Computer Vision in 3D Interactivity

Organizer
Mark Holler
Intel Corporation

Panelists
Ingrid Carlbom
Bell Laboratories, Lucent Technologies

Steven Feiner
Columbia University

George Robertson
Microsoft Research

Demetri Terzopoulos
University of Toronto/Intel Corporation

With microprocessor clock rates in excess of 350MHz, SIMD integer instructions commonplace, and shared memory multiprocessing available for under $3,000.00, integration of computer vision with 3D graphics is now more practical than ever. Tracking the user's head, hands, and body, and detecting gestures, is one obvious direction to explore to eliminate encumbering sensors and enable new modes of interaction. Another direction is using computer vision techniques to understand 3D structure and camera parameters in multi-view image-based scenes for the purpose of re-rendering the scenes as a user directs. Yet another is giving animated characters visual awareness of users and other characters to enable richer interactions. What will be the most compelling integration of computer vision with 3D graphics?

The panelists address a subset of the following questions: What information besides human-user attitude/gesture can be extracted from images to enhance 3D interactivity. What other input modes are compatible with gesture and when? Is computer vision technology good enough today to be applied in commercial applications? If not, when? Is there a set of computer vision software components that would be useful to people working in 3D interactivity. What are the best applications for image-based rendering. Is the compute load small enough to run on today's machines? If not, when? Does system architecture need to change? What about memory and bus bottlenecks when multiple-video input channels are added to a system nearly bandwidth limited rendering graphics. Is computer vision + 3D graphics a big enough combination to drive the need for multiprocessing? Are there other standards, performance improvements, or specialized functions needed in video and multi-channel video capture for computer vision applications?

Audio-Visual Tracking for 3D User Interfaces

Ingrid Carlbom
Bell Laboratories, Lucent Technologies
Carlbom@research.bell-labs.com

Interactive virtual environment systems combine graphics, acoustics, and haptics to simulate the experience of immersive exploration of three-dimensional virtual worlds. Most such systems require users to wear cumbersome sensors for input and display units, eye- and headphones for the visual and auditory experience. However, the long-term goal for 3D interactivity is an interface more closely resembling human-to-human communication, depending more on multi-modal, unencumbering sensor and display technologies.

Tracking is a key technology for hands-free (unencumbered) 3D interactivity. Tracking can be used to determine user position and orientation, as well as user actions, such as gestures, facial expressions, and lip movements. While visual tracking with cameras alone has met with some success, the robustness of tracking can be increased if combined with acoustic tracking using microphones. Integrated acoustic and visual tracking can drive visual and auditory input, as well as output, to enhance the sense of immersion in a virtual world.

Camera and microphone-based tracking can be both complementary and cooperative to achieve accurate user localization. Camera-based tracking is particularly useful in acoustically noisy or reverberant environments, or to continue tracking a user who has temporarily stopped speaking while continuing to move. Similarly, acoustic tracking information from a microphone array can be used to localize the person who is speaking when several persons are present. This is particularly important under poor lighting conditions. User localization enables foveated processing for more detailed analysis of a user's gestures and expressions, as well as focusing of microphone beams on a user for high-fidelity speech input.

Accurate localization allows visual and auditory output to be directed to the user. The visual focus can be changed to the user's location (e.g., perspective vanishing point opposite the viewer, gaze of an avatar directed to the user). Auditory display in the form of spatialized sound can complement and enhance visual cues to aid in navigation, communication, comprehension, and sense of presence in virtual environments. Maximum fidelity and minimum disturbance to others is achieved if the acoustic output signal can be steered towards the listener. With a known user head position and orientation, combined with loudspeaker crosstalk cancellation, it will become possible to produce 3D spatialized sound for a moving user with virtual loudspeakers.

Steven Feiner
Columbia University
Feiner@cs.columbia.edu

Augmented reality refers to the use of see-through displays to overlay graphics, audio, and other media on the user's experience of the surrounding world. To accomplish this so that virtual objects are spatially registered with physical objects, we must be able to precisely track the 3D position and orientation of the user's head. As cameras and the compute power needed to process their input rapidly decrease in size and cost, the prospect of using computer vision for tracking becomes increasingly brighter. I discuss some of the issues involved in tracking for augmented reality, and potential advantages and disadvantages of using vision-based approaches. For example, one significant distinction of vision-based systems is the rich nature of the raw sensor data itself. Unlike other tracking technologies, input from one or more cameras can be used to perform object recognition, to build up a model of the surrounding environment, or just to document the user's experience.

Mark Holler
Intel Corporation
mark_holler@ccm.sc.intel.com

As computer performance marches forward according to Moore's Law, entirely new application domains are enabled. Digital imaging is currently going through a spurt of growth and will soon be followed by digital video processing. 3D graphics performance in PCs is also going through rapid growth now as 3D graphics accelerators proliferate. In addition to Moore's Law, there has been the addition of Single Instruction Multiple Data Instructions to most microprocessors. These instructions perform four or eight operations on four or eight pairs of 16-bit or eight-bit integers in parallel, typically in one clock, enabling a number of image-processing functions used in computer vision to be accelerated by 2-4X. Optimized libraries to achieve this acceleration are available for download on the Web [Performance Libraries]. Support for symmetric multiprocessing in mainstream CPUs such as the Pentium II and operating systems such as Windows NT has also provided a quantum leap in compute power available for integration of computer vision with 3D computer graphics. Bradski (1998) has reported a four-degree-of-freedom, 30fps head tracker using under 30 percent of one Pentium II CPU in a multithreaded app where head position/orientation controls fly above a 3D model of Hawaii. The second CPU and an E&S RealImage 3D accelerator are fully utilized for 3D rendering.

Immersive VR using HMDs requires a participant to wear the display and most often cumbersome sensors on head, hands, and body. "Fish-tank VR" (non-stereo) using computer-vision-based head tracking offers a less immersive experience but still provides control of motion parallax while freeing the user from wearing hardware. Arthur et al. (1993) and Rekimoto (1995) have shown that fish-tank VR enables users to understand complex 3D scenes more accurately than when given just static views. Ware et al (1993) have shown that motion parallax is a stronger cue for understanding 3D structure than stereopsis, suggesting that fish-tank VR is more effective in providing 3D cues than a stereo display, in addition to not requiring the user to wear shutter glasses. The narrower field of view of typical fish-tank VR systems is less likely to produce motion sickness.

Intuitive navigation in 3D spaces fundamentally requires more input than a mouse can provide. The mouse provides two degrees of freedom simultaneously while full 3D navigation requires six degrees of freedom, or more if viewing is de-coupled from navigation. Hand-controlled devices with six degrees of freedom require more attention to control than may be available during a 3D interactive game. Computer vision can extract some or all of the degrees of freedom from head position and orientation to reduce the required attention to hand coordination. Head movement such as peeking around corners to produce view changes is very intuitive for humans because we do it all the time in the real world. Used conservatively, tracking also promises to lower the interactivity bar for young children because of reduced requirements for fine-motor control.

Computer vision is capable of extracting 3D structure information from stereo views or motion sequences. With the view morphing approach [Seitz, Dyer 1996], a full 3D model of the scene need not be extracted to produce the novel views. This information is useful in producing novel views of an image-based scene. One can imagine an interactive telepresence application in which trackers know the positions and head orientations of participants and morph available view images to achieve eye contact and motion parallax cues. We have demonstrated such a capability in our labs.

Performance Library Suite: MMX technology optimized libraries in Image Process, Pattern Recognition, Signal Processing and Linear Algebra can be downloaded from developer.intel.com/design/perftool/perflibst/index.htm

G.Bradski. Computer Vision Face Tracking For Use in a Perceptual User Interface, Intel Technology Journal, developer.intel.com/technology/itj/q21998/articles/art2.htm Q2, 1998

K.Arthur, S. Kelogg, S.Booth, C.Ware. Evaluating 3D task performance for fish tank virtual worlds. ACM Transactions on Information Systems, Vol 11, No. 3, pp 239-265, 1993.

J.Rekimoto. A Vision-Based Head Tracker for FishTank Virtual Reality - VR without Head Gear. Proceedings of the 1995 IEEE Annual Virtual Reality International Symposium, March 1995 pp: 94-100.

C. Ware, K.Arthur, and K.S. Booth. Fish tank virtual reality, in INTERCHI'93 Conference Proceedings, pp37-42, 1993.

S.Seitz, C.Dyer. View Morphing. Computer Graphics Proceedings, SIGGRAPH 96.

George Robertson
Microsoft Research
Ggr@microsoft.com
www.research.microsoft.com/ui/ggr/ggr.htm

Virtual Reality apparently attains its power by captivating the user's attention to induce a sense of immersion. This is usually done with a display that allows the user to look in any direction (like HMDs or CAVEs), and that updates the user's viewpoint by passively tracking the user's head motion. However, there are other forms of VR where immersion occurs. Fish-tank VR uses a desktop stereo display rather than surrounding the user visually. Desktop VR uses animated interactive 3D graphics to build virtual worlds with desktop displays and without head tracking.

Current HMD-based VR techniques suffer from poor display resolution, display jitter, and lag. These problems tend to inhibit the illusion of immersion. Fish-tank VR uses desktop stereo displays to solve display resolution and jitter problems. Desktop VR solves all three problems, but at the expense of losing stereo and head tracking. Studies have shown that head-motion parallax is a stronger depth cue than stereopsis. Hence, adding head-motion parallax to a Desktop VR system could bring it quite close to fish-tank VR capabilities. Computer vision can track the user head motion without the user wearing any tracking sensors. This has additional benefits of eliminating fatigue and making it easier (and more desirable) to use, thus enabling everyday or extended use.

Computer vision enables other capabilities that may make 3D interactivity more effective and enjoyable. Adding awareness to our systems becomes possible. The system can know whether the user is present, whether the user is facing the screen, whether the user is engaged in some other activity (like talking on the phone or to another person in the room), and what the user is looking at on the screen.

Combining computer vision and 3D does involve solving some problems. The devices (cameras) are not expensive and are becoming ubiquitous. In the near future, the standard PC will likely include a camera. However, computer vision is computationally expensive. We currently use multiprocessors, which are a bit more expensive. We are nearing a point when computer vision and 3D interfaces can be effectively integrated and enable a number of exciting new interface capabilities.

Demetri Terzopoulos
University of Toronto/Intel Corporation
dt@cs.toronto.edu

Interactive 3D virtual worlds populated by autonomous characters with realistic behaviors rely on perceptual information processing, especially computer vision, so that the characters can sense one another and the user. I review the state of the art of perceptual modeling for behavioral characters and discuss how new vision algorithms promise to couple interactive characters much more closely to the user.