Video: Stereo Vision Overview

[00:01] [slide 1] How does the human visual system construct a three-dimensional representation of the world from the two-dimensional images projected onto the left and right eyes, and how can a computer vision system take two views of a scene from a stereo camera and compute a depth map that captures the three-dimensional distance to surfaces in a scene? This video describes the geometry of projecting a three-dimensional scene onto the eyes or cameras, and it introduces the concept of stereo disparity and how this quantity can be used to compute depth, and provides an overview of the stereo process. The next video elaborates on the most challenging part of this process, which is known as the stereo correspondence problem.

[00:49] [slide 2] The first question you might ask is, why do we even need more than one view of a scene, to infer its three-dimensional structure? These two images show in a humorous way, that depth is inherently ambiguous from a single view. If you didn't know better, you might think the man on the left is really pushing on a tiny leaning tower of Pisa, or that these farmers really grew this gargantuan pumpkin. An object that we see in a single image could be at any distance from the eye, with an actual size that's proportional to its distance. [slide 3] The creation of stereo photographs and devices for viewing them, called stereoscopes, started with Charles Wheatstone in 1838. He invented the stereoscope shown in the upper right. Below it is a stereogram of Wheatstone himself. Stereoscopes like the one shown in the middle, and creation of stereo cards like those on the left, were a big industry and a popular source of entertainment in the late 1800s and into the next century. These days, magic-eye stereograms are the latest craze in the world of stereo entertainment.

[02:06] [slide 4] What is stereo disparity and how does it convey information about depth? Consider the eclectic scene at the top. Objects in the scene are at different depths, and you can sense that to some extent in a single image, from cues like the way objects in front occlude parts of objects behind them. But the left and right eyes provide different perspectives on the scene that cause objects at different depths to project to different locations in the two eyes, and we can use the relative position of objects in the two images to infer depth. On the left, I also superimposed the two images and I'm going to flip back and forth between the two. The movement that you see here is due to the changing position of the objects in the left and right views, and this relative movement also gives a stronger sense of depth than you see in a single image. That sense of three dimensions would be even stronger if you could view these two images stereoscopically, with the left eye looking at the left image and the right eye looking at the right image. The two cameras are focused on the Joy of Trivia book in the middle of the image, and this book doesn't move very much between the two views, but notice that objects in the back shift in one direction, while objects in the foreground shift in the opposite direction. I'll let you just experience this for a moment. Why does that occur? Before getting into the geometry of stereo projection, just a last bit of entertainment - people are taking famous paintings like Starry Night here, and artificially constructing multiple views that give you an impression of the three-dimensional structure of the scene, like this.

[04:15] [slide 5] So, let's get serious. On the left here is a bird's eye view of the left and right eyes, viewing three objects at different depths. Imagine that the eyes are focused on a point on the front of this blue circle here. On the front of each eye, the black oval depicts the lens and the direction the eye is looking. The black dot inside the eye is the focal point, or center of projection. Any point on a visible surface is projected through the focal point, onto the back of the eye, where it appears in the retinal image. The point where the eye is focused is projected to the center of the field of view. On the right, we're taking the curved retinal image at the back of the eye, and laying it out flat. In each image, the point on the blue circle that projects to the center of each eye, is portrayed as the blue bar here that appears in the center of the left and right images.

[05:25] While the eyes stay focused on the blue circle, let's now consider where this red triangle projects onto the two eyes. We draw a line from a point on the triangle through the focal point of each eye onto the back, and a line through that focal point to the right eye here. Then we can eyeball where this point appears in our flattened left and right images. Note that its position in the right eye is shifted to the right relative to its position in the left eye. Now let's consider a point on the green square at the back. We again draw lines from that point through the focal points of each eye, onto the back of the eyes, and then recreate where that appears in the left and right flattened images over here. Note that in this case, the green bar in the left image is shifted to the left in the right image. We define this shift in position as stereo disparity. We'll refer to a shift to the right in the right image as positive stereo disparity, and we get a positive disparity in position for any feature that's closer to us than the depth where the eyes are focused, like in the red triangle here. A shift to the left is defined as negative stereo disparity, and we get a negative disparity like this for any feature that's further away in depth relative to the distance where our eyes are focused, like the green square here. A feature that appears at the same location in the two eyes, like the blue points here, is said to have zero disparity.

[07:36] [slide 6] A critical point I'd like to stress here is that this stereo disparity that arises when we project a three-dimensional scene onto the eyes depends on where the eyes are focused in space. This point is reinforced in this next example. We're again looking at a bird's eye view of the two eyes, with the same three objects out in space, but now the eyes are turned to focus on the red triangle, closer to the eyes. The point of focus on the red triangle now projects to the center of each eye, so the red bar here now both appears in the middle of the flattened images on the right. Let's repeat the same exercise as before, to see where the blue circle and green square appear in the images. In both eyes, a point on the blue circle appears to the right of the red point in the center, but they appear at different distances from the center of the eye. A point on the green square also appears to the right of the center point, like this. But notice in this case, the disparity in position of both the blue and green bars is negative, both are shifted to the left in the right image. But the green square is further away in depth from the point of focus than the blue circle, and in the images, the green bar here is shifted more to the left than the blue bar. When we see a feature with a larger stereo disparity, or larger shift in position between the two eyes, this means that this feature is further away in depth from the fixation point.

[09:44] I encourage you to repeat this exercise on your own, for the case where the two eyes are turned to focus on the green square in the back. You'll find in this case, that both the blue circle and red triangle will have positive stereo disparity, because they're both now in front of the object of focus, but the red triangle will have a larger positive disparity, a larger shift between the two images, because it's further away in depth from our point of focus on the green square. The key takeaway message here is that the stereo disparity of objects in a scene will change as we move our eyes around to focus on points at different depths. We'll take advantage of this observation when we talk about models of human stereo vision.

[10:42] [slide 7] I'd just like to mention a second way to define stereo disparity that's often used in the literature on stereo vision in biological systems. We're going back to the case where the eyes are focused on the blue circle in the middle, and this time, we'll consider the angle between the two lines of sight that meet at our point of focus. I'll call this angle alpha. For a point on the red triangle, we can construct this same angle between the two lines of sight, and I'll call this angle beta. We can do the same thing for the green square, and I'll again call this angle beta. As I added on the right here, for any location in front of fixation, that angle beta will be larger than the angle alpha, and this difference (beta - alpha) will be greater than zero, or positive. For any point that's behind the distance of fixation, that angle beta will be smaller than the angle alpha, so this difference (beta - alpha) will be negative, or less than zero. The angular difference (beta - alpha) is referred to as angular stereo disparity, and that can also be used to infer the relative depth of points in the scene - the sign of it tells us whether a point is in front or in back of the point of fixation.

[12:30] [slide 8] An added note on the special case of zero disparity - there's actually a whole surface of points out in space, called the horopter, where each point on that surface projects to the same location in the two eyes, so it has zero stereo disparity. In this bird's eye view here, the eyes are focused at this point labeled fixation at the top, and that projects to the center of the left and right eyes. If you consider, for example, this point with the yellow lines here to the left of fixation, that point projects to a location in the left image that's slightly to the right of the center, and that point projects to the same location in the right image, slightly to the right of the center. In fact all the points around this circle that people refer to as the Vieth-Muller circle, they project to the same location in the left and right images. We'll also come back to the horopter when we talk about human stereo vision.

[13:45] [slide 9] So points at different depths can project to different locations in the left and right eyes, and our visual system senses this disparity in position and uses this information to infer the depths of surfaces in the environment. Computer vision systems can also make these inferences, and this is an example of results from an automated stereo system. These are two aerial views of the Pentagon taken from opposite wingtips on an airplane, and these images were processed to identify the shifts in position of features in the left and right images. The stereo disparities were then used to create this three-dimensional reconstruction of the scene. You can see the building above the ground, with the lower courtyard in the middle of the Pentagon, and it's hard to see here, but the system is sensitive enough that cars in the parking lot are shown above the ground. What's the process of getting from these input images to this depth map?

[14:52] [slide 10] There are three main steps to the process. The first is to extract features from the left and right images, whose stereo disparity will be measured. There are choices here, for example, do we want to measure the shift of small patches of brightness, or the shift of edges of the sort you learned about earlier, or whole objects, like the Arizona can or the Hobbit box. The second step is to match up these features in the left and right images and measure their disparity in position. For example, for each feature in the left image, like the left border of this sculpture in the foreground, we want to match that up with its corresponding feature in the right image, the same left border of the sculpture, so we can then measure the difference in position of that particular feature in the two images. We can again see that shift in position if I flip back and forth between the two images. That's the step of matching up the left and right images, and this is a problem that's often referred to as the stereo correspondence problem. The last step is to use that stereo disparity to compute depth. By far, the most difficult step in this process is the step in the middle, which is referred to, as I said, the stereo correspondence problem. The next video will elaborate on this step, what makes it challenging, and some strategies for solving this problem. But before exploring stereo correspondence, I'd just like to touch briefly on the third step of using stereo disparity to compute depth.

[16:52] [slide 11] We're going to assume a simple scenario illustrated in the upper right, where we have two stereo cameras facing in the same direction. The optical axis for each camera is shown with a dotted white line, and in this case, the two optical axes are parallel. On the left is a bird's eye view of the configuration of two cameras. The black dots at the bottom are the focal points inside the cameras, and the horizontal bars are the left and right images seen from above. The center of each image is where the optical axes intersect the image. We're going to assume that we know the distance between the two cameras, that I labeled T here, and we'll also assume that we know the distance from the focal point to the images, which I labeled f. The three-dimensional scene is projected onto the image using perspective projection, illustrated in the lower right corner. To construct the image in this case, we again draw lines from the three-dimensional scene, this time through an image plane in front of the eye or camera, to the center of projection or focal point. On the diagram on the left, let's add a point p in space. We'll also construct the lines from p to the focal point in each image, and see where it intersects our image plane, and those are the open circles that are shown in blue here. x_l will refer to the horizontal coordinate of that point in the left image, and it's to the right of the center of the image, so it's going to have a positive value. We'll use x_r to refer to the coordinate of the projection of p in the right image, and in this case, it's to the left of the center of the image, so it's going to have a negative value.

[19:05] What do we want to compute here? We want to compute the depth of the point p, which I'll define as the distance from the line connecting the two cameras out to that location p, and it's labeled Z in red. How can we compute Z? There are two similar triangles here - a larger triangle formed by the two focal points and the point p, and there's a smaller triangle formed by the two projections and that point p. We can take advantage of these similar triangles to create an equation that says that T is to Z as the distance between the two projections is to the height of the shorter triangle. The height of the shorter triangle is just Z - f, and this distance between the two projected points is T minus the quantity x_l that's positive, plus the quantity x_r, which is actually a negative value. That's what we have in the numerator here. Shuffling things around, we can see that the depth Z that we want, is a function of the two quantities that we know, f and T, and this quantity x_l - x_r, that's the stereo disparity that we can measure after we've matched up features in the left and right images. The equations for the actual projection of features in the left and right eyes in the human visual system are a bit more complex than this, but this should at least give you the sense that it's basically just a geometry problem that we need to solve to get depth from disparity.

[slide 12] Returning to the main steps of the stereo process, the next video will explore the stereo correspondence problem, and simple strategies for solving this problem in a computer vision system. In our next class, we'll learn about the human stereo system and some key ideas about how we solve this problem - some of these ideas have been integrated into computer stereo systems.