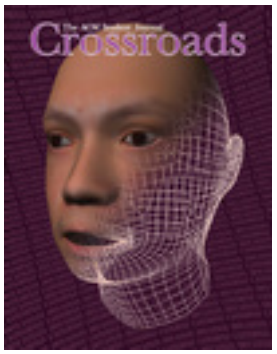
Crossroads
The ACM Student Magazine

Introduction: Computer Vision and Speech

by ***Niels Ole Bernsen***



If you are interested in computers with human capabilities, vision and speech open an entirely new world of computers that can *see* and *talk* like we do. Computer vision is the moody input cousin of computer graphics-in graphics, you have all the time you can afford to program the rendering, but visual input is an unpredictable and messy reality. Computer speech is both input and output, like in systems capable of spoken dialogue.

Viewed as enabling technologies, computer speech arguably holds the lead over computer vision. Even though a speech signal is enormously rich in information and we are still far from mastering important aspects of it like online recognition and generation of speech prosody, it is still much easier to shut up the people in a room in order to get a clear speech signal than it is to control the room's lighting conditions and to identify and track all of its 3-D contents independently of the viewing angle.

Given the state of the art, it makes good sense that the papers in this issue of *Crossroads* are about speech *or* vision. Two articles address different stages of the process of making computers understand what is commonly called the speaker's communicative intention, i.e., what the speaker really wishes to say by uttering a sequence of words.

Deepti Singh and Frank Boland discuss approaches to the important pre-(speech)-recognition problem of detecting if and when the acoustic signal includes speech in the first place. If no speech is present, there is no reason to spend computational resources on recognition or speaker identification, nor, perhaps, to steer a camera towards the source. Nitin Madnani's introduction to natural language processing, or NLP, is likely to tempt computer scientists to try out NLP for themselves.

The two articles on machine vision make two equally interesting points about the present state of the field. Taking human faces as an example, Justin Solomon compares the relative ease with which it is possible to solve complex face rendering problems with the difficulty of modeling the unique face each one of us has. Gang Gao and Paul Cockshott describe how smart use of computer image processing promises a robust shortcut solution to the integration of magnetic resonance images of the same object generated using two different imaging techniques.

Movies such as Arnold Schwarzenegger's *Terminator* series and Will Smith's *I, Robot* have engendered an overly ambitious public view of what robots and computing systems may accomplish when it comes to emulating human speech and vision. But perhaps we are closer to Hollywood's fantasy than we think. Honda's humanoid robot ASIMO is capable of visually recognizing faces, gestures, moving objects, and environmental terrain. ASIMO also has rudimentary audio capabilities, including recognizing voices and environmental sounds. Since there are only 20 ASIMO units worldwide at the moment, a more relevant example is Sony's robotic toy, AIBO. The mass production of a hearing and seeing robotic dog is exciting not just because it will not leave a mess behind, but because this field is beginning to shape everyday life. The next ten years promise to hold many more exciting developments.

This issue ends *Crossroads'* 13th year of continuous publication. Our four issues this year have included many highlights, such as a new cover style, updated layout, and exclusive interviews with Sid Meier and the inventor of the DNA sequencer. Next year will bring many new changes, including nonthemed issues, allowing us to cover a larger variety of topics within each issue. As always, please contact us (crossroads@acm.org) if you would like to submit an article, inquire about staff opportunities, or chat about technology.

Biography

Niels Ole Bernsen has worked in the field of spontaneous-speech, spoken dialogue systems for more than 15 years, from task-oriented flight ticket reservation through in-car spoken navigation and hotel reservation, to a system that engages kids in speech and gesture conversation with 3-D animated fairytale author Hans Christian Andersen. He is currently a professor in the Department of Mathematics and Computer Science at the University of Southern Denmark, and leads the department's Natural Interactive Systems Laboratory.