EXERCISES 27

COMPUTER VISION

27.1 Introduction

Exercise 27.CVDT

Define the following terms in your own words.

a. Active and passive sensing
b. Image feature
c. Object model
d. Rendering model

a. Active and passive sensing: Passive sensing is what we normally do with our eyes–we receive stimulus from the world. Active sensing is what bats do with their ultrasound–send out a signal and detect what effect it has on the world. We can also do this with a flashlight.
b. Image feature
c. Object model
d. Rendering model

Exercise 27.OTCV

Name as many perceptual channels as you can that are used by robots today.

Vision, sound, touch (with capacitive or resistive sensors), force (with strain gauges, piezoelectric sensors, or other types), range finding (with ultrasonic or laser range finders), proximity sensing (with electromagnetic detector), temperature, acceleration, and global positioning have all been used by various robots.

27.2 Image Formation

Exercise 27.PINH

In the shadow of a tree with a dense, leafy canopy, one sees a number of light spots. Surprisingly, they all appear to be circular. Why? After all, the gaps between the leaves

through which the sun shines are not likely to be circular.

The small spaces between leaves act as pinhole cameras. That means that the circular light spots you see are actually images of the circular sun. You can test this theory next time there is a solar eclipse: the circular light spots will have a crescent bite taken out of them as the eclipse progresses. (Eclipse or not, the light spots are easier to see on a sheet of paper than on the rough forest floor.)

Exercise 27.SPHR

Consider a picture of a white sphere floating in front of a black backdrop. The image curve separating white pixels from black pixels is sometimes called the "outline" of the sphere. Show that the outline of a sphere, viewed in a perspective camera, can be an ellipse. Why do spheres not look like ellipses to you?

Consider the set of light rays passing through the center of projection (the pinhole or the lens center), and tangent to the surface of the sphere. These define a double cone whose apex is the center of projection. Note that the outline of the sphere on the image plane is just the cross section corresponding to the intersection of this cone with the image plane of the camera. We know from geometry that such a conic section will typically be an ellipse. It is a circle in the special case that the sphere is directly in front of the camera (its center lies on the optical axis).

While on a planar retina, the image of an off-axis sphere would indeed be an ellipse, the human visual system tries to infer what is in the three-dimensional scene, and here the most likely solution is that one is looking at a sphere.

Some students might note that the eye's retina is not planar but closer to spherical. On a perfectly spherical retina the image of a sphere will be circular. The point of the question remains valid, however.

Exercise 27.CORR

Consider an infinitely long cylinder of radius r oriented with its axis along the y-axis. The cylinder has a Lambertian surface and is viewed by a camera along the positive z-axis. What will you expect to see in the image if the cylinder is illuminated by a point source at infinity located on the positive x -axis? Draw the contours of constant brightness in the projected image. Are the contours of equal brightness uniformly spaced?

Recall that the image brightness of a Lambertian surface (page 850) is given by $I(x, y) = k\mathbf{n}(x, y).\mathbf{s}$. Here the light source direction $\mathbf{s}$ is along the x-axis. It is sufficient to consider a horizontal cross-section (in the x–z plane) of the cylinder as shown in Figure S27.1(a). Then, the brightness $I(x) = k\cos\theta(x)$ for all the points on the right half of the cylinder. The left half is in shadow. As $x = r\cos\theta$, we can rewrite the brightness function as $I(x) = \frac{kx}{r}$ which

reveals that the isobrightness contours in the lit part of the cylinder must be equally spaced. The view from the z-axis is shown in Figure S27.1(b).

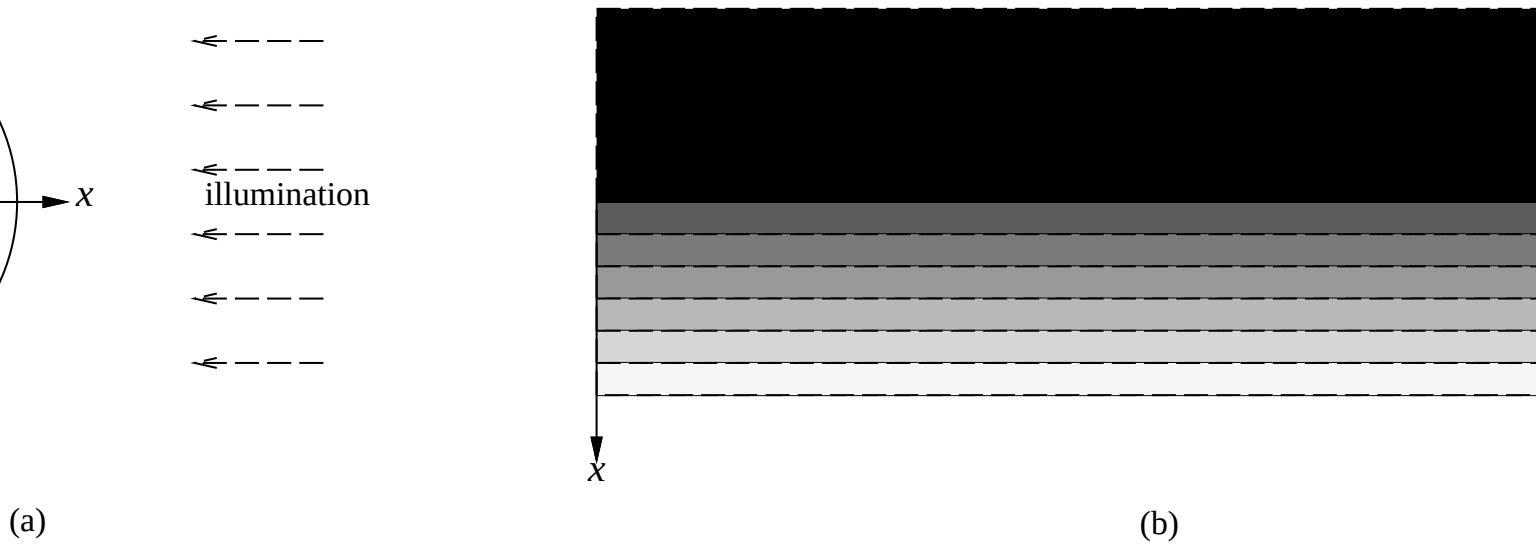

Figure S27.1 (a) Geometry of the scene as viewed from along the y-axis. (b) The scene from the z-axis, showing the evenly spaced isobrightness contours.

Exercise 27.DOFC

Normally we think of more depth of field as a good thing: we want more of the image in focus. But some photographers prize shallow depth of field, because it makes the subject of the photograph stand out from a blurred background. It is known that if two cameras with the same aperture number (say, f/2) take a picture of the same scene with the same angle of view, then the cameras with a small sensor (such as a cell phone) will have a deeper depth of field and less background blurring, while a large-sensor camera will have shallower depth of field and more background blurring. Why is that?

Depth of field is proportional to $1/f^2$ where f is the focal length. So images with the same focal length (and a few other factors being equal) will have the same depth of field regardless of sensor size. But to achieve the same angle of view–the same scene–as a full-frame camera with a 28mm focal length, a cell phone camera would have a focal length of about 4mm. This makes sense: think of the diagram with the triangle tapering down from the image to the lens and then a similar but smaller triangle going to the sensor. To get the same angle of view a small sensor has to be close and a large sensor farther away. Another way to look at it is that an aperture number of f/2 means the actual lens aperture diameter is 28mm/2 or 14mm for the full-frame camera and 4mm/2 = 2mm for the cell phone. If the diameter is larger then there is more opportunity for light from slightly different image patches to be farther apart on the lens and thus farther apart when projected on the sensor.

Exercise 27.FTCL
Does a small cell phone camera lens with an aperture of f/2 gather as much light from each patch in the scene as an f/2 lens on a large SLR camera that is taking an image of the same scene?

For each scene patch, a smaller lens gathers less light. But the smaller lens has a smaller focal length to achieve the same scene, so the light is concentrated into a smaller area on the sensor. Thus the amount of light per unit area is the same.

27.3 Simple Image Features

Exercise 27.EDGE
Edges in an image can correspond to a variety of events in a scene. Consider Figure 27.4 (page 994), and assume that it is a picture of a real three-dimensional scene. Identify ten different brightness edges in the image, and for each, state whether it corresponds to a discontinuity in (a) depth, (b) surface orientation, (c) reflectance, or (d) illumination.

We list the four classes and give two or three examples of each:

a. *depth*: Between the top of the computer monitor and the wall behind it. Between the side of the clock tower and the sky behind it. Between the white sheets of paper in the foreground and the book and keyboard behind them.

b. *surface normal*: At the near corner of the pages of the book on the desk. At the sides of the keys on the keyboard.

c. *reflectance*: Between the white paper and the black lines on it. Between the "golden" bridge in the picture and the blue sky behind it.

d. *illumination*: On the windowsill, the shadow from the center glass pane divider. On the paper with Greek text, the shadow along the left from the paper on top of it. On the computer monitor, the edge between the white window and the blue window is caused by different illumination by the CRT.

Exercise 27.WALZ
The Waltz algorithm takes as input a list of the vertices in a sketch diagram of one or polyhedrons. Each vertex is described by the number of lines that meet and whether the lines are colinear. From that, the algorithm outputs a description of the 3D scene, by propagating constraints about the vertices through the connecting lines. Implement a version and report on how well it works.

An example implementation is in Chapter 16 of `https://github.com/norvig/paip-lisp`.

27.4 Classifying Images

Exercise 27.CLIM

Build an image classification model, by following a pre-built recipe, and report on what worked well and what you had problems with.

(One example: `https://www.kaggle.com/rohandeysarkar/ultimate-image-classification-guide-2020`.)

Answers will vary by student.

Exercise 27.CVML

Compare three architectures for computer vision image classification: (1) convolutional neural networks (CNN), (2) vision transformers (ViT), and (3) multi-layer perceptron mixers (MLP-Mixer). Find relevant research and report on your findings.

Convolutional neural networks are emphasized in the book. Vision transformers are covered in `https://arxiv.org/abs/2010.11929`. Mult-layer perceptrons are covered in the key paper `https://arxiv.org/abs/2105.01601`.

CNNs have an inductive bias that is well-suited to understanding images: the idea that nearby pixels in the image are often nearby elements in the world, and that elements may appear at any location in the image. Vision transformers lack that inductive bias and thus do worse when trained on a small or medium number of images, but start to do well with hundreds of millions of images and can exceed CNN performance. The MLP-Mixer model emphasizes speed of training, which is important for very large models. It is said to achieve similar accuracy as other models with 1/3 the training time.

Students should go into more detail about the various differences.

27.5 Detecting Objects

Exercise 27.YOLO

Research the YOLO (You Only Look Once) approach to object detection, and compare it to the RCNN region proposal approach described in the book.

The key paper is `https://arxiv.org/abs/1506.02640v5` YOLO's big advantage is speed: it can operate in real time (45 frames per second). It is a nice advantage that the code is open source. It achieves this impressive speed by training a single end-to-end network that predicts bounding boxes and category labels all at once. YOLO is not quite as accurate as RCNN methods at locating boxes with great accuracy, but it is resistant to false positives coming from background noise.

Exercise 27.MAPV
Object detectors are commonly evaluated with a metric known as mean average precision (mAP). Research this method and explain how it works.

We explained in the book how the "area under the curve" (AUC) metric works: it graphs recall on the x-axis and precision on the y-axis and measures the area under this curve. Average precision (AP) does this, but instead of computing the exact area by integration, it samples the precision at a number of recall points and sums the areas of rectangles whose height is the precision and whose width is the difference from one sample point to the next. The mean average precision (mAP) is then the average of the AP scores over all object classes.

Exercise 27.CNTN
A new object detection approach called CenterNet models an object as a single point, the center of its bounding box. Research the approach and report on your findings.

The key paper is `https://arxiv.org/abs/1904.07850` and the github repository is `https://github.com/xingyizhou/CenterNet`. CenterNet gives strictly less information than other approaches: it doesn't provide the size or shape of the bounding box, or the pose. Bit its simplicity allows it to be faster, running in real time.

Exercise 27.I
mage classification relies on labeled image data. Where does that come from? How accurate are the labels? How hard is it to create the labels?
Read `http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/` to see how one machine learning researcher, Andrej Karpathy, spent time labeling ImageNet data and measured his own accuracy at 94%. Try the task yourself and see what accuracy you achieve. Can you think of better ways of collecting labels for images?

Each student will have their own reactions to this task.

27.6 The 3D World

Exercise 27.STRO
A stereoscopic system is being contemplated for terrain mapping. It will consist of two CCD cameras, each having 512×512 pixels on a 10 cm $\times$ 10 cm square sensor. The lenses to be used have a focal length of 16 cm, with the focus fixed at infinity. For corresponding points (u_1, v_1) in the left image and (u_2, v_2) in the right image, $v_1 = v_2$ because the x-axes in the two image planes are parallel to the epipolar lines—the lines from the object to the

camera. The optical axes of the two cameras are parallel. The baseline between the cameras is 1 meter.

a. If the nearest distance to be measured is 16 meters, what is the largest disparity that will occur (in pixels)?

b. What is the distance resolution at 16 meters, due to the pixel spacing?

c. What distance corresponds to a disparity of one pixel?

Before answering this exercise, we draw a diagram of the apparatus (top view), shown in Figure S27.2. Notice that we make the approximation that the focal length is the distance from the lens to the image plane; this is valid for objects that are far away. Notice that this question asks nothing about the y coordinates of points; we might as well have a single line of 512 pixels in each camera.

a. Solve this by constructing similar triangles: whose hypotenuse is the dotted line from object to lens, and whose height is 0.5 meters and width 16 meters. This is similar to a triangle of width 16cm whose hypotenuse projects onto the image plane; we can compute that its height must be 0.5cm; this is the offset from the center of the image plane. The other camera will have an offset of 0.5cm in the opposite direction. Thus the total disparity is 1.0cm, or, at 512 pixels/10cm, a disparity of 51.2 pixels, or 51, since there are no fractional pixels. Objects that are farther away will have smaller disparity. Writing this as an equation, where d is the disparity in pixels and Z is the distance to the object, we have:

$$d = 2 \times \frac{512 \text{ pixels}}{10 \text{ cm}} \times 16 \text{ cm} \times \frac{0.5 \text{ m}}{Z}$$

b. In other words, this question is asking how much further than 16m could an object be, and still occupy the same pixels in the image plane? Rearranging the formula above by swapping d and Z, and plugging in values of 51 and 52 pixels for d, we get values of Z of 16.06 and 15.75 meters, for a difference of 31cm (a little over a foot). This is the range resolution at 16 meters.

c. In other words, this question is asking how far away would an object be to generate a disparity of one pixel? Objects farther than this are in effect out of range; we can't say where they are located. Rearranging the formula above by swapping d and Z we get 51.2 meters.

Exercise 27.TFCV

Which of the following are true, and which are false?

a. Finding corresponding points in stereo images is the easiest phase of the stereo depth-finding process.

b. Shape-from-texture can be done by projecting a grid of light-stripes onto the scene.

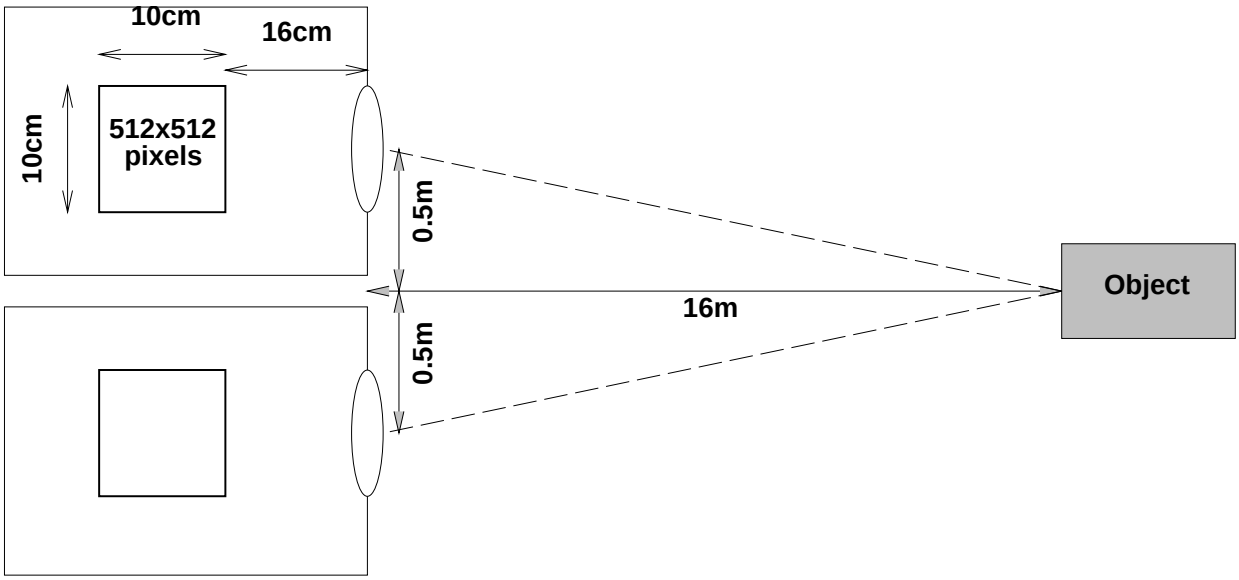

Figure S27.2 Top view of the setup for stereo viewing (Exercise 24.6).

c. Lines with equal lengths in the scene always project to equal lengths in the image.

d. Straight lines in the image necessarily correspond to straight lines in the scene.

a. False. This can be quite difficult, particularly when some point are occluded from one eye but not the other.

b. True. The grid creates an apparent texture whose distortion gives good information as to surface orientation.

c. False.

d. False. A disk viewed edge-on appears as a straight line.

Exercise 27.BOTT

(Courtesy of Pietro Perona.) Figure **??** shows two cameras at X and Y observing a scene. Draw the image seen at each camera, assuming that all named points are in the same horizontal plane. What can be concluded from these two images about the relative distances of points A, B, C, D, and E from the camera baseline, and on what basis?

A, B, C can be viewed in stereo and hence their depths can be measured, allowing the viewer to determine that B is nearest, A and C are equidistant and slightly further away. Neither D nor E can be seen by both cameras, so stereo cannot be used. Looking at the figure, it appears that the bottle *occludes* D from Y and E from X, so D and E must be further away than A, B, C, but their relative depths cannot be determined. There is, however, another

possibility (noticed by Alex Fabrikant). Remember that each camera sees the camera's-eye view not the bird's-eye view. X sees DABC and Y sees ABCE. It is possible that D is very close to camera X, so close that it falls outside the field of view of camera Y; similarly, E might be very close to Y and be outside the field of view of X. Hence, unless the cameras have a 180-degree field of view—probably impossible—there is no way to determine whether D and E are in front of or behind the bottle.

27.7 Using Computer Vision

Exercise 27.ODED

Suppose you have a CNN object detection algorithm that runs very well in a data center on a large cluster of computers. Now you want to deploy it onto cell phones (without draining users' batteries). What can you do?

A good survey is `https://arxiv.org/abs/1704.04861`, which describes MobileNets, which were designed for this purpose. Some things to try:

a. Use a smaller network with fewer convolutional blocks and fewer parameters.

b. Experiment with hyperparameters for the size of the model, such as a width multiplier, and understand the tradeoffs of accuracy versus model size, so that you can pick an appropriate size based on the available device.

c. Quantize your network: Instead of using 32 bit or 64 bit floating point numbers, scale that down to 8 bit integers and reduce memory and CPU demand, at the cost of some loss in accuracy. Quantization is directly supported by platforms such as TensorFlow Mobile and Caffe2Go.

d. You can reduce the resolution of images you process (and thus the size the network). You may only need 100 pixels on a side, not thousands of pixels.

Exercise 27.STYT

Examine one of the available code repositories for **style transfer**, such as `https://colab.research.google.com/github/tensorflow/models/blob/master/research/nst_blogpost/4_Neural_Style_Transfer_with_Eager_Execution.ipynb` or `https://www.tensorflow.org/tutorials/generative/style_transfer` or `https://github.com/hnarayanan/artistic-style-transfer` or `https://www.digitalocean.com/community/tutorials/how-to-perform-neural-style-t` Try it on some combinations of content image an style reference images. Experiment with changing some aspects of the model, such as the number of content layers and style layers, or the loss functions. What works well and what doesn't work? What does "works well" even mean in this context?

The main point of this exercise is to get experience running networks in TensorFlow/Keras/PyTorch, and to think about the resulting output. Students will have different conclusions based on the images they work on and their interests. In general, styles that have distinctive profiles in

terms of low-level frequencies (as seen in Van Gogh, Kandinsky, and Hokusai) work best. "Works well" means that the viewer sees the result as interesting, and that the style and content are both recognizable.

Exercise 27.CRWD

Crowd counting is the task of estimating the number of people in an image (or images). Whenever there is a big event, organizers of the event tend to exaggerate the crowd size, and competitors try to minimize it. It would be helpful to have an accurate estimate.

Experiment with a crowd counting computer vision implementation such as one found at `https://github.com/topics/crowd-counting`. How well does it work?

Contrast the computer vision approach with an approach described in the articel "Quantifying crowd size with mobile phone and Twitter data" (`https://royalsocietypublishing.org/doi/10.1098/rsos.150162`). Crowd counting has societal benefits in traffic control and public safety, but it also raises some ethical issues. Discuss.

The implementations give reasonable numbers, but it is difficult to establish ground truth. It is hard work to look at images and count them by hand, and we don't know how accurate humans are at this task. Also, an image is just one view of a crowd at one instance in time, and hence is an incomplete record.

One ethical issue is that people are not very good at understanding large numbers, so promoting more use of numbers (as contrasted with other ways of covering an event) may not be helpful. The main issue is one of privacy: people expect to be able to attend an event without revealing their identity. Crowd counting does not detect identities, but it is a step towards more general surveillance.

The approach based on mobile phone data seems to be useful, particularly for events where it is difficult to obtain images (perhaps because it is dark or the event is indoors and aerial photos are not possible). Combining both methods may lead to better estimates.