John Garofolo, R. Travis Rose,
and Rainer Stiefelhagen

Eval-Ware: Multimodal Interaction

In this issue, "Best of the Web" presents online resources for research in and evaluation of multimodal interaction (MI). The goal of MI is to increase the user experience and the usability of applications through multiple interfaces/modes of interaction. MI work includes: modeling of human-human interaction, human-computer interaction (HCI) and multimodal dialog; design and implementation of multimodal interface agents and human-computer interfaces; and modeling and implementation of emotions in emotion-sensitive HCIs. Consequently, MI builds on and contributes to the extensive research and development work in the areas of multimodal processing; data structuring, fusion, fission, and presentation; and multimodal annotation, indexing and retrieval.

In the vast area covered by MI, several multimodal evaluations, workshops, research programs, and standardization efforts have been established and are described here. We include the type of MI effort and the available Web resources in square parentheses.

Please send suggestions for web resources of interest, proposals for columns, as well as general feedback, by e-mail to "Best of the Web" Associate Editor Alen Docef at adocef@vcu.edu.

MI EVALUATIONS AND WORKSHOPS

CLEAR

http://clear-evaluation.org/

[open evaluation; workshop; online proceedings; training/test/ground truth data]

The CLassification of Events, Activities, and Relationships (CLEAR) evaluation and workshop is an international collaborative effort to evaluate systems that are intended to recognize people, activities, interactions and relationships in human-human and/or human-object interaction scenarios. The participation of researchers in the yearly CLEAR evaluations is open (with free registration). Evaluation tasks have included person and face tracking, head pose estimation, acoustic event detection and classification. CLEAR Web resources include online proceedings and sample data.

RT

http://www.nist.gov/speech/tests/rt/

[open evaluation; workshop; evaluation tools; training/test/ground truth data]

The Rich Transcription (RT) program is held annually as a National Institute of Standards and Technology (NIST)-sponsored evaluation of language technologies within the meeting domain. RT and CLEAR work on multimodal signals and sensor fusion experiments, and their events are held jointly. RT 2006 included speech-to-text translation, diarization (the identification of the number of participants in a meeting and the creation of a list of speech time intervals for each participant), speech activity detection, and text recognition in video. RT Web resources include information about speech benchmark tests, tools, test beds, and publications.

TRECVID

http://www-nlpir.nist.gov/projects/trecvid/

[open evaluation; workshop; evaluation tools; training/test/ground truth data]

The Text Retrieval Evaluation Conference video track (TRECVid) is a NIST-sponsored, annual multimodal retrieval evaluation workshop, which was presented in this column in September 2006. TRECVid's multimodal aspects involve multimodal recognition and querying of audiovisual data. Recently, TRECVid has begun to explore BBC rushes (raw materials used to produce videos), which typically contain highly redundant clips with natural audio; therefore they present a new opportunity for training and evaluating automatic recognition systems. TRECVid Web resources include conference proceedings and evaluation tools.

MI RESEARCH PROGRAMS

AMI

http://research.amiproject.org/

[research program; meeting data sets]

The Augmented Multi-party Interaction (AMI) program researches new MI technologies to support human interaction, specifically in the context of smart-meeting rooms and remote meeting assistants. The areas of interest include multimodal input systems, meeting dynamics and human-human interaction modeling, dialogue acts, video summarization, browsing, and indexing. AMI has routinely conducted MI evaluations for person detection and tracking, visual focus of attention, and speech detection. AMI Web resources include meeting data sets (with free registration) and a sample meeting for inspection online (without registration).

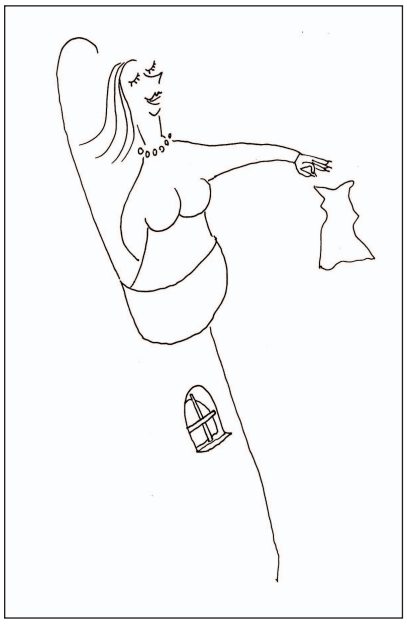
CHIL

http://chil.server.de/

[research program; evaluation]

The Computers in the Human Interaction Loop (CHIL) program is an integrated, collaborative project funded by the European Union that explores MI in modeling/interfaces for human interactions through face/person detection and tracking, head pose estimation, speech activity detection, speech

1053-5888/07/$25.00©2007IEEE



Multimodal interaction (Romeo and Juliet revisited) Cartoon by Tayfun Akgul (tayfun.akgul@ieee.org)

recognition, speaker diarization, multimodal person identification and tracking. The CHIL program has accumulated many hours of seminar and meeting data that are employed in its yearly MI evaluations (coordinated recently with CLEAR and RT.) CHIL Web resources include publications and a technology catalog.

CALO

http://caloproject.sri.com/

[research program]

The Defense Advanced Research Projects Agency (DARPA) project for Cognitive Assistant that Learns and Organizes (CALO) has the goal to create cognitive software systems that can reason, learn from experience, be told what to do, explain actions, reflect on their experience, and respond robustly to surprise via MI. The project is led by Stanford Research Institute International and brings together researchers in artificial intelligence, machine learning, natural language processing, knowledge representation, HCI, flexible planning, and behavioral studies. CALO Web resources include publications and tools (with free registration).

IMIM (OR IM2)

http://www.im2.ch/

[research program]

The National Center of Competence in Research on Interactive Multimodal Information Management (IM2) is an international multipartner program that researches and develops prototypes in human-machine interaction, focusing on technologies to coordinate natural input modes (such as speech, image, pen, touch, gestures, head and/or body movements, and physiological sensors) with multimedia system outputs (such as speech, sounds, images, three-dimensional graphics and animation.) IM2 Web resources include research and education information.

SIMILAR

http://www.similar.cc/CMS/

[research program; workshop]

The SIMILAR network of excellence is a European Union collaboration and annual workshop that includes special interest groups for fusion-fission research, adaptation, and usability, with application to HCI in medical and "edutainment" domains. SIMILAR research collaborations focus on recognition of speech, gesture, expression, handwriting; brain-computer interfaces, and other. SIMILAR Web resources include tools and links to a related project that supports the design and development of an open source platform for MI systems.

MI STANDARDIZATION EFFORTS

W3C MI

http://www.w3.org/2002/mmi/

http://en.wikipedia.org/wiki/W3C_MMI

[standards; use cases; MI framework]

Given the heterogeneous nature of the World Wide Web (WWW), it is no surprise that the WWW Consortium—W3C, its primary standardization body—is leading efforts to support multiple modes of interaction on the Web. The targeted applications of W3C's Multimodal Interaction Activity include mobile devices, automotive telematics, office automation, and home entertainment systems. A study of use cases and requirements led to the creation of the W3C MI Framework. Within this framework, the MI working group developed the Extensible Multi-Modal Annotation (EMMA) markup language as a standard data interchange format for multimodal systems, the Ink Markup Language (InkML) for electronic pen/stylus manipulations, and a Multimodal Architecture that specifies MI component interfaces and application control.

AUTHORS

John Garofolo (john.garofolo@nist.gov) is manager of the Speech Group in the National Institute of Standards and Technology's Information Access Division, in Gaithersburg, Maryland.

R. Travis Rose (travis.rose@nist.gov) is a computer scientist in the NIST Speech Group.

Rainer Stiefelhagen (stiefel@ira.uka.de) is an assistant professor with the Department of Computer Science at the University of Karlsruhe, Germany.

Disclaimer: Identification of commercial entities, equipment, or materials is not intended to imply recommendation or endorsement by the National Institute of Standards, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose they were presented. [SP]

Multimodal interaction (Romeo and Juliet revisited) Cartoon by Tayfun Akgul (tayfun.akgul@ieee.org)