Multimedia Information Retrieval at a Crossroad

Qing Li
City University of Hong Kong, China

Jun Yang
Carnegie Mellon University, USA

Yueting Zhuang
Zhejiang University, China

INTRODUCTION

In the late 1990s, 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 multiple types of data, 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, multimedia techniques have been penetrating into 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. In other words, multimedia data have been 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 became a non-trivial problem. This problem is defined 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 key words, like, "George Washington photo"; or a media object, like a picture of George Washington. Moreover, the multimedia data are also represented by a certain form of summarization, typically called an index, which is directly matched against queries. Similar to a query, the index can take a variety of forms, including key words and features such as color histograms and motion vectors, 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 key words as an index, multimedia data lack an effective, semantic-level representation (or index) that can be computed automatically, which makes multimedia retrieval a much harder research problem. On the other hand, the diversity and complexity of multimedia offer new opportunities for its retrieval task to be leveraged by the state of the art in various 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 following sections, we will overview the techniques for multimedia information retrieval and review the applications and challenges in this area. Then, future trends will be discussed. Some important terms in this area are defined at the end of this article.

MULTIMEDIA RETRIEVAL TECHNIQUES

Despite the various techniques proposed in literature, there exist two major approaches to multimedia retrieval; namely, text-based and content-based. Their main difference lies in the type of index: The former approach uses text (key words) as the index, whereas the latter uses low-level features extracted from multimedia data. As a result, they differ from each other in many other aspects, ranging from feature extraction to similarity measurement.

Text-Based Multimedia Retrieval

Text-based multimedia retrieval approaches apply mature IR techniques to the domain of multimedia retrieval. A typical text-IR method matches text queries posed by users with descriptive key words extracted from documents. To use the method for multimedia, textual descriptions (typically key word annotations) of the multimedia objects need to be extracted. Once the textual descriptions are available, multimedia retrieval boils down to a text-IR problem. In early years, such descriptions were usually obtained by manually annotating the multimedia data with key words (Tamura & Yokoya, 1984). Apparently, this approach is not scalable to large datasets, due to its labor-intensive nature and vulnerability to human biases. There also have been proposals from computer vision and pattern recognition areas on automatically annotating the images and videos with key words based on their low-level visual/audio features (Barnard, Duygulu, Freitas, Forsyth, Blei, D. & Jordan, 2003). Most of these approaches involve supervised or unsupervised machine learning, which tries to map low-level features into descriptive key words. However, due to the large gap between the multimedia data form (e.g., pixels, digits) and their semantic meanings, it is unlikely to produce high-quality key word annotations automatically. Some of the systems are semi-automatic, attempting to propagate key words from a set of initially annotated objects to other objects. In other applications, descriptive key words can be easily accessible for multimedia data. For example, for images and videos embedded in Web pages, the text surrounding them is usually a good description, which has been explored in the work of Smith and Chang (1997).

Since key word annotations can precisely describe 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 (or key words) 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 key word annotations, since there are no indexing techniques that guarantee both efficiency and accuracy if the annotations are not directly 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, Gorkani, Hafner, Lee, Petkovic, Steele & Yanker, 1995; Smeulders, Worring, Santini, Gupta & Jain, 2000). Gradually, the idea has been applied to 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 bottom-level representation of the data, such as pixels for bitmap images, MPEG bit-streams for MPEG-format video and so forth. Content-based retrieval, as opposed to a text-based one, 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 key words are extracted from the content of documents.) The low-level features used for retrieval depend on the specific data type: A color histogram is a typical feature for image retrieval, motion vector is used for video retrieval, and so forth. Despite the heterogeneity of the features, in most cases, they can be transformed into feature vector(s). Thus, the similarity between media objects can be measured by the distance of their respective feature vectors in the vector space under certain distance metrics. Various distance measures, such as Euclidean distance and