Chapter 12

Lecture 12: Homogeneous coordinates, calibration, stereo

Monday, March 14, 2011 MIT EECS course 6.869, Bill Freeman and Antonio Torralba
Slide numbers refer to the file "lecture12CalibrationStereo.key"

slides 1-30, see notes for slides 77-100 of previous lecture

slides 31 Now we've calibrated the camera, so we know how to relate points in an image with points in the world.

One application is to recover depth from stereo images. Of course, stereo is an important visual cue for shape. Many animals have stereo vision.

slides 32 We have many visual cues for depth. Here's another one: familiar objects. Here's a face mask, and there are many cues that tell us what the shape is–stereo, familiar shape, motion, etc.

What happens in we pit these cues against each other? We can do that with this demo. For the hollow mask illusion, familiar shape wins over stereo and motion.

slides 33

It's possible to separate out these various cues. A special stimulus, called a random dot stereogram, creates an impression of shape without any familiar shapes. They were developed by Bela Julez at Bell Labs, in one of the early applications of computer graphics to vision science. If you look at either eye's view by itself, you just see random dots with no form. But when viewed one by each eye, the disparity of some of the dots lets us see a raised square (if that is coded into the dot disparities).

slides 34 In general, of course, we have many visual cues in our stereo images, and we can make photographs look 3-dimensional by showing spatially offset views to the left and right eyes. Shown here are some methods to show those to our eyes.

slides 35, 36 In a kind of self-referential form, here is an early 3-d display of early 3-d displays. The red-blue anaglyph lets you present one image to the right eye and another to the left.

This is what you'll do with your pinhole camera boxes, by the way.

slide 37

People have become very creative with ways to present 3d information to people, by the way. This is an autostereo gram, which requires no special equipment to perceive a 3d form. But it requires some special viewing gymnastics–you have to relax your eyes a bit until they cross.

slide 38 For stereo, we look at the parallax induced by viewpoint change, and use that to infer depth. To understand the relationship between that parallax change and the objects' depth, we need to understand the calibration of the stereo cameras. To measure the parallax change itself, we need to automatically find point correspondences across images. Let's examine a number of the steps in those tasks.

slides 39-43 First, let's look at the main idea of stereo, how to measure depth from disparity.

Assume the optical axes of the two cameras are parallel. We see a point p in both cameras, at position x_l in the left camera, and position x_r in the right camera. The focal length of the cameras is f, and their optical centers are O_l, O_r. How do we infer depth from their relative positions?

By simliar triangles (triangle p, p_l, p_r is similar to triangle p, O_l, O_r), the ratio of their heights to their bases are equal. Re-arranging terms then gives

$$Z = f \frac{T}{x_r - x_l}. \tag{12.1}$$

That's the main idea: parallax tells depth. Now the remaining problems are: where to look for point matches across images, how to find the matches, and how to infer probable depth images. We'll finish up with depth estimates using structured light.

slides 44-54 Ok, so you have the tools you need to figure this out: suppose you see a point in one image, what is the locus of points that it could appear in in the other image?

A point from one location in one camera, going out into the world, forms a ray of possible 3-space locations that the point could have originated from. A line in 3-space renders as what, under perspective projection? As a line in the image plane of the second camera. That line is called an epipolar line for that feature point and that camera pair.

Likewise, the image of the 3-d point in the 2nd camera gives rise to a epipolar line in the first camera. Those lines, the rays from the camera, and the line connecting the two centers of projection of each camera all lie in a plane, called the epipolar plane.

slides 55-66

To calculate where those epipolar lines are, we want to look at the associated epipolar algebra. We've calibrated both cameras, so we know their relationship to the world coordinates, and thus to each other. Let one be related to the other by a rotation, R, and a translation, T.

By taking the appropriate cross products (see slides for equations), we can find a matrix, E, that satisfies a relation,

$$X'^T E X = 0, \tag{12.2}$$

where X' is the (non-homogeneous) 3-d coordinate of a point in the image plane of the primed camera system, and X is the corresponding image plane point in the other camera. The matrix, E, relating those two is called the essential matrix.

slides 67-69

Once the camera calibration is known, it's often convenient to "rectify" the images–to reproject them onto a pair of cameras oriented in a common direction, perpendicular to the line between the camera centers. This allows for the epipolar lines to be oriented along scanlines. See p. 56 of Szeliski book, in Sect. 2.1.5, for more details on this procedure.

slides 70-74 Now with a calibrated, rectified stereo rig, we are back to the problem we discussed at the beginning of the stereo section: how to find the matches for points between a pair of images?

Very naively, you can match pixel intensities between the two images. But that's not a very unique identifier for a good match. Better is to match image patches. We can match by sum-of-squared differences (SSD), or by a richer set of matching criteria.

slides 74-78

People have put a lot of effort into the local evidence framework. Listed are various of the steps that are taken.

slides 79

To get even better results, people resort to spatial priors on depth for images.

slides 80

In the energy minimization formulation of the regularization problem, we have a shape-dependent energy term that gets added-in to help determing the best solution.

slides 81-85

If we just look along a scan-line then optimal depth estimation is just a 1-d problem, and it's straightforward to find the globally optimal set of disparities, using dynamic programming methods.

slides 86-95

For the more general energy function/Markov random field, 1-d solutions don't suffice. So we can solve them using graph cuts, or in the probabilistic framework, using any of the methods we talked about previously.

slides 96-97

A real contribution of Szeliski to stereo (and computer vision in general) has been to maintain a controlled dataset and scoreboard rating algorithms. (note how many BP's are in the leaders)

slides 98-106

A related method for infering depth is, instead of using a second camera to find point correspondences, to project structured light from an offset projector position. Once unique projected positions have been found, the math is exactly as it was for the stereo camera case. That's how the Microsoft Kinect works.