Guest Editor's Introduction

Integrating and Using Large Databases of Text, Images, Video, and Audio

Alexander G. Hauptmann, Carnegie Mellon University

//Author: I've rearranged the material to provide what I hope is a more logical structure. Is it acceptable?//

With the advent of relatively cheap, large online storage capacities and advances in digital compression, comprehensive sources of text, image, video, and audio (TIVA) can be stored and made available for research and applications. The processing of a single medium has seen significant progress, especially for pure text sources. Also, images are frequently processed and made available through a query-by-example procedure (that is, find another image that has similar colors, textures, and shapes as this one).

However, the processing of a combination of multiple types of data has not been explored as thoroughly. Most TIVA sources were not produced with computer processing in mind. In contrast with text processing, few effective methods exist for understanding or even searching the content of combined TIVA sources. Intelligent, content-understanding systems can greatly improve the usefulness of the huge quantities of existing material from these sources. Collecting and intelligently integrating several of these media sources open up opportunities for novel applications of existing AI techniques and for further development of intelligent technologies. Unfortunately, there is no clear categorization or organization of the various research efforts concerning mixed-media databases.

In this issue

However, in recent years several workshops have focused on multimedia databases, learning, and their integration, thus spurring research. This special issue presents examples of current research and potential future contributions of intelligent, integrated systems using TIVA sources. These articles demonstrate exchange and cross-fertilization across the fields of vision, speech processing, natural-language processing, machine learning, and information retrieval. They all describe interesting combinations across multiple media, looking at how large amounts of data can be extracted, integrated into another system, and used in applications.

In "Named Faces: Putting Names to Faces," Ricky Houghton elegantly combines a variety of approaches—face recognition in images, OCR over the text on the screen, and Web spiders. The resulting application constructs a database and allows queries to that database.

In "Learning to Recognize Speech by Watching Television," Photina Jang and I describe a method that leverages the closed-captioned text to provide training data for any speech-recognition system.

"Image Retrieval Agent: Integrating Image Content and Text," by Jesus Favela and Victoria Meza, looks at ways to search for images found on the Web. **//Author: I finished the next sentence; is it accurate?//** The authors combine query by example in the visual domain with traditional text retrieval.

Finally, in "Retrieving Related TV News Reports and Newspaper Articles," Yasuhiko Watanabe, Yoshihiro Okada, Kengo Kaneji, and Yoshitaka Saka discuss a way to align television and newspaper articles on the same news item.

TIVA opportunities

In terms of potential impact, fields ranging from medicine (mixed-media patient records and data, evolving over time), to entertainment (video, audio, and images accessed over the Web), to education (multimedia training materials, searching historical and scholarly collections), to business and military information gathering could all benefit from advances in the processing of combined voice, image, video, and audio data.

Ample opportunities exist for cross-disciplinary work: digital signal processing is used for the basic processing of images, voice, and video. Research in very large databases provides clustering techniques and various tree-based access methods. AI and machine learning provide tools for classification and learning. Statistics provides tools to discover trends and analyze the data. Information retrieval provides time-tested fundamental text indexing and search techniques, which can be combined with visual and audio material.

Open issues

Several generic long-term problems are open research questions:

- What is the best way to pose multimedia queries to a system? This question has implications for both mixed media and human factors. Can users use multimedia queries more effectively than existing text-only queries?
- Can we do data mining on mixed-media databases? How can we exploit advances in data mining?
- What kind of information can we learn or extract from multimedia databases? Interesting opportunities exist for research on cross-media training and learning. Can we obtain better performance on a task by leveraging information from another source? For example, we want to be able to correlate speech segments to text, or speech to images, faces, or specific persons.
- How can we deal with more data? Too many techniques look really good on paper for small data sets but do not scale up to larger, real-life databases. For example, if a comparison of the similarity of two faces takes one-tenth of a second, and the process is linear, the system will not scale beyond a few thousand faces in an application. A related second issue concerns the quality of the process. The precision retrieval of an image-matching process might be quite good for 500 images, but for a 500,000-image database, the results could be unusable.
- **//Author: I combined parts of two bullets here because they both discussed query by content.//** How can a mixed-media database be processed to allow query by concept, rather than the query by keyword, pixel, or image statistics that we currently use? For example, if the user gives a sample image of a football game (human-like blobs in a green background), current systems will find images with similar amounts of green, and so on. The true goal is to find all images related to the concept of football, even if the colors and shapes are completely different.

//Author: We need a short biographical sketch containing, in the following order, current position and technical interests, prior applicable professional experience, education, professional affiliations, and address. I've started the sketch using material from your Web page.//

Alex Hauptmann is a senior systems scientist at Carnegie Mellon University's School of Computer Science. His research interests are in speech recognition, speech synthesis, speech interfaces, and language in general. **//Author: From where did you receive your degrees, and in what area are they?//** Contact him at the School of Computer Science, Carnegie Mellon Univ., 5000 Forbes Ave., Pittsburgh, PA 15213-3890; alex@cs.cmu.edu; http://www.cs.cmu.edu/afs/cs.cmu.edu/user/alex/www/HomePage.html.