Chapter 11

Hardware-Supported Instructor Extraction

Although the previous chapters present a pretty robust and usable approach, image segmentation heuristics are always at risk of failing in certain situations. Since the logic behind human vision is not yet understood, the only alternative is provided by spezialized hardware. Although more expensive in acquistion, range-sensing devices might provide a fail-safe alternative for E-Chalk's lecturer segmentation task. 3D laser scanners use triangulation, which is computationally expensive and not real time capable. Stereo cameras obey the same rules as texture-based segmentation approaches. A rather new kind of device is called *time-of-flight 3D camera*. They promise to make real-time segmentation of objects easier, avoiding the practical issues resulting from 3D imaging techniques based on triangulation or interferometry. This chapter presents investigations on using a time-of-flight 3D camera for extracting the instructor acting in front of an electronic chalkboard.

11.1 The Time-of-Flight Principle

Time-of-flight 3D cameras are now becoming available (see for example the offers by 3DV Systems, Inc. [1], Canesta, Inc. [11], CSEM [13], or PMD Technologies, Inc. [39]). I tested a miniature camera called "SwissRanger SR-2" [13] built by the Swiss company CSEM and a prototype camera called "Observer 1K" built by the German company PMD Technologies [39].

A schematic view of the design of time-of-flight 3D cameras is shown in Figure 11.1. A time-of-flight camera works very similar to radar. The camera consists of an amplitude-modulated infrared light source and a sensor field that measures the intensity of backscattered infrared light. The infrared source is constantly emmitting light that varies sinusoidal. Object A reflects almost the maximum intensity while object B, having a greater distance to the camera, reflects less light. This is because at any specific moment, objects that have different camera distances are reached by different parts of the sinus wave. As shown in Figure 11.1, the incoming light is then compared to the sinusoidal reference signal which triggers the outgoing infrared light. The phase shift of the outgoing versus the incoming sinus wave is then proportional to the time of

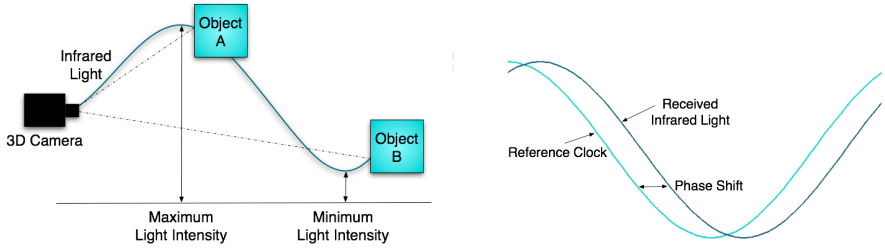

Figure 11.1: Left image: Two objects reflect amplitude-modulated infrared light. Object A reflects more light than object B because at the point of time when the photons hit object A, they were emitted with maximum light intensity. The photons that hit object B at the same time where emitted before that, with lower intensity. Right image: The actual distance can be calculated by measuring the phase shift between the emmitted and the reflected light. If the distance of the reflecting object were zero, the two curves would have no phase shift. The farther the object away, the greater the phase shift.

flight of the light reflected by a distant object. This means, by measuring the intensity of the incoming light, the phase-shift can be calculated and the cameras are able to determine the distance of a remote object that reflects infrared light. The output of the cameras consists of depth images and a conventional low-resolution gray-scale video, as a byproduct. A detailed description of the time-of-flight principle can be found in [Luan et al., 2001, Oggier et al., 2004, Göktürk et al., 2004].

The depth resolution depends on the modulation frequency. For the experiments a frequency of 20 MHz was used which gives a depth range between 0.5 m and 7.5 m, with a theoretical accuracy of about 1 cm. Usually, the cameras allow to configure frame rate, integration time, and a user-defined region of interest (ROI) by writing to internal registers of the camera. The cameras then calculate the distances in an internal processor. The resolution of the SwissRanger camera is 160×124 non-square pixels. The resolution of the Observer 1K is 64×16 non-square pixels. Both cameras behaved similarly in my experiments, but its low resolution made the Observer 1K unusable for instructor segmentation purposes. The following descriptions therefore refer to the SwissRanger camera.

11.2 Setup

The setup is essentially the one described in Section 9.3. The SwissRanger 3D camera is mounted on top of a video camera, and both cameras capture data synchronously. Figure 11.2 shows the 3D camera as well as the camera stand for two cameras.

In the experiments, the achieved frame rate varied around eleven frames per second. When an object is too close to the camera lense, overflows occur due to the large amount of light that is reflected. When an object is too far away from the lense, the distance measurement becomes imprecise. Ideally, the camera should be located at the end of the room to minimize the optical disparity between the video and the 3D camera without requiring a mirror setup.

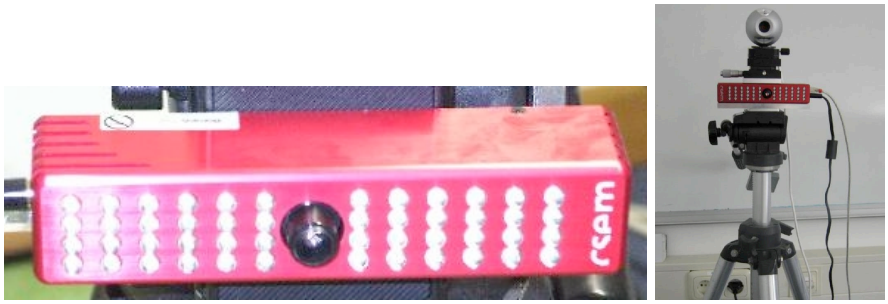

Figure 11.2: Left image: The SwissRanger time-of-flight 3D camera by CSEM. One can see the infrared light emmiting diodes on both sides of the receiving lense. Right image: The camera stand that allows a 2D camera and the Swissranger to capture the same scene.

Due to the restricted range of 7.5 m, the disparity is noticeable. However, a practical range for segmentation is between 2 m and 4 m in front of the electronic chalkboard. This range is optimal in terms of minimal visual noise and low overflow probability.

11.3 Technical Issues

Theoretically, the entire segmentation problem is reduced to a simple depth-range check. In the image captured by the 2D camera, only those pixels are considered to belong to the foreground that correspond to pixels with 3D camera depth coordinates smaller than the distance of the electronic chalkboard. In practice, however, several issues have to be solved first. They are currently investigated by Neven Santrac as part of a diploma thesis. A summary of his current results can also be found in [Santrac et al., 2006].

Camera Synchronisation

Unfortunately, the tested time-of-flight cameras do not support hardware-triggered synchronisation. It is therefore impossible to guarantee that the 2D camera and the 3D camera capture the frames at exactly the same time. While this is a technical issue and may be solved by the manufacturers in future, it is hard to match the images of the two devices in software. Measuring the time-difference between the two cameras manually is cumbersome, especially when the frame rates of the two cameras differ.

Resolution Inequality

The two cameras have different resolutions. In order to provide for a smooth viewing experience, the resolution for the lecturer extraction video should be at least 640×480. The maximum resolution of the SwissRanger SR-2 is 160×124. The resolution of the camera can be raised using different approaches. Cüneyt Göktekin and Frank Darius have found out[1] that the resolution of the 3D camera

[1] Unfortunately, they have not published the results of their experiments at the time of writing this text.

Figure 11.3: Two examples of raw depth-range segmentation (i. e., using only a camera calibration and depth-range treshold). Noise, motion blur, and camera desynchronization are the three most disturbing factors.

could be increased by *pixel shifting*, a technique that is commonly used for 2D photo cameras and recently also for video cameras [Ben-Ezra et al., 2005]. The resolution is increased for each dimension at the cost of a lower frame rate by combining several pictures from slightly shifted camera positions. Pixel shifting, however, requires additional hardware to shift the sensor device for each frame by a few micrometers. In [Diebel and Thrun, 2005] Markov Radom Fields are used to increase the resolution. They, too, use a combination of a 2D Camera and a 3D camera, exploiting the fact that discontinuities in range (i. e., in the 3D image) and coloring (i. e., in the 2D image) tend to co-align. Experiments with their approach indicated that it works reasonably well. Figure 11.4 shows an example. However, the method is computationally expensive and far from real-time performance. A downside is that the method sometimes blurres the edges between instructor and chalkboard. Sometimes the influence of the 2D camera image gets too strong and reflections from the board surface appear in the final segmentation result.

Lense distortion

As the cameras are positioned close to the board in order to avoid noise and overflows, the cameras' radial distortions have to be eliminated. Having fixed both cameras in the right position, lense distortion can be elimanated using well-known calibration methods, for example [Tsai, 1987]. However, the calibration has to be repeated whenever the position of a camera changes. This makes the solution very impractical for a mobile setup.

Light Scattering and Motion Blur

Objects reflect light in different angles and in different intensities. The properties of light reflection depend on the form and on the material of a given object. In the result, the depth measurement is not texture and material-independent. Since darker objects reflect less light, the output of the camera is noisier than in the measurement of brighter objects. Light scattering is a prelimiary cause for edge bluring. 3D time-of-flight cameras tend to blur motions even more than regular 2D cameras. Motion interferes with the measurement of the reflected light. A quickly moving arm, for example, reflects light from two positions in one frame. Motion blur is particularily observed at the borders of moving objects.

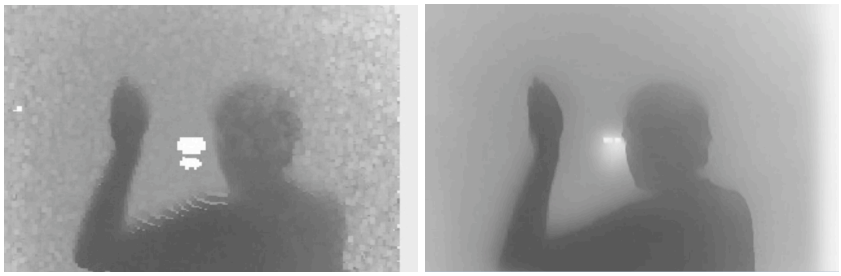

Figure 11.4: Left: Depth image obtained by the camera. Right: Depth image enhanced using the method proposed by [Diebel and Thrun, 2005]. The white spot in the center is an artifact produced by reflection.

Noise

Noise is one of the biggest issues. The signal-to-noise ratio shrinks radially from the center to the corners. The highest signal-to-noise ratio is found in the center of the camera view and the smallest in the corners. The measurement error induced by noise is up to 4 % (this means 30 cm in a total scale of 7.50 cm). When an electronic chalkboard with back-projection is used, noise becomes even worse because residual light of the electronic chalkboard is caught by the lense, which interferes with the measurement of the reflected infrared light. The hands of the instructor can usually not be distinguished from the background by the camera since they are too close to the board surface and moving too fast. Figure 11.3 shows two sample frames with typical noise distortion and other issues discussed here.

11.4 Segmentation Approach

While the above-mentioned problems make a direct segmentation based on the output of the camera virtually impossible, an advantage of the time-of-flight principle is that the method does not make use of motion or texture-based segmentation techniques. It is therefore possible to use the 3D camera in combination with complementary texture-based methods. For this reason, a combination of the output of the 3D camera with the method presented in Chapter 9 provides a working solution. The idea is to find a subset of the instructor as well as a subset of the background using the depth information provided by the camera and then to use the color signature-based segmentation approach.

The depth image is mapped onto the 2D image by using a grid calibration. The resoltion of the depth image is increased using pixel duplication. In order to get a subset of the background, a connected component search is performed on the depth picture of the camera. The biggest connected set of pixels that is in a predefined depth range is considered to be a mixture between background and instructor. A gracefully chosen bounding box of the corresponding area in the 2D image is considered to be a superset of the instructor. The regions outside the bounding box are considered background. This heuristics sometimes fails for the hands of the instructor when these are close to the corners of the image where the camera's signal-to-noise ratio is very small. In most cases, however, the method gives good results. A small subset of the instructor is chosen with

Figure 11.5: Top left: The depth data of the 3D camera is used to construct both a superset and a subset of the foreground. Other images: A few sample frames showing the segmentation result using the color segmentation approach described in the previous chapter. The instructor is robustly segmented while working with any content on the electronic chalkboard even if he or she is moving quickly. Results taken from [Santrac et al., 2006].

the following strategy. The biggest connected component is shrunk radialy from the edges until the variance of the corresponding depth information in the area is below a threshold. This way, an input is generated that is similar to the input described in Section 10.6. Figure 11.5 shows a sample input image with the constructed bounding box and the subset of the instructor along with a few sample results of the following color segmentation. The color signatures are regularily updated in order to adapt to changing background and foreground.

The approach works in real time with 25 frames per second on a video with 640×480 pixels and is rather resistant against blurred edges and desynchronisation effects. Only a subset of the noisy 3D camera image is used. The fine details are segmented by the color segmentation.

11.5 Conclusion

3D time-of-flight cameras initially promise an efficient way to solve the instructor segmentation problem. In practice however, the exact calibration and synchronization of the two cameras is tricky. The 3D cameras do not yet provide any explicit synchronization capability, such as those provided by many FireWire cameras. The low x and y-resolution of the 3D camera results in coarse edges. The z-resolution is just not high enough, since the instructor usually stands very close to the board and the range of interest usually is about 50 cm. The low signal-to-noise ratio does not allow for direct segmentation. Adequate segmentation is only possible in combination with other techniques. Besides overflows,

there are other artifacts caused by quickly moving objects, light scattering, background illumination, or the non-linearity of the measurement. Last but not least, using a time-of-flight camera requires a large budget.

For instructor segmentation, the ideal time-of-flight camera should offer a higher depth range (for example 15 m) and a z-axis resolution of a few millimeters. The image resolution should be at least PAL. It would be ideal if a color video chip could be combined with the depth-measurement chip in a single unit. Given such a camera at a low price, the computational costs of image segmentation could be dramatically reduced. Because a time-of-flight 3D camera captures depth and intensity information at acceptable frame rates and rather texture-independent, the segmentation problem is theoretically reduced to camera calibration and a simple depth-range check. At the time being, the software segmentation approach presented in Chapter 9 and its generalization in Chapter 10 proved to be an adequate tool for combination with 3D time-of-flight cameras in order to facilitate segmentation.