Multimedia Information Extraction Roadmap

Greg Myers[1], Gokhan Tur[1], Lynn Voss[1], Bob Bolles[1], Sachin Kajarekar[1], Elizabeth Shriberg[1,2], Dilek Hakkani-Tür[2]

SRI International, Menlo Park, CA 94025
International Computer Science Institute (ICSI), Berkeley CA 94704

Critical Technical Challenges

Information about people, their activities and communication is one of the most important types of content to extract from video data. This information is expressed in both the audio and visual domains. The critical technical challenges for extracting such content include:

1) Understanding interactions between people – their relationships, functional roles, hierarchies and dominance; and understanding their activities.
2) Broadening the robustness of multimodal information extraction techniques beyond narrowly constrained sets of actions, limited venues, and reliance on highly instrumented collection environments.
3) Obtaining sufficient amounts of annotated data for training models and classifiers.
4) Developing novel multimodal fusion techniques for semantically complex tasks involving human behavior.

While humans naturally process the multimodal information, there has been little effort on building a fusion framework specifically for semantically complex tasks involving human-human communication. For such tasks, information fusion is especially critical as they involve co-dependent and integrated multimodal information. The different modalities typically provide supporting information toward a goal, and handling the task by considering only one of the modalities is sometimes impossible. A framework for an extensible and effective protocol for fusion is clearly needed for long-term progress in this area.

Existing Approaches

Methods and Techniques

Approaches for multimedia information extraction can build upon existing content extraction components, such as: detection, tracking, and identification of people and faces, gesture recognition, 3-D scene modeling, scene classification, diarization, audiovisual speaker detection, speaker identification, and automatic speech recognition.

Existing multimodal fusion techniques such as feature-level combination (early fusion), decision-level combination (late-fusion), and pipelined and cascaded approaches can also be utilized.

In the literature the importance of exploiting information from multiple modalities has been shown for a number of tasks such as addressee detection [1] or speaker verification [2]. However in most studies, individual modality processing systems are treated as black boxes. Much research is needed to come up with more effective ways of multimodal information extraction.

Data Sets

Recent advances in data-driven image, speech and language processing techniques combined with machine learning algorithms enable us to build high-performance, robust, portable, statistical models. However, data-driven classifiers are trained using large amounts of in-domain task data that is usually transcribed and annotated by humans. While it is cheap and easy to collect vast amounts of video, there are only a few highly annotated data sets used in the literature. Most broadcast data is simply transcribed and annotated with simple information such as anchor id. On the other hand, multimodal multiparty meetings are extensively used in the literature. The AMI (Augmented Multi-party Interaction) Consortium [3] and DARPA-funded CALO (Cognitive Assistant that Learns and Organizes) [4] projects concentrate on conference-room meetings with small numbers of participants. The CHIL (Computers in the Human Interaction Loop) [5] project collected a series of lectures dominated by a single presenter with shorter question/answer portions, as well as some "interactive" lectures involving smaller groups. AMI and CHIL also produced corpora of time-synchronized media, generally including close-talking and far-field microphones, microphone arrays, individual and room-view video cameras, and output from slide projectors and electronic white boards. While automatic speech recognition and diarization technologies for meeting processing are increasing in maturity because of regular NIST evaluations, deeper meeting understanding exploiting multiple modalities is still awaiting further investigation.

Remaining Gaps

As mentioned above, one of the key gaps for multimodal

information processing is the lack of publicly available annotated data. Given that regular government evaluations on certain fields such as speech and speaker recognition enabled significant advances and benchmark datasets, the area of multimodal information extraction is in desperate need for similar evaluations and datasets.

Note that this does not underestimate the importance of research on adapting models to unknown domains and tasks. We believe a collaborative approach for providing annotations for different tasks given the corpora is a very effective method for building datasets on top of existing ones. Actually this is how most of the AMI and ICSI [6] corpus annotations are gathered.

The key technology gaps that require focused research are:

- classification of venue, that is, understanding the environment of the video
- understanding a larger range of human activities and gestures, such as addressee detection exploiting content of the conversation and image data.
- understanding emotional content, using cues such as prosody, facial expression, and head and body movements
- developing multimodal fusion techniques that can combine features from multiple modalities and deal with one or more intermittently degraded or missing modalities
- exploiting different modalities to improve each other using supervised and/or unsupervised techniques.
- dealing with poor audio fidelity and multiple simultaneous speakers
- processing data recorded in under-instrumented environments, such as with far-field microphones and limited camera views.

References

[1] N. Jovanovic, R. op den Akker, and A. Nijholt. Addressee Identification in Face-to-Face Meetings. In Proceedings of the EACL. Trento, Italy, pp:169-176. 2006.

[2] Sanderson, C., Bengio, S., Bourlard, H., Mariethoz, J., Collobert, R., BenZeghiba, M., Cardinaux, F., and Marcel, S., "Speech and face based biometric authentication at IDIAP," in Proceedings of IEEE Intl. Conf. on Multimedia and Expo (ICME'2003), Baltimore, July 2003.

[3] Augmented multi-party interaction. http://www.amiproject.org.

[4] SRI Cognitive Agent That Learns and Organizes (CALO) Project. http://www.ai.sri.com/project/CALO.

[5] Computers in the Human Interaction Loop. http://chil.server.de.

[6] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI Meetings Corpus. In Proceedings of International Conference on Audio, Speech, and Signal Processing (ICASSP), Hong Kong, April, 2003.