Kinect Body Tracking Reaps Renown

By Rob Knies
September 26, 2011 6:30 PM PT

By any standard, Kinect for Xbox 360 has proved to be a technological sensation. Kinect, the controller-free interface that enables users to interact with the Xbox 360 with the wave of your hand or the sound of your voice, sold 8 million units in its first 60 days on the market, a figure that makes it the fastest-selling consumer-electronics device in history, as confirmed by Guinness World Records.

That, of course, is a tribute to the Interactive Entertainment Business (IEB), which produced Kinect—and, by extension, to Microsoft Research, which made several key contributions to the technology.

For many, the most noteworthy part of Kinect is its ability to track the body movements of users and provide natural interaction as a result, and critical portions of that work came from Microsoft Research Cambridge.

Jamie Shotton, Andrew Fitzgibbon, Andrew Blake, Toby Sharp, and Mat Cook of Microsoft Research Cambridge each made seminal contributions to the success of Kinect, and the value of their research has been widely praised, in particular during early June, when they received the MacRobert Award, the most significant honor bestowed for innovation by The Royal Academy of Engineering. Another member of the Cambridge contingent who helped shape the skeletal-tracking feature was Oliver Williams, who worked at that facility before transferring to Microsoft Research Silicon Valley.

Researchers from Microsoft Research Cambridge beam during the ceremony in which they won the MacRobert Award: (*from left*) Mat Cook, Jamie Shotton, Andrew Blake, Andrew Fitzgibbon, and Toby Sharp.

"This technology is a radical development in human-computer interaction," says Blake, a Microsoft distinguished scientist and managing director of Microsoft Research Cambridge. "First, we had the green screen, then mouse and keyboard, then touch and multitouch, and now, what could be called 'no-touch' interaction.

"It has radically advanced the state of the art in gaming, but that is just the beginning. This is a new kind of technology that could have far-reaching implications for the way we interact with different kinds of machines."

Fitzgibbon, a researcher at the facility, has an extensive background in computer-vision research connected with films and video, so he performed an invaluable role in consulting for the various components of the skeletal-tracking research. Shotton, also a researcher, devised the algorithm that takes an image from Kinect's depth camera and identifies the different parts of the body. Sharp, senior research

The Research Behind Kinect
- *Silicon Valley's Kinect Contributions*
- *Helping Kinect Recognize Faces*
- *Kinect Audio: Preparedness Pays Off*

Microsoft Research Anniversary
- 20th Anniversary website
- *20 Years On, a Future Brighter than Ever*
- *Changing the World of Science*
- *Why I Work for Microsoft Research*

Inside Microsoft Research
- *In Beijing, a Rousing Welcome*
- *Anniversary Momentum Hits the Subcontinent*
- *Exciting New Research in Merry Olde England*
- *Silicon Valley Talk Focuses on Security*
- *A Special, Interdisciplinary Approach*
- *20th Anniversary Gets a Fitting Finale*

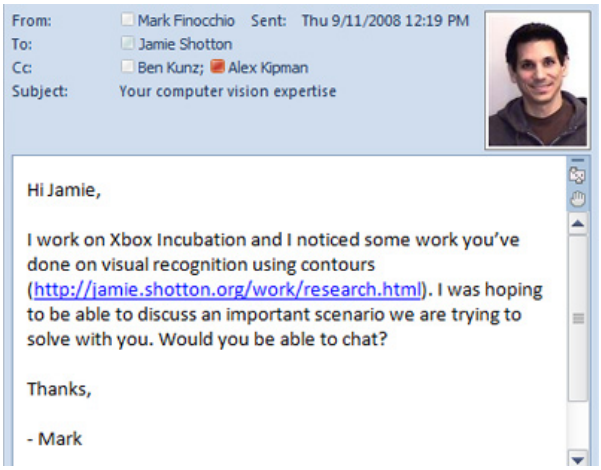
Products
- Kinect for Xbox 360

Facilities
- Microsoft Research Cambridge
- Microsoft Research Silicon Valley

People
- Jamie Shotton
- Andrew Fitzgibbon
- Andrew Blake
- Toby Sharp
- Oliver Williams
- Antonio Criminisi
- John Winn
- Carsten Rother
- Mihai Budiu

software-development engineer, worked on high-speed implementations of the algorithm for the Kinect system, and Cook, a contractor with extensive experience in working on computer games, was called in to provide a massive amount of training data.

"Kinect takes a stream of images coming off the camera," Shotton explains, "and quickly works out where the joints in your body are in 3-D. It can use that to animate characters and to manipulate objects on the screen."



This email began the collaboration that culminated in Kinect's skeletal-tracking ability.

The research effort that led to the skeletal-tracking ability of Kinect began in September 2008 with an email from Mark Finocchio, principal software engineer with IEB, to Shotton, asking for help with a planned accessory that could track bodies in real time for use in gaming.

The Xbox team already was using a depth camera and had written a tracking algorithm that could track a body's movements quickly. A video sent to Cambridge left the researchers there impressed.

"We saw that video," Fitzgibbon recalls, "and thought, 'OK, you've solved it, so why are you talking to us?'"

Soon enough, the reasons became apparent. To begin with, to start the recognition process, the user had to strike a particular pose with arms extended. More important, the algorithm typically would work well for a while, then, when the motion input became unpredictable, the body tracking could "crumple" into unusable disarray, and the only way to reset the tracking system was to strike the pose again. Because the system sometimes failed after only about a minute, it simply was not feasible for extended game play.

"What they wanted," Fitzgibbon recalls, "was a way of initializing the algorithm, recovering from these crumplings—preventing them in the first place, ideally—and making it work for everyone, no matter your body shape and size."

The algorithm's handicap was that, basically, it was relying on analysis of past motions to predict the future. When a user began moving deliberately, the algorithm would assume that was the natural pace of motion, so when certain types of fast motion occurred—for example, when a person moved too far within a 30-millisecond span, which easily can occur when people are playing an exciting video game—it couldn't keep up.

One at a Time

"We realized that we had to look at a single image at a time," Shotton says. "We couldn't rely exclusively on context, your history, or your motion in the past. We had to just look at an image and decide what your body pose was. We knew that this was, in theory, possible, because a person can do this. If you look at a photo, you can draw the position of the joints."

Related work in this area existed in the computer-vision literature, and the team tried one promising approach.

"It would try to match your whole body at once," Shotton says. "It would have a big database of the way the body appears in different positions. You'd take the image coming off the camera and search through this big 'flipbook' of different body positions and try each of them against the image until you find the one that matches

Publications

- *Cambridge Engineers 'Kinect' with Judges to Land U.K.'s Most Valuable Engineering Innovation Prize*
- *TextonBoost: Joint Appearance, Shape and Contect Modeling for Multi-Class Object Recognition and Segmentation*
- *Implementing Decision Trees and Forests on a GPU*
- Real-Time Human Pose Recognition in Parts from Single Depth Images

best.

"That kind of worked. We got Xbox very excited quickly. But we realized early on that it wasn't going to scale up to our needs. It's essentially a brute-force approach, and you have to represent every possible combination of human pose and body shape and size into the training set."

In addition, the number of joints under such analysis led to a gigantic number of potential body poses.

"Let's say I can bend my right elbow into 10 positions," Shotton continues, "and I can move my right shoulder in a hundred different positions. That's 10 times 100—a thousand. I've got another 1,000 on the left side, so when I multiply those, we're at a million. And then we've got maybe a hundred more positions for this joint or that joint. You get this exponential growth in the number of positions the human body can make. This was obviously quite a major issue."

Existing literature pointed a way out, though, by breaking up the problem: trying to match the hand separately to the shoulder to avoid the exponential compounding. Shotton decided to alter his approach. Instead of trying to predict the positions directly, he'd just examine every pixel in an image and try to determine which part of the body it represents.

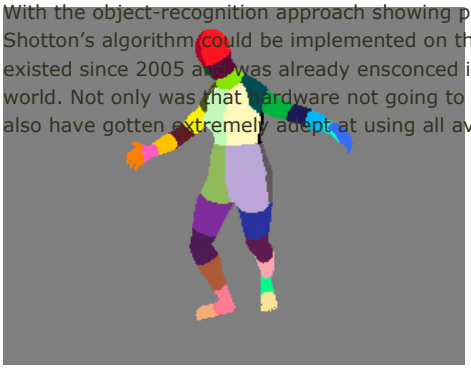
That, the team soon realized, was akin to earlier work on object recognition in images by Shotton, Sharp, and Microsoft Research Cambridge colleagues Antonio Criminisi, John Winn, and Carsten Rother on object recognition in images.

The sheep-and-grass example of object recognition in images.

"We usually use the sheep-and-grass example," Shotton says. "You take a photo of a sheep on grass and try to segment the sheep against the grass and label those automatically as sheep and grass. We knew this worked. The key realization for Kinect was that we can take that, apply it to this new problem, where we were trying to effectively color in the body. We got an image from the camera, and we're trying to color in the body with different body-part colors.

"If we can take an image coming off the depth camera, a gray-scale image where the different pixels represent depth to the sensor, and convert it to this body-part coloring, we have an extremely good idea of where the joints in the body are, because we defined these parts to be near the joints. If you can get all of the pixels that are labeled as the person's left hand and color them in correctly, by clustering those together and using the depth information, that gives us a really good idea about the 3-D XYZ coordinate of the left hand."

With the object-recognition approach showing promise, Sharp began to consider how Shotton's algorithm could be implemented on the Xbox 360 hardware, which had existed since 2005 and was already ensconced in millions of living rooms around the world. Not only was that hardware not going to be retrofitted, but game developers also have gotten extremely adept at using all available processing capability and memory to optimize their titles.


Body-part coloring identifies areas near joints in the body.

"It becomes a challenge to squeeze implementation of a state-of-the-art tracking algorithm onto pre-existing hardware," Sharp says, but he was able to apply more of his software-engineering expertise.

"Some work I'd done became relevant," he says, "about how to run these decision-forest algorithms on a graphics processor using DirectX [Microsoft technologies for running and displaying multimedia applications]. The sheep-and-grass example translated to the Kinect domain. The implementation on a graphics card translated to the Kinect domain, as well. A paper I published on that in 2008 was used as a starting point for the development of the Kinect implementation."

The researchers also needed training data. Enter Cook, in December 2008.

"We were going to use machine learning," Shotton says, "because we're in the machine-learning group, we have experience with it, and we know it can work really well. But the thing about machine learning is that you need data. That's why we brought Mat on board. His job was to develop a way of generating images that we needed.

"We couldn't capture real images of people and label them by hand. That would take too long, be too expensive, and we'd never get enough data. But if we could synthesize images using computer graphics, we could use those synthesized images for our training data. Synthesizing depth images, as opposed to the usual RGB color images, turns out to be a sweet spot of what we can achieve reliably. "

That enabled the collection of training data to begin in earnest.

"We began to acquire motion-capture data, 3-D joint positions," Fitzgibbon says. "Mat would feed that motion-capture data into a computer-graphics tool that generated depth images so we'd have something to test on where we knew the right answer. We had a ground-truth answer associated with each image."

At the outset, Cook had no idea exactly what he was helping to achieve. The real nature of the project was a secret.

"It wasn't immediately clear exactly what the whole system was going to be and how well it would work," Cook smiles. "I was told that we're working on a system that would plug into an Xbox and track people. The first challenge was to generate anything, to get some data that was useful fairly quickly."

Hundreds of Thousands of Poses

Off-the-shelf software was proving too slow or unable to generate a good depth image, but one tool was up to the task. It could work with existing motion data sets, enabling the use of motion data using different people and different poses. And it included a feature that fogs out items that are relatively further away. That provided image depth. A series of tweaks eventually enabled the synthesis of millions of images. Eventually, the training data ballooned into hundreds of thousands of body poses.

What followed was a lengthy period of incremental improvements in getting accuracy up to snuff.

"We could plot graphs of accuracy versus the number of training images," Shotton says. "You'd see it going up and up and up. We knew we had to keep extending the training set, but it was taking a week to train, and that wasn't fast enough. We spent a long time working with our colleagues at Microsoft Research Silicon Valley to make it train quickly."

Once they did, they realized they had an algorithm, dubbed Exemplar, that would work fast enough and accurately enough to run on every frame of the stream of data from the Kinect depth camera.

To obtain realistic body positions, they went global.

"We told the Xbox people we'd need real-world test data," Fitzgibbon says. "Amazingly, they sent a team of people with the prototype depth camera to 10 houses across the planet and asked the residents to dance around as if they were playing Kinect games. No one was allowed to see the data unless they were testing the algorithm."

In June 2009, the researchers attended an offsite meeting near Microsoft's Redmond, Wash., headquarters. By this point, Kinect was moving from incubation into production. The researchers explained the algorithm in more detail to the platform team, which was a bit wary of the machine-learning component. How, the product team asked, are we supposed to debug this? The team from Cambridge convinced its counterparts that machine learning was just another form of test-driven development and that more training data would improve performance.

Shortly after the meeting, the Xbox team wrote an algorithm of its own, one that takes the output from the Exemplar algorithm and applies additional wizardry to provide a full skeleton rather than a collection of joints and to provide the use of

temporal coherence.

Help from the Valley

The latter work was aided significantly by the work of Williams of Microsoft Research Silicon Valley. That facility also supplied Mihai Budiu, who collaborated with Shotton on distributing the training algorithm.

By January 2010, the final algorithm was basically complete. The GPU implementation had enabled the processing of 30 frames per second while using only 10 percent of the hardware resources.

"The algorithm is highly parallel," Fitzgibbon says. "Computation can be run on every pixel of an image independently, making it suitable for a graphics processor, and can be implemented using graphics primitives. Those primitives are based on rendering triangles to cover pixels in an image buffer. The result is that every pixel in the body is labeled, in real time, according to which part of the body it belongs."

Adds Shotton: "Had Toby not done this work on the GPU runtime of the decision forest, we may not have even considered it as a possibility, because we wouldn't have known that it was something you could do fast enough."

That, Sharp says, is a hallmark of Microsoft Research.

"It is an example of research," he observes, "that turned out to be crucial for a product almost immediately after it was done, even though it wasn't done with the product in mind."

At the time, the effort that eventually became Kinect was known as Project Natal. It was enthusiastically received once unveiled in Los Angeles in June 2009 during the Electronic Entertainment Expo. Renowned film director Steven Spielberg was particularly effusive in his praise.

It was around then that the Cambridge researchers got their hands on the first games designed to use the capabilities they had helped to create.

"We could immediately see," Shotton recalls, "that even if it doesn't work 100 percent, the games are going to be fun!"

Adds Fitzgibbon: "That was the first time we realized that, yes, we probably are going to get there."

That they did, as evidenced by the fact that Real-Time Human Pose Recognition in Parts from Single Depth Images, the paper that resulted from their work, received the best-paper award during the Institute of Electrical and Electronics Engineers' 2011 Computer Vision and Pattern Recognition conference, held June 21-23, 2011, in Colorado Springs, Colo.

Black-Tie Time

Two weeks earlier, before a black-tie-clad audience of hundreds at the Guildhall in London, Messrs. Blake, Fitzgibbon, Shotton, Sharp, and Cook had heard their names read as the MacRobert Award winners.

"It was an evening," Fitzgibbon relates, "where there was a lot of smiling."

That good humor was the result of lots of hard work—and effective teamwork—within the Cambridge team and with the Xbox product and incubation groups.

"It was a really good, constructive collaboration," Shotton says. "Mark Finocchio, the guy I was working with directly, is a very clever coder and did a fantastic job of taking what we were throwing at him and integrating it."

Matt Bronder, principal software engineer for IEB, also collaborated with the researchers. It didn't hurt that he and Finocchio had been aware of Sharp's work on decision trees on GPUs.

"I had stumbled upon some interesting image-classification work online from a researcher named Jamie Shotton," Finocchio recalls. "I found that he actually worked at Microsoft and contacted him immediately. He felt that there was a possibility of

something there and that it was worth pursuing. As I got to know Jamie, I learned that he is unique. He's an incredible researcher *and* developer. Because of this, he can see the academic side through a practical perspective. A quality like that is rare, and this company is lucky to have him."

That mutual respect is a common thread among those who worked on this project.

"For me,' Cook says, "it's been an amazing opportunity to work with so many fantastic, great people across the company. To see the research we do in a real product, having a real impact, is an amazing experience."

Sharp offers an additional perspective.

A Big Leap

"Technology progresses sometimes in big leaps and sometimes in small steps," he says. "With Kinect, particularly with the machine learning in Kinect, it's definitely one of the big leaps. It's great to have been a part of that story, which I think will stand out as a milestone long into the future."

That's the kind of excitement and wonder the project has brought to those who helped bring it to life—let alone to those thousands playing a Kinect game at this very moment.

For Blake, this is the culmination of a long process.

"People have been working on vision for decades," he observes, "since long before it was practical to build real-time vision machines, and the human-body-motion-tracking problem has been open for the last two decades or so. Now, Kinect is a prime example of a computer-vision system that has really impacted mainstream technology."

As for Shotton, who had been employed by Microsoft Research Cambridge a mere three months when he received Finocchio's fateful email, the project became something that he could socialize proudly—once the word was out.

"It was very exciting to work on something that was new and different, with world-class engineering that would change the world," Shotton says. "I couldn't share that excitement with any of my friends or family, because it was all top-secret. But eventually, as the press reports came out, I could say, 'Yeah, this is what I'm working on.'"

Cook knows the feeling.

"It was great to be working on something that you could talk about down the pub and people would have some idea what it did," he smiles. "The game thing where you have to jump around vigorously for three hours in order to play it—parents approve very much of this."

Share E-mail this Print Back to top