Gesture Recognition Crosses the Chasm

By Sinclair Voss

Flip your hand to change channels on your TV? Dance your way to win a video game? Move your fingers to perform a distant surgery? For human-machine interaction, we are starting to command our tools without intermediate hardware: without mice, keyboards, joysticks, or other remote controls. We are beginning to control technology in a natural manner — with our voices and our gestures. And the use of 3D gesture recognition (GR), the technology that enables controlling machines with body movements and even facial expressions, has recently broken through from isolated applications to mainstream consumer electronics.

GR is a broad term that includes the use of accelerometers and other types of sensors that a user holds or wears. This article discusses GR that does not use remote sensors on the user and goes into the current uses of hands-free GR, how it works, and the markets for increasingly diverse applications.

Current Applications

Gaming is the first truly mass-market application for GR. Microsoft® Kinect replaces the remote control for the Xbox 360 game and entertainment console, calculating user commands with a set-top box by sensing and interpreting body movements. Sony, Nintendo, and other game console manufacturers are likely to quickly follow up with similar or next-generation offerings, due simply to the irresistible attraction of a more immersive and intuitive user experience.

In particular, fitness activities strongly lend themselves to GR-enabled gaming, and fitness games are the fastest growing segment of a market that PriceWaterHouseCooper estimates will be $68B USD in 2012. Typically, a fitness game tracks user activity closely and represents it with an on-screen avatar exercising in a variety of immersive visual realities: running by a lake, boxing in a ring, or dancing in a disco. For the avatar to truly represent the user, movement data needs to be gathered from all parts of the body, from head to toe. Holding an accelerometer while standing on a sensor pad provides very limited data, and from a user standpoint, holding any kind of remote control is too clunky. GR lets the user work out in a realistic way in a virtual environment without using any intermediate devices.

GR is turning marketing signage, kiosks, and even tradeshow exhibits into touch-free, interactive displays where a user can guide the information flow with pointed fingers and other hand gestures. Shell Oil recently launched a product demo roadshow using a 53 ft. trailer lined with interactive displays — visitors could look inside engines and browse screens of product information. Sprint is another user, with GR adding a considerable wow factor to digital signage in its brick-and-mortar retail outlets. The 2008 Beijing Olympics featured a hand-gesture-activated flight simulator to help market the 2010 Olympics in Vancouver.

High-volume consumer electronics other than gaming platforms have yet to implement GR capabilities, but it is safe to assume that the market for touch-free control of home entertainment systems, mobile devices, and the like will soon be flooded now that the Kinect device has broken the dam.

How It Works

The roots of simple gesture recognition technology go back to the use of a light pen to manipulate onscreen objects — moving them, changing sizes, and using constraints. The light pen communicated 2D hand gestures to the computer. In the first Apple computers, the mouse replaced the light pen, still working in 2D, but adding several degrees of precision and a button to execute commands. More recently, capacitive touch-screens have proliferated to virtually all smart-phone designs, and remote controls with accelerometers, such as Wii controls, are now commonplace gaming devices.

So how does a GR sensor pick up commands without any physical connection to the user?

By constantly sensing changes in light patterns reflected back to the eyes, the human brain is able to create a very exact 3D map of its immediate vicinity. This 3D map allows the human to seamlessly interact with this environment, be it to accurately (or not so accurately) hit a golf ball or dance a tango with a partner. Controller-less 3D GR systems work in the same way.

Let's take the example of a player swinging a golf club in computer game. A light source in the GR sensor head illuminates the whole area in front of it, including the player, with infrared light. An optical receiver then detects the light that reflects back. Optical filters in the sensor ensure that only this reflected light is detected, with spurious and ambient light filtered out.

Fast electronics in the sensor then process the received information, turning it into a 3D map which the computer or gaming console can interpret and use to insert the player real-time into the game. Different proprietary techniques can produce this 3D map. For example, some techniques use the round-trip time of the reflected light, while others use patterns encoded onto the light. Whatever the method, the end result is always the same, namely a very realistic gaming experience without any wires to trip over our handheld devices to throw through the TV.

Despite the number of different technologies that support these systems, most GR sensors share a basic component list:

- *Light Source* — an LED or a diode generating infrared or near-infrared light. This light isn't normally noticeable to users and is often optically modulated to improve the resolution performance of the system.
- *Controlling Optics* — optical lenses help optimally illuminate the environment as well as focus reflected light onto the detector surface. A bandpass filter lets only reflected light that matches the illuminator's light frequency reach the light sensor, eliminating ambient and other stray light that would degrade performance.
- *Light detector* — a high-performance optical receiver detects the reflected, filtered light and turns it into an electrical signal for processing by the firmware.
- *Firmware* — very-high-speed ASIC or DSP chips process the received information and turn it into a format which can be understood by the end-user application (for example, the computer game software).

Simply using software with a camera is a relatively low-cost GR solution that can be appropriate for simpler applications such as mobile phones. However, the increased bandwidth and processing speeds

needed for high-integrity video compression limit this kind of solution to recognizing the most basic of movements. Yet algorithmic software can play an important role in refining GR data output. Human joints have predictable limits, and there are only so many ways that body appendages can rotate. Reducing the number of possible command gestures can accelerate response times.

The Future

3D GR systems are an elegant and effective way to free up users from the burdens of electronic remote controls or wands or keyboards — hence, the intrinsic market demand for such systems is clear. The speed at which 3D GR is adopted will, however, depend on a number of factors:

- Performance
  The performance of human interface devices such as the gaming wand, the remote control, or the computer mouse have been perfected over decades. For 3D GR systems to replace these devices, they must offer a similar degree of performance. The first generation of gesture recognition systems is already close, if not equal, in terms of speed, resolution, and accuracy but further investment in high-speed optics and electronics will be necessary to ensure widespread adoption.

- Price
  The price premium that the end user is prepared to pay for the 3D GR functionality will of course be application dependent, but an attractive cost model is vital to drive market penetration into consumer markets. The optics and much of the electronics know-how within 3D GR systems has been leveraged from the more-mature telecommunications industry, enabling immediate competitive pricing as well as high-volume manufacturing capability. However, challenges still remain for particularly cost-sensitive applications such as cell phones.

- Size
  Particularly for mobile applications (for example, notebooks), the 3D GR hardware will need to be very compact and robust. First-generation 3D GR systems are still relatively large but investment has already started to develop more integrated components to shrink the size of the sensor head significantly.

The applications for compact, low-cost GR solutions are virtually limitless. Intel recently released a video in which a researcher conducts his laboratory like a symphony — adjusting light levels, his instruments, and his music with dramatic flourishes and gestures. Keyboards and mice may well be replaced by GR sensors integrated into monitors. Joysticks to manipulate heavy machinery may go the way of steam engines. The bottom line is replacing relatively complicated, expensive manufactured items with a compact, inexpensive unit consisting of a light source, sensor, and firmware.

Conclusion

3D GR is a fundamentally inexpensive, rapidly emerging technology. The hardware requires little in the way of raw materials, and it lends itself to very high-volume manufacturing. Whenever a control goes much beyond on/off in terms of complexity, the cost threshold for implementing 3D GR will be very low

and these new systems will significantly improve the ease with which every one of us interacts with our electronic environment. This will be an exciting and fast-moving market over the next few years.

Touting Kinect, Microsoft posts: “technology evaporates, letting the natural magic in all of us shine.” It is a bit difficult to use the term “natural” when talking about virtual-reality gaming and the manipulation of computer-based environments. However, it is indeed the intuitive, instinctual use of body gestures to control the environment that makes GR an ultimate solution for human-machine interaction. Mice, keyboards, and joysticks are simply primitive GR solutions. Devices with accelerometers and gyroscopes were the next generation. Being able to recognize motion from an entire body, from two legs, two arms, 10 fingers, not to mention a face with its own library of complex expressions, is an entirely new frontier. Pairing GR with voice recognition, with its ability to detect emotional subtleties, will enable command complexity that will make our current implementations appear stone-aged.

Sinclair Vass received his BSc with honors in physics from the University of Edinburgh. He is the director of EMEA sales at JDSU, a major supplier of optical communications components, optical network test and measurement tools, and professional optical network services. Reach him via sinclair.vass@jdsu.com.