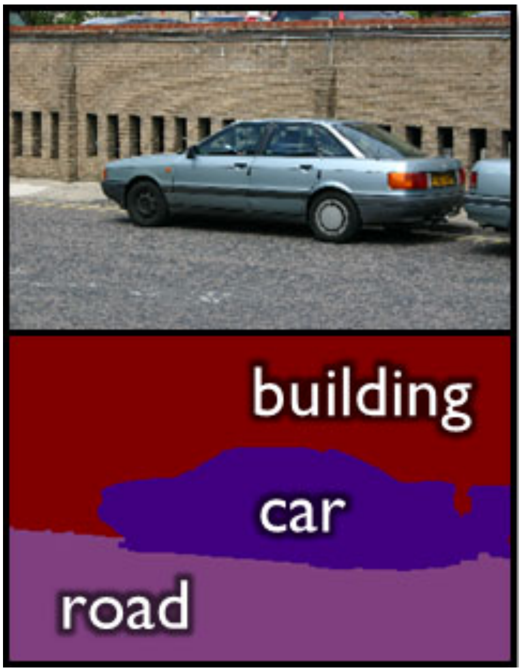
Alumnus Dr Jamie Shotton and the development of Kinect for Xbox 360

20 January 2011

Dr Jamie Shotton

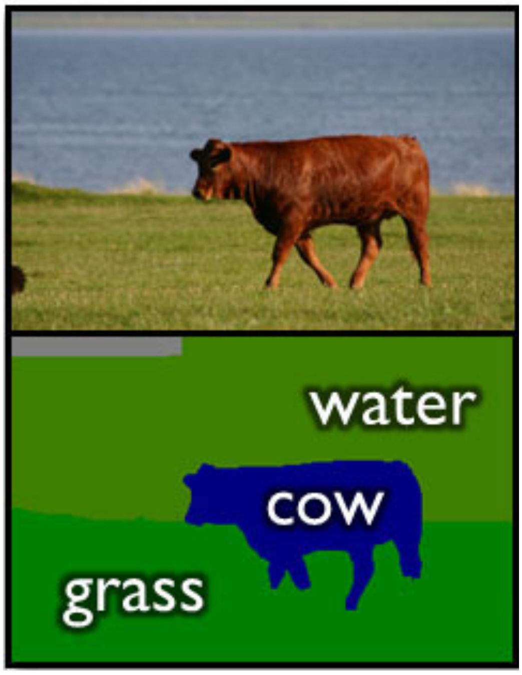
Dr Jamie Shotton completed his PhD in the Machine Intelligence Lab with Professor Cipolla, here at the Department of Engineering from 2003-2007. Jamie now works for Microsoft at their Cambridge research laboratory, where he has been intimately involved in the development of Kinect for Xbox 360. Kinect makes you the controller, allowing you to jump in and play games using your whole body, without holding or wearing anything special. Jamie came back to the Department to lecture the 4th year undergraduate students on this in November 2010. Below, he tells his behind-the-scenes story about Kinect.



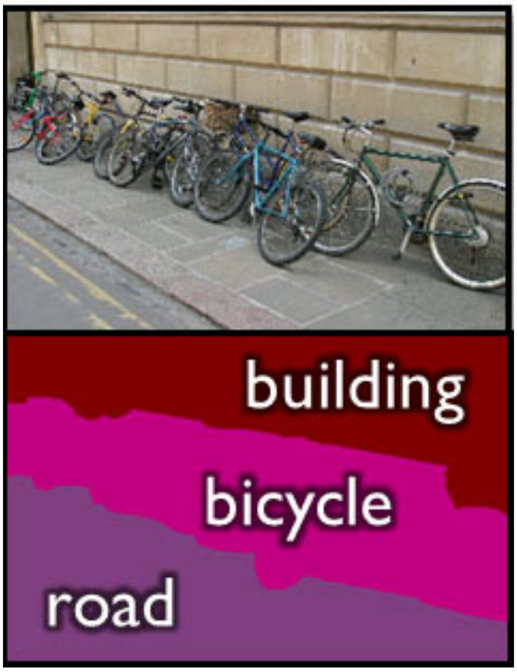
"I joined the Machine Learning & Perception group at Microsoft Research Cambridge (MSRC) in June 2008 as a post-doc to continue my PhD research in computer vision. In this, I had focused on automatic visual object recognition: teaching computers how to recognise different types of object in photographs such as cars, sheep, trees, etc. Little did I know at that point how quickly I would get pulled into the frenzy of research and development around Kinect, and how this blue-skies research could be applied to such a practical problem.

"I had taken a machine learning approach to visual object recognition in photos, which works as follows. First, you build up a varied training set of images where you label each pixel with a colour, according to which object category it belongs to. So, for example, you hand label all 'cow' pixels in blue, and all 'tree' pixels in green, using a simple painting application. Second, you give this training data to a machine learning algorithm that does some number crunching to automatically work out patterns of image appearance that correlate with the presence or absence of various object categories. The learning algorithm gives you a trained 'model' that efficiently encodes these correlations, and hopefully generalises to new unseen data. Finally therefore, you show the model a new image and it works out to which object category each pixel belongs.

"A couple of months into my post-doc at MSRC, I got a call out of the blue from the Xbox product group who, having come across some of my earlier publications from the Department of Engineering, wanted to discuss an 'important, top-secret scenario' with me. They described their goal, that of real-time robust human body tracking, and how it could be used for playing computer games. Now, this had been a dream of science fiction for many years, and still is a hugely active topic in the computer vision community - several of my fellow students in Professor Cipolla's group, including Dr Bjorn Stenger, had had this as their PhD topic. But it was always seen as being 'five years away' from being commercially viable. So of course I was rather sceptical anything could come of this, especially given Xbox's ambitious plan to launch by Christmas 2010.



"But then they mentioned the new depth-sensing camera hardware they were busy developing. I had seen depth cameras before but only at very low resolution (about 10x10 pixels). The new Kinect camera worked at 320x240 pixels and 30 frames per second, and the depth accuracy really got me excited - you could even make out the nose and eyes on your face. Having depth information really helps for human pose estimation by removing a few big problems. You no longer have to worry about what is in the background since it is just further away. The colour and texture of clothing, skin and hair are all normalised away. The size of the person is known, as the depth camera is calibrated in metres. Further, since the depth camera is 'active', shining out its own structured dot pattern of infra-red light into the room, the camera can work with the lights turned off.



"But even with depth cameras, it's not all plain sailing. There is still the whole gamut of human body shapes and sizes, and, worse, people can get themselves into an incredible variety of poses (body positions). Just think about how many positions you can put your right arm in, then multiply by the number of positions for your left arm, your right leg, and so on, and you rapidly end up with a combinatorial explosion.

"The Xbox group also came to us with a prototype human tracking algorithm they had developed. It worked by assuming it knew where you were and how fast you were moving at time t, estimating where you were going to be at time t + 1, and then refining this prediction by repeatedly comparing a computer graphics model of the human body at the prediction, to the actual observed depth image on the camera and making small adjustments. The results of this system were incredibly impressive: it could smoothly track your movements in real-time, but it had three limitations. First, you had to stand in a particular 'T'-pose so it could lock on to you initially. Second, if you moved too unpredictably, it would lose track, and as soon as that happened all bets were off until you returned to the T-pose. In practice this would typically happen every five or ten seconds. Third, it only worked well if you had a similar body size and shape as the programmer who had originally designed it. Unfortunately, these limitations were all show-stoppers for a possible product.

"And so our brief back at MSRC was to overcome these limitations somehow. I sat down with colleagues Dr Andrew Fitzgibbon and Professor Andrew Blake and we brainstormed about how we might solve the problem. A first observation was that when you look at a photo of a person, you can tell where their limbs are even though the person is not moving. If we could remove the temporal dependency, we would remove the need for the initial T-pose, and be able to recover if we lost track. Another thought was that to cope with the variations in human size and shape we should use machine learning, rather than try to somehow directly program for all possibilities by hand: instead, we would encode these possibilities in the training data.

"Having studied with the Department's Professor Cipolla, I knew about Dr Stenger's research which uses a technique called 'chamfer matching' to a whole image of the body against the training set of body images. By finding the closest match (the 'nearest neighbour') you can transfer the known 3D human pose from the training image to the test image. We tried this technique out, and had some success getting a coarse human pose out without using any temporal information. The problem was, however, that to get the level of detail we needed would have required so many 'exemplar' training images to cover all possible body shapes and sizes that the matching process could not run in real-time on the limited processing hardware we had available.

"So we went back to the whiteboard. What was now clear was that we had to divide up the body into parts and somehow match each part independently to avoid the combinatorial problems with matching a whole pose at once. I hit on the idea of revisiting my PhD work on object recognition, but this time instead of object categories, we were going to use body parts such as left hand or right ankle. We designed a pattern of 31 different body parts as you see colour-coded on the right here, and then trained an efficient decision tree classifier to predict the probability that a given pixel belongs to each part of the body. If you can accurately predict these part probabilities from a single depth image, regardless of body shape, size, or pose, then you get 3D proposals for the locations of many body joints at extremely low computational cost.

"This turned out to be the winning formula, but it still needed a lot of engineering to scale up to the level of accuracy we needed. The larger and more varied we could make the training set, the better it was likely to perform in your living room. So we turned to Hollywood, who have been building advanced computer graphics models of the human body for their movies for many years. We recorded hours of footage at a motion capture studio of several actors doing various moves that could be useful for gaming: dancing, running, fighting, driving, etc. This 'mo-cap' data was then used to automatically animate computer graphics models of different human shapes and sizes. We ended up with a vast training set of millions of synthetically generated depth images. Moreover, the graphics algorithm could easily render the corresponding body part images we needed for training as a texture map.

"The final piece of the puzzle was how to deal with these millions of training images. My previous work on recognition in photographs had taken a day or two to train from only a few hundred images, and using this approach directly on millions of images would have taken weeks if not months, prohibitive on our tight schedule. We enlisted the help of our colleagues at Microsoft Research in Silicon Valley who had been developing an engine called 'Dryad' for efficient and reliable distributed computation. Together, we built a distributed training algorithm that divided up the millions of training images into smaller batches and trained off each batch in parallel on a networked cluster of computers. Using a cluster of about 100 powerful machines, we were able to bring the training time down to under a day.

"All the pieces were in place now, and we worked with the Xbox team to put everything together. Our body part recognition algorithm gives fast and accurate proposals about the 3D locations of several body joints which are then taken and processed by the Xbox group's tracking algorithm to stitch the skeleton together (another bit of engineering magic!). This skeletal tracking, together with other new technologies such as voice recognition, give game designers the platform on which to build the magical experiences you get with games such as Kinect Sports and Dance Central.

"But of course, gaming is just the beginning, and I foresee this technology fuelling rapid advances in augmented reality and tele-presence, Internet and personalised shopping, and healthcare, to name just a few. We are even looking at how touch-free interaction could find its way into the operating theatre so that the surgeon can navigate the patient's data much more quickly and without risk of contamination from a mouse or keyboard."

If you are interested in finding out more, please contact Dr Jamie Shotton by email: jamie@shotton.org