Video: Human Stereo Vision

[00:01] [slide 1] This video examines some key aspects of human stereo vision, and briefly touches on the neural mechanisms underlying stereo processing in the monkey visual system. The companion video explores a computational model of human stereo processing that incorporates many of the observations described here.

[00:21] I said in our very first class that stereo is our most accurate means of sensing the three-dimensional structure of the scene from the two-dimensional image. So how accurate is our stereo ability? Consider the diagram in the bottom right and imagine the eyes are focused on this distant point, and we place another point at some distance in front of it. Or vice versa, maybe the eyes are focused on the closer point and another point is placed slightly behind it. How much of a difference in depth does it take, in order for us to reliably say which point is in front? That's the essence of stereoacuity, and it's often defined as the difference between the two angles here, formed by the lines of sight - the difference between the two angles marked with the blue arcs. Our stereoacuity is truly remarkable - it's only a few seconds of visual angle, and an example of what this means is that if we view an object at a distance of 30 cm, we can sense a difference in depth of one hundredth of a cm. Why would we ever need this kind of precision? In our man-made world, a task like threading a tiny needle may need that kind of precision. If we were still swinging among the trees, maybe we'd need that accuracy to make an accurate landing. One implication of this stereoacuity is that when we're matching features between the left and right images, and computing their disparity in position, we need to be able to determine the location of the features in each image at a very high resolution. We'll come back to this point when we talk about a model for human stereo processing.

[02:17] [slide 2] We already said that our ability to fuse random-dot stereograms tells us that the human stereo system can function independently of other visual processes, and is capable of matching very simple features. All we have here are dots, there's no distinct edges or objects. If some of the dots are shifted in position in the right image, like the square region outlined in red here, and we view the stereogram in such a way that the left eye views the left pattern and the right eye views the right pattern, we sense the difference in position of the dots between the left and right views, and we're able to perceive surfaces in depth, as we do somewhat when I show the two images in motion.

[03:07] [slide 3] The features also need to be similar between the two images. If we present a positive image to one eye and a negative of its stereo pair to the other eye, we're not able to fuse the two images together and see depth. We only match features that have the same sign of contrast in the left and right images. In the case of random-dot stereograms, we only match white dots in the left image to white dots in the right, and the same with black dots.

[03:41] [slide 4] A very important property that was described in the excerpt that you read earlier from the book by Wolfe and colleagues, is that we can only actually fuse together the images of objects in the two eyes that lie within a limited range of depth around our current fixation

distance. What do we really mean by “fuse” here? We mean that objects that are viewed in the left and right eyes appear as single solid objects - it’s the sense we experience for the object that we’re looking directly at, in the center of our field of view. Objects that are some distance away in depth from the surface we’re focused on actually appear double, like the hand here. We’re generally not aware that we’re seeing things double, but if you focus on one of your fingers and place another finger in front or behind it, you’ll experience this double vision as you separate the two fingers in depth.

[04:45] In a previous video that introduced stereo, I described the horopter - a surface in space that’s formed by points that project to the same locations in the left and right eyes. These points have zero disparity, similar to the point of focus. The gray region in the diagram here shows the limited range of depths where we fuse the left and right images together stereoscopically. How can we fuse together the images of objects that are located at distances that are much closer to the eyes, or much further away, outside the area of fusion? We move our eyes around to focus on objects at different depths. These are referred to as vergence eye movements and they involve rotating the two eyes inward to focus on closer surfaces, or rotating them outward to focus on more distant surfaces, as suggested by these pictures on the bottom. As we focus our eyes on surfaces at different depths, we can fuse together extended parts of the scene at these different depths, and we remember where these surfaces are in space as we move our eyes to other places in the scene. In the next video, we’ll build an algorithm that captures this behavior, and we’ll also show a simple demonstration of the creation of a complete depth map of a scene as the eyes focus on different depths.

[06:21] [slide 5] We’ve been focusing here on the disparity in position of features in the horizontal direction, and when we introduced the stereo correspondence problem, we mentioned the epipolar constraint. Depending on the geometric arrangement of the two cameras, or the two eyes, features in the left eye may have matching features in the right eye that are not along a horizontal line at the same height in the image. The stereo matching algorithms that we’ll explore assume that the images have been transformed so that matching features do lie along the same horizontal line in the two images. But we can ask whether the human visual system is able to fuse patterns in the left and right images that are offset in the vertical direction, as suggested by the different vertical positions of the regions outlined here in red. We refer to this offset as vertical disparity. In one perceptual experiment, random-dot patterns were presented to viewers for a very brief moment, only about 125 ms, which is enough to activate stereo, but doesn’t give us enough time to move our eyes. At one eye position, we can only tolerate a small amount of vertical shift in position between the left and right images. If we’re given more time to view a stereogram like this, we can actually rotate the two eyes independently in the vertical direction, and bring the two patterns into vertical alignment. Then the two images can be fused stereoscopically.

[08:07] [slide 6] The last property of human stereo vision that I’d like to highlight is the role of multiple operator sizes in the stereo correspondence process. Earlier, you learned that the human visual system analyzes the incoming images at multiple scales - each region of the

visual field is analyzed by retinal ganglion cells with different size receptive fields. We also described this neural processing as capturing intensity variations over different ranges of spatial frequency. The larger operator sizes capture low frequency intensity variations and small operator sizes capture the high frequency variations. These different operator sizes, or spatial frequency channels, play an important role in stereo vision.

[09:02] [slide 7] The next two slides provide a taste of some of the perceptual evidence for this role. Imagine that you start with a random-dot stereogram, maybe one that has a square region in the left image that's shifted on the right, so when fused, it gives the impression of a square surface floating in front of a background. Let's say you filter the original pattern to preserve only the low spatial frequencies, or coarse spatial structure, as shown in the top left and right images here. Suppose you then add to that image, fine scale intensity variations that have no correlation between the left and right images - think of this as high frequency noise. The left and right images on the bottom show an example of this. The top and bottom images are then added together, similar to the hybrid images that you saw earlier. What happens when you view these hybrid images stereoscopically? We can still fuse them and see depth, and there can be a large shift in position between corresponding regions in the left and right. But we only get a rough sense of distinct surfaces at different depths - the borders of the surfaces, and their depths, are a bit fuzzy.

[10:30] [slide 8] Now imagine the opposite manipulation. We again add together stereo images with information at multiple scales. This time, at the coarse scale, the low spatial frequencies, there's no correlation between what's presented in the left and right eyes. So if you look at the pattern of blobs in the two images at the top, they're very different. But we add in fine scale patterns that are correlated between the left and right images, with regions from the left image copied and shifted in the right image. In this case, we can also fuse the images, but it's more difficult. We can only tolerate small shifts in position between the two eyes, and we spend more time moving our eyes around to focus at different depths. What does this all mean for human stereo vision?

[11:30] [slide 9] In this summary of perceptual observations, jump first to the bottom section about stereo matching at multiple scales. First, these demonstrations tell us that stereo information at different scales can be processed independently. We can fuse coarse scale features in the presence of interference by junk at the fine scale, and we can fuse fine scale features in the presence of junk at the coarse scale. The visual information that's available at multiple scales doesn't need to be correlated across scales. The other points we touched on are, the image features used for matching are simple features that look similar in the two eyes and we need to measure the positions of the features very precisely, in order to account for our high stereoacuity. At a single eye position, we can only match features over a limited range of disparity in position in the horizontal and vertical directions. We need to move our eyes in order to fuse together features that initially appear over a large range of disparity. These are mostly vergence eye movements that bring our point of focus to objects at different depths in the scene.

[12:54] [slide 10] I'd like finally to briefly mention some observations from neuroscience regarding the neural mechanisms underlying stereo processing. You saw this diagram earlier, and in the context of stereo, I'd like to highlight the paths taken by input from the two eyes. Information from the right side of the visual field is processed initially by neurons in visual cortex on the left side of the brain, as shown by the green pathways here, and information from the left side of the visual field is processed by cortical neurons on the right side of the brain, as shown by the red pathways. The key thing for stereo is that each side of primary visual cortex combines information from both the left and right eyes, so it has the input it needs to perform stereo processing.

[13:52] [slide 11] There's three areas of visual cortex that we'll mention in this context. In the monkey, they're referred to as areas V1, V2, and V4. You can think of neurons in area V1 as providing input to neurons in V2, and the results of processing in V2 being sent to area V4, although the actual connections between areas are more complex.

[14:22] [slide 12] Early studies by Gian Poggio and colleagues examined the response of neurons in area V1 of the monkey, using visual images that resemble random-dot stereograms. They wanted to engage the stereo system using images that had no other cues to depth, just stereo disparity. The monkeys were trained to fixate on a target that appeared at a particular distance from the eyes, while stereo images were displayed that portrayed surfaces at different depths relative to the fixation distance. A surface at the same depth as the fixation point has zero disparity. A surface placed in front of the fixation point, closer to the monkey, was said to have "near" disparity, and a surface behind the fixation point was described as having "far" disparity. On the left here, are plots of tuning curves that indicate how particular neurons respond to surfaces at different disparities, or depths. The vertical axis is the neural response, or firing rate of the neuron. On the horizontal axis, the 0 point in the middle, marked by the dashed vertical line, that corresponds to zero disparity. Disparities on the left of the zero point correspond to near surfaces, closer to the monkey, and disparities on the right side of each figure correspond to far surfaces. They found that some simple and complex cells in area V1 were selective for stereo disparity, but with different patterns of response. The tuning curves shown in the middle here portray neurons that will respond for disparities very close to the fixation distance, right around that fixation plane. Below it are neurons that are inhibited when you place the surface at a disparity that puts it very close to the fixation distance. Other neurons responded when the surface was placed within a narrow range of depth in front of fixation, or a narrow range of depth behind fixation. These were labeled tuned near and tuned far neurons. Finally, there were neurons that responded to a large range of disparities in front of the fixation point, or a large range of disparities behind ir. A general observation that they made is that neurons with larger receptive fields tended to be selective for a larger range of disparity, they had broader tuning curves. Neurons with these general behaviors are also found in area V2, and appear to provide the building blocks for stereo processing in the brain, but the stereo correspondence problem itself is not solved in area V1, and perhaps not in area V2 either. I'll come back to this point in a moment.

[18:00] [slide 13] I'd first like to turn to another important discovery about the behavior of some neurons in area V2. We generally perceive the world as comprised of objects with clear borders that define their overall shape and extent, and help us to recognize objects, navigate around them as we move through the scene, and manipulate objects with our hands. An important goal of stereo processing is to locate potential object boundaries where there's a change in depth, and also to determine which objects the boundaries belong to. This is a property that's often referred to as border ownership, and you'll read more about the study of border ownership in the article by Williford and von der Heydt that's linked from the schedule page. To understand the meaning of border ownership, look at the two simple images at the top here, particularly the one in the upper left-hand corner here, for example. Imagine that you have a neuron whose receptive field is defined by this dotted black curve here, this oval figure. In both cases here, what's seen within the receptive field is exactly the same - there's an edge here that's light gray on the left and dark gray on the right, in both cases. But our interpretation of the scene is very different. On the far left here, that edge seems to belong to a light square on the left of the receptive field. Whereas in the right pattern, that same edge appears to belong to a dark square object on the right of the receptive field. In area V1, neurons don't make this distinction - a neuron that responds to an edge of this orientation and contrast will respond the same in both scenarios. But there are neurons in area V2 that do make this distinction - they might, for example, only respond to this edge if we perceive it as belonging to an object on the right, even though the evidence for this interpretation exists far beyond the receptive field of this neuron.

[20:43] Bringing the discussion back to stereo, some of these neurons also respond to a border in their receptive field that's defined by stereo disparity instead of brightness. Imagine a monkey viewing a random-dot stereogram that has a tilted square region that's shifted in position in the right image, so it gives rise to the percept of a surface floating out in front. The dashed red lines portray the location of the border, but don't actually appear in the image itself. Some neurons that respond to a luminance edge like this that's part of an object to the right, also respond to borders defined by stereo disparity, but only when the border belongs to an object at a closer depth on the right. Note that we could also create this random-dot stereogram with disparities that cause the square area to appear behind a surrounding surface, as if we're looking at the surface through a window. A neuron of the sort I just described would not respond to the stereo border in this case, because it belongs to a surface on the left.

[22:03] Finally, I'd like to return to the comment on the last slide, that the stereo correspondence problem is not solved in area V1. I said earlier on, that we only match similar features in the left and right images, for example, dots of the same color. In the stereogram here, the positions of the dots are the same in the left and right images, but their contrast is flipped. You can see this in the small region at the top outlined by the white line. All the white dots in this region are shown as black dots on the right, and vice versa. We can't actually fuse these stereo images together and see depth. But neurons in the early stages of cortical processing, like areas V1 and V2, will vary their response as the disparity of the dots is adjusted. You can think of these responses as capturing the possibility that certain white dots on the left might match certain

white dots on the right that happen to have disparities in position that are preferred by the neuron, but these are not the correct correspondences between the dots in the left and right that we actually perceive in the end. As the visual system tries to resolve what are the correct correspondences between left and right dots, it eventually figures out that the early neural signals here do not reflect correct matches, and it suppresses these signals, and we don't perceive any variations in depth in these patterns. In later stages of cortical processing, in area V4 for example, there's a strong correlation between how the neurons respond to depth, and our perception of the depths of surfaces in a scene.

[24:10] That's a brief introduction to some key observations from perception and physiology that provide insight into the processing of stereo information in the human visual system, and in monkeys that have stereo vision similar to ours. The next video provides an example of how you could design a stereo correspondence algorithm that captures many of these observations.