From late 1990s to early 2000s, the availability of powerful computing capability, large storage devices, high-speed networking, and especially the advent of the Internet, led to a phenomenal growth of digital multimedia content in terms of size, diversity, and impact. As suggested by its name, "multimedia" is a name given to a collection of data of multiple types, which include not only "traditional multimedia" such as images and videos, but also emerging media such as 3D graphics (like VRML objects) and Web animations (like Flash animations). Furthermore, relevant techniques have been developed for a growing number of applications, ranging from document editing software to digital libraries and many Web applications. For example, most people who have used Microsoft Word have tried to insert pictures and diagrams into their documents, and they have the experience of watching online video clips such as movie trailers from Web sites such as YouTube.com. Multimedia data have been available in every corner of the digital world. With the huge volume of multimedia data, finding and accessing the multimedia documents that satisfy people's needs in an accurate and efficient manner becomes a nontrivial problem. This problem is referred to as *multimedia information retrieval*.

The core of multimedia information retrieval is to compute the degree of relevance between users' information needs and multimedia data. A user's information need is expressed as a *query*, which can be in various forms such as a line of free text like "*Find me the photos of George Washington*," a few keywords like "*George Washington photo*," a media object like a sample picture of George Washington, or their combinations. On the other hand, multimedia data are represented using a certain form of summarization, typically called *index*, which is directly matched against queries. Similar to a query, the index can take a variety of forms, including keywords, visual features such as color histogram and motion vector, depending on the data and task characteristics.

For textual documents, mature information retrieval (IR) technologies have been developed and successfully applied in commercial systems such as Web search engines. In comparison, the research on multimedia retrieval is still in its early stage. Unlike textual data, which can be well represented by term vectors that are descriptive of data semantics, multimedia data lack an effective, semantic-level representation that can be computed automatically, which makes multimedia retrieval a much harder research problem. On the other hand, the diversity and complexity of multimedia data offer new opportunities for the retrieval task to be leveraged by the techniques in other research areas. In fact, research on multimedia retrieval has been initiated and investigated by researchers from areas of multimedia database, computer vision, natural language processing, human-computer interaction, and so forth. Overall, it is currently a very active research area that has many interactions with other areas.

In the coming sections, we will overview the techniques for multimedia information retrieval, followed by a review on the applications and challenges in this area. Then, the future trends will be discussed, and some important terms in this area are defined at the end of this chapter.

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

MULTIMEDIA RETRIEVAL TECHNIQUES

Despite the various techniques proposed in literature, there exist three major approaches to multimedia retrieval, namely text-based approach, content-based approach, and hybrid approach. Their main difference lies in the type of index used for retrieval: the first approach uses text (keywords) as index, the second one uses low-level features extracted from multimedia data, and the last one uses the combination of text and low-level features. As a result, they differ from each other in many other aspects ranging from feature extraction to similarity measures.

Text-Based Multimedia Retrieval

Text-based multimedia retrieval approaches apply mature information retrieval techniques to the domain of multimedia retrieval. A typical text-IR method matches text queries issued by users with descriptive keywords extracted from documents. To use this method for multimedia retrieval, textual descriptions in the form of "bag of keywords" need to be extracted to describe multimedia objects, and user queries must be expressed as a set of keywords. Given the text descriptions and text queries, multimedia retrieval boils down to a text-IR problem. In early years, such descriptions were usually obtained by manually annotating the multimedia data with keywords (Tamura & Yokoya, 1984). This approach is not scalable to large data if the number of human annotators is limited, but is applicable if the annotation task is shared among a large population of users. This is the case of several image/video sharing Web sites, such as YouTube.com and Flickr.com, where users add (keyword) tags on their photos or videos such that they can be found by keyword search. The vulnerability to human bias is always an issue with manual annotations. There have been also proposals from computer vision and pattern recognition areas on automatically annotating the images and videos with keywords based on their low-level visual/audio features (Barnard, Duygulu, Freitas, Forsyth, Blei, & Jordan, 2003; Jeon, Lavrenko, & Manmatha, 2004). Most of these approaches involve supervised or unsupervised machine learning, which tries to map low-level features into descriptive keywords. However, due to the large gap between multimedia data forms (e.g., pixels, digits) and their semantic meanings, these approaches cannot produce high-quality keyword annotations. Some of the systems are semi-automatic, attempting to propagate keywords from a set of initially annotated objects to other objects. In some other applications, descriptive keywords can be easily accessible for multimedia data. Particularly, for images and videos embedded in Web pages, the text surrounding them as well as the title of the Web pages usually provide good descriptions, an approach explored both in research (e.g., Smith & Chang, 1997) and also in commercial image/video search engines (e.g., Google Image Search).

Since relatively speaking keyword annotations can precisely capture the semantic meanings of multimedia data, the text-based retrieval approach is effective in terms of retrieving multimedia data that are *semantically relevant* to the users' needs. Moreover, because many people find it convenient and effective to use text (keywords) to express their information requests, as demonstrated by the fact that most commercial search engines (e.g., Google) support text queries, this approach has the advantage of being amenable to average users. But the bottleneck of this approach is still on the acquisition of keyword annotations, especially when there is a large amount of data and a small number of users, since no techniques provide both efficiency and accuracy in acquiring annotations when they are not available.

Content-Based Multimedia Retrieval

The idea of **content-based retrieval** first came from the area of content-based image retrieval (CBIR) (Flickner, Sawhney, Niblack, Ashley, Huang, Dom, et al., 1995; Smeulders, Worring, Santini, Gupta, & Jain, 2000). Gradually, the idea has been applied to the retrieval tasks for other media types, resulting in content-based video retrieval (Hauptmann et al., 2002; Somliar, 1994) and content-based audio retrieval (Foote, 1999). The word "content" here refers to the low-level representation of the data, such as pixels for bitmap images, MPEG bit-streams for MPEG-format video, and so on. Content-based retrieval, as opposed to text-based retrieval, exploits the features that are (automatically) extracted from the low-level representation of the data, usually denoted as low-level features since they do not directly capture the high-level meanings of the data. (In a sense, text-based retrieval of documents is also "content-based," since keywords are extracted from the content of documents.) Obviously, the low-level features used for retrieval depend on the type of data to

7 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/multimedia-information-retrieval-crossroad/17508

Related Content

Analysis of Platforms for E-Learning
Maribel-Isabel Sánchez-Segura (2009). *Encyclopedia of Multimedia Technology and Networking, Second Edition (pp. 22-34).*
www.irma-international.org/chapter/analysis-platforms-learning/17378

Cross-Media Publishing and Storytelling
(2019). *Cross-Media Authentication and Verification: Emerging Research and Opportunities (pp. 135-154).*
www.irma-international.org/chapter/cross-media-publishing-and-storytelling/208004

Towards Robust Invariant Commutative Watermarking-Encryption Based on Image Histograms
Roland Schmitz, Shujun Li, Christos Grecosand Xinpeng Zhang (2014). *International Journal of Multimedia Data Engineering and Management (pp. 36-52).*
www.irma-international.org/article/towards-robust-invariant-commutative-watermarking-encryption-based-on-image-histograms/120125

Automatic Pitch Type Recognition System from Single-View Video Sequences of Baseball Broadcast Videos
Masaki Takahashi, Mahito Fujii, Masahiro Shibata, Nobuyuki Yagiand Shin'ichi Satoh (2010). *International Journal of Multimedia Data Engineering and Management (pp. 12-36).*
www.irma-international.org/article/automatic-pitch-type-recognition-system/40983

Towards Fusion of Textual and Visual Modalities for Describing Audiovisual Documents
Manel Fourati, Anis Jedidi, Hanen Ben Hassinand Faiez Gargouri (2015). *International Journal of Multimedia Data Engineering and Management (pp. 52-70).*
www.irma-international.org/article/towards-fusion-of-textual-and-visual-modalities-for-describing-audiovisual-documents/130339