CS 280 Computer Vision
Spring 2018 Efros, Malik, Yu Lecture 15

1 Lecture

The focus of the next few lectures will be 3D computer vision – i.e. how we perceive the world in 3D.

Binocular Stereopsis

Consider an object that is being viewed from two different locations (e.g. a left and a right eye). The left and right images will depict slightly different views, and allow us to calculate **disparity** (the difference between the object location in the left image and the object location in the right image). Disparity is a signal for depth, so we can use it for 3D estimation.

Camera Configurations

We can arrange the cameras in multiple ways. Maybe there is a single point of fixation where the optical axes intersect. This is very common in biological vision: we want an object of interest to be in the center of each eye's field of view, since this is the area of highest resolution, and therefore the optical axes of each eye (which extend from the centers) will intersect at the object which we are fixating at.

As another configuration, the two cameras might be looking at infinity (with the two optical axes parallel to each other). This is common for robots and situations where we don't want to move the two eyes.

Finally, in the general case there is no direct relationship between the two cameras. They can be looking in arbitrary directions relative to each other. This is common too, and arises for example in cases where we're reconstructing a scene based on multiple photographs.

I. Fixating Binocular System

Let's calculate the disparity of the fixation point for the first case, where the left and right eyes fixate at the same point. We need to look at the image in each eye and see if they're different. By definition, the fixation will be at $(0, 0)$ in each eye, so the fixation has no disparity. But other points might have disparity, and this might give us some signal.

Take a general scene point Q. What is its disparity? There's going to be a coordinate system for the left eye, with axes X_L, Y_L and origin at the center of projection of the optical axis. Likewise, there will be a coordinate system X_R, Y_R for the right eye.

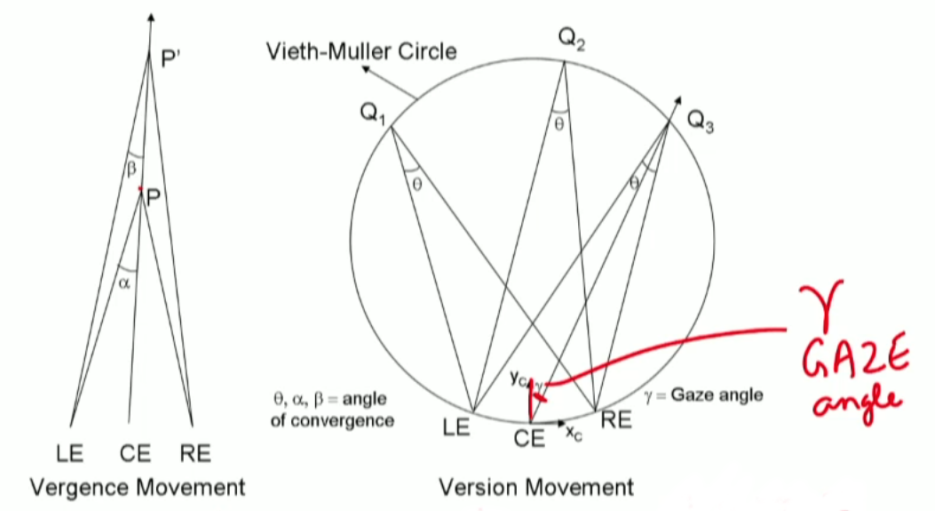
Meanwhile, we'll use the cyclopean coordinate system X_C, Y_C for points in the world. This is based on the cyclopean eye – an imaginary eye at the midpoint of the left eye and the right eye. To recap: scene points will be represented in the cyclopean coordinate system. Projected points will be represented in the left and right coordinate systems.

The left and right eyes can rotate in order to look at different objects. **Vergence movement** describes a situation where we're looking at a near point or a far point in a certain direction. We call the near point P and the far point P'. Then the half angle for P (between the left eye and the cyclopean eye) is α and the half angle for P' is β.

Vergence movements are characterized by α and β. α is large for near fixation and small for far fixation.

Then we have **version movements**, where we change the direction of gaze (where is the eye looking?). We can change our gaze direction γ while keeping the angle of vergence constant.

So the two types of eye movements are *pure vergence movements* (keep looking in the same direction, but near or far) and *pure version movements* (change the eye direction while keeping the angle of vergence constant).

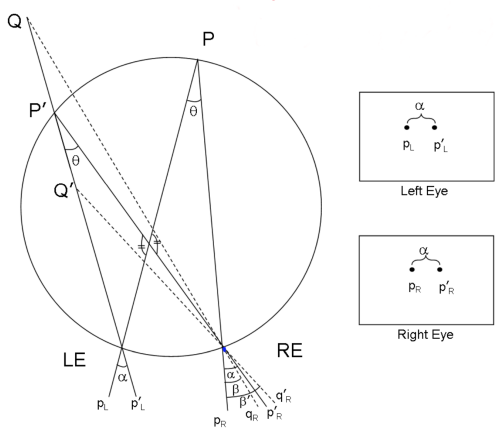


source: slides from Jitendra Malik's CS 280 lecture

Say we have points Q_1, Q_2, Q_3. These are three different fixation points, all of which correspond to the same angle of vergence θ. It can be proved that all three points must then lie on a circle passing through the left eye and the right eye. This circle is called the Vieth-Muller circle.

Disparity is zero on the Vieth-Muller circle. Recall: we have a point of fixation P and we're interested in the disparity of point Q. This can be calculated as $q_R - q_L$, i.e. the difference between the coordinates in the left eye and the coordinates in the right eye. It can be measured in length (pixels) or angle (degrees).

We now consider a general point P' which is also on the circle. We need to figure out where it projects in the left eye and the right eye. We will measure its location by measuring an angle. Now: there is p_L, i.e. the center (the fixation point), and p'_L, the projection of P' in the left eye. Likewise, we have p_R and p'_R. We want to measure α and β (the angles that P' makes with respect to the optical axis in each eye). Then the disparity is $\beta - \alpha$.

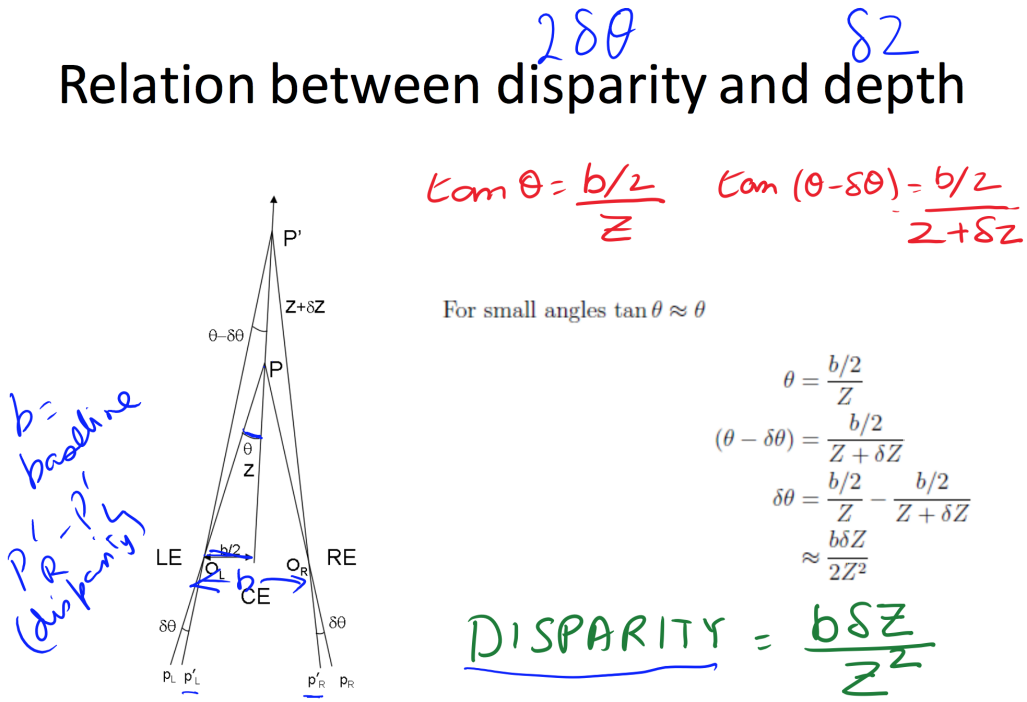


source: slides from Jitendra Malik's CS 280 lecture

Via geometry, we compute $\beta - \alpha$ as 0. Therefore, disparity is 0 if P' lies on the Vieth-Muller circle.

(Consider the point Q lying outside the Vieth-Muller circle, and the point Q' lying inside the Vieth-Muller circle. If we do the same geometric calculations, we will get positive disparity in one case, and negative disparity in the other.)

What if there *is* disparity? Again, we have a point P on the circle. Take a point P' in the same direction (as the ray from the cyclopean eye to P), except further. Also define b, the baseline, as the distance between the left eye and the right eye, and θ, $\delta\theta$, Z, and δZ according to the following figure. By trigonometry, the disparity of P' is $b(\delta Z)/Z^2$.



source: slides from Jitendra Malik's CS 280 lecture

So the disparity allows us to measure the difference in depth: δZ is how far past P P' is. If we measure the disparity in our camera, we can use this quantity to get δZ, and then add that to Z to get the depth.

II. Parallel Optical Axes (Fixation at Infinity)

Again, we'll have a coordinate system at the pinhole of each eye's camera, and a cyclopian eye at the midpoint of the two pinholes. Then there will be some scene point (X, Y, Z) which we view using cyclopian eye coordinates. The point will have a projection x_L in the left eye and a projection x_R in the right eye. We call f the Z-distance from the pinhole to the image plane, while b is again the distance between the two pinholes. At the end of it all, the disparity will be $x_R - x_L$.

If we work through it, the disparity comes out to be $x_R - x_L = \frac{-bf}{2}$, or $x_L - x_R = \frac{bf}{2}$. The disparity increases linearly with baseline; if the two eyes are further apart, we'll get bigger disparity. This makes sense, because if the two eyes were on top of each other ($b = 0$), we ought to get no signal. As we do.

The disparity also increases linearly with f. (Any distances we measure on the image plane are magnified by f. If we put the image plane twice as far away, all the x-coordinates will be doubled.)

Finally, disparity increases as $1/Z$ ("nearness," since Z is depth). As $Z \to \infty$, disparity goes to zero.

Example: range sensors. The Kinect depth camera employs **active stereopsis**: it replaces one of the cameras with a projector, and then images the projector's emitted pattern of light using the other camera. The geometry is still that of stereopsis, however, so we simply label it as "active" stereopsis – as opposed to "passive" stereopsis, where there are two cameras. The passive counterpart doesn't emit light itself; it simply waits for light in the environment to come to it. The mathematics of disparity in both cases are exactly the same; we'll get the same bf/Z formula.

There are two types of technologies for measuring depth in an active way.

- One technology is that of the Kinect, and is based on triangulation. You have either two cameras or one camera and one projector, and the disparity follows the formula bf/Z. *Note: the Kinect is best used indoors, because the pattern of infrared light it sends out will get swamped by all the sunlight outside and will not be visible.*
- The other technology is **time-of-flight**. Here, we send out a light signal and measure how long it takes to bounce and come back. If it comes back in one nanosecond, we thus know – based on the speed of light – that the point is x feet away. If it takes longer, the point is further away. And so on.

Disparity is proportional to 1/depth, which means that depth is proportional to 1/disparity. If we can measure disparity to a certain accuracy, how accurately can we measure depth? Using calculus, we find that error in the distance estimate increases quadratically with the distance, i.e. error $\propto$ distance^2.

$$Z = \frac{c}{d}$$
$$\frac{\delta Z}{\delta d} = -\frac{c}{d^2} = -\frac{1}{c}\frac{c^2}{d^2} = -\frac{1}{c}Z^2$$

Thus $\delta Z = k_1 Z^2 \delta d$, where δZ is the measurement error in the depth (the distance), k_1 is some constant, and δd is the measurement error in the disparity. We see that δZ is proportional to Z^2. So we expect any measurement device based on stereo or triangulation to have an error that grows as Z^2. Basically, stereo is a sense that becomes useless for objects which are far away. Conversely, stereo is very good for objects that are nearby. *Stereo works in the space around us. We exhibit fine motor control for tasks in the near field.*

If we want to measure depth accurately with stereo, we should have as big of a baseline as possible.

III. Arbitrary Camera Configuration

Now, there are two cameras with some arbitrary rigid body motion (a rotation R and translation t) between them. Say we have a point M in the world which is projected onto both the left image and the right image. Based on the measurements we make in each camera, we want to figure out how far away M is.

Let's say we don't know the rotation and translation between the cameras, but we do have n point correspondences between the two images. Our goal might be to estimate R, t, and depths for each of the points...