Kris Rivera – 12/4/10

Explain how the 3D camera in the Microsoft Kinect sensor operates

The Microsoft Kinect is an accessory for the XBOX 360 console that turns the user's body into the controller. It is able to detect multiple bodies simultaneously and use their movements and voices as input. The hardware for the Kinect is comprised of a color VGA camera, a depth sensor, and a multi-array microphone. The VGA camera is used to determine different features of the user and space by detecting RGB colors. It is mainly used for facial recognition of the user. The multi-array microphone is a set of four microphones that are able to isolate the voices of multiple users from the ambient noises in the room, therefore allowing users to be a few feet away from the device but still be able to use the voice controls. The third component of the hardware, the depth sensor (generally referred to as the 3D camera), has two parts to it: an infrared projector and a CMOS (complimentary metal-oxide semiconductor) sensor. The infrared projector casts out a myriad of infrared dots that the CMOS sensor is able to "see" regardless of the lighting in the room. This is, therefore, the most important portion of the Kinect which allows it to function. But there is a second component that would render the Kinect quite useless otherwise: the software that interprets the inputs from the hardware. These two components will be the main focus of the paper.

User videos posted on Youtube show the Kinect's large array of scattered infrared dots literally painting the user's living room in a swathe of green lights. The rays are cast out via the infrared projector in a pseudo-random array across a large area. The CMOS sensor is able to then read the depth of all of the pixels at 30 frames per second. It is able to do this because it is an active pixel sensor (APS), which is comprised of a two-dimensional array of pixel sensors. Each pixel sensor has a photo detector and an active amplifier. This camera is used to detect the location of the infrared dots. Following this, depth calculations are performed in the scene using a method called Stereo Triangulation. Stereo Triangulation requires two cameras to be able to perform this calculation. The depth measurement requires that corresponding points in one image need to be found in the second image. Once those corresponding points are found, we can then find the disparity (the number of pixels between a point in the right image and the corresponding point in the left image) between the two images. If the images are rectified (along the same parallel axis), then, once we have the disparity, we can then use triangulation to calculate the depth of that point in the scene.

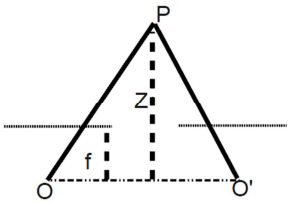

Figure 1.

$$Z = \frac{f(O' - O)}{(p'_x - p_x)}$$

Figure 2.

$$d = p_x - p'_x$$

Figure 3.

This Stereoscopic Triangulation requires two cameras, but the Kinect is unique in that the depth sensor only has one camera to perform these calculations. This is because the infrared projector is, in and of

itself, a “camera” in the sense that it has an image to compare with the image taken from CMOS sensor camera. The projected speckles are semi-random in the fact that they are a generated pattern that the Kinect knows where they can be found. Since the device knows where the speckles are located, it has an image which can be compared to find the focal points. The CMOS sensor captures an offset image to detect differences in the scene where the disparity between dots can be analyzed and the depth can therefore be calculated. I am assuming that the images are rectified, making it simple to calculate the depth with the equation in Figure 2.

Now that we have our depth calculations, we have all of this data that needs to be interpreted and used in the system. This is where the software, loaded onto the Kinect, comes into play. The software goes through a series of steps to make sense of the input from the camera and have the user’s body be the controller. Before those steps can be described, a separate but completely related process needs to be explained: teaching the computer how to understand what it is seeing. Teams of Microsoft employees went to homes all around the world and set up special rigs that captured the participant’s motions in everyday life. The difficult part of this was that a computer cannot understand anything of the human form just from images alone; they need to be labeled. So, Microsoft programmers had to analyze each frame and label the different important components of the people in the frame. Once this information was completed, the images and labels were fed into a learning algorithm so that it could create probabilities and statistics about the human form and movement. This information is loaded onto the Kinect for later use.

Figure 4.

Beginning with the user stepping in front of the Kinect, a 3-D surface is generated using the aforementioned means of retrieving the depth at each different infrared dot, creating a point cloud of the user. From this, the Kinect creates a rudimentary guess at the user’s skeleton is made (Figure 4. Left image). Next, using the information gathered and the probabilities of human form, the Kinect makes a guess at which parts of the user’s body are which. A level of confidence is also assigned to each guess based on how confident the algorithm is about guessing the correct parts (Figure 4. Bottom). Once this is done, the Kinect finds the most probable skeleton that would fit these body parts and their confidence levels assigned to them. Next, it creates a 3D Avatar (Figure 4. Right), which is skinned and clothed on an in-game avatar displayed on the screen. This is performed at 30 frames per second over and over again. All of these steps put together are what allows the user to become the controller and what makes the Kinect so powerful.