Multimedia Information Extraction Roadmap

Rohini K. Srihari

Dept. of Computer Science & Eng, University at Buffalo
CEDAR/UB Commons, 520 Lee Entrance, Buffalo, NY 14228

Critical Technical Challenges

What are the critical technical challenges in multimedia information extraction (MMIE)?

There are several challenges, on several fronts. Some of these include:

- Detecting events of interest in video where there is no accompanying sound or text; examples include surveillance video. Further advances in computer vision, perhaps combining multiple 2D views are necessary. It is interesting to note that in the UK, it is almost impossible to walk outside for 5 minutes without being captured by some surveillance video system
- Content extraction from noisy media, such as telephone conversations, home videos (as seen on youtube)
- Correlating multimedia data to other data sources, especially text sources. Frequently, multimedia data (such as video, maps, charts) are referred to in email and chat conversations. The ability to automatically and dynamically correlate these sources enhances understanding
- Semantic access to multimedia data, enabling granular search; perhaps this requires a comprehensive ontology of entities, relationships and events which is more geared to multimedia search.
- Efficient indexing and retrieval: much of multimedia content is being captured and/or delivered via mobile devices. Such devices have limited processing capabilities.

Existing Approaches

What are the important existing methods, techniques, data and tools that can be leveraged?

Methods and Techniques

The techniques range from natural language processing (NLP) to 3D computer vision. NLP techniques are reaching a degree of sophistication enabling the semantic indexing and retrieval of video and images in cases where there is accompanying text. This allows us to go far beyond traditional keyword based information retrieval. Much of the focus has switched to noisy text such as blogs, chat and speech.

Significant advances have also been made in image processing, including feature detection, object detection. Extraction from video has also seen significant progress in terms of identifying entities from video, as well as change detection and even limited event detection.

DataSets

It would be nice to have a data set of youtube video (analogous to the TREC blog data set) that represents consumer generated content. This should be annotated and query sets should be developed (reflecting real user queries) to evaluate the state of the art in multimedia information access due to content extraction.

Tools

Tools include NLP systems, speech detection, audio/video indexing, segmenters, and feature detectors.

Remaining Gaps

What key technology gaps remain that require focused research? [If possible, forecast when you believe these gaps will be filled in terms of 1 to 5 or 10 years in the future]

Besides advances in areas such as NLP and computer vision, speech recognition, etc. there is another area that should be looked at. It is necessary to develop a *semantic map* that maps low-level visual percepts into concepts that can be expressed through language: in other words, a fundamental mapping between language and vision. This is a grand challenge. Cognitive scientists such as Jackendoff and George Miller have discussed this, but have not attempted any practical implementation. With advances in machine learning, it may be possible to attempt such a connection. It is interesting to see that many research efforts in content-based image retrieval are attempting to combine text and image features. Although this is a start, a much more fundamental mapping, which also includes time-varying visual data is necessary.