Computer Vision 1, Final Exam

23 Oct 2020

1 Image Capturing (15pt)

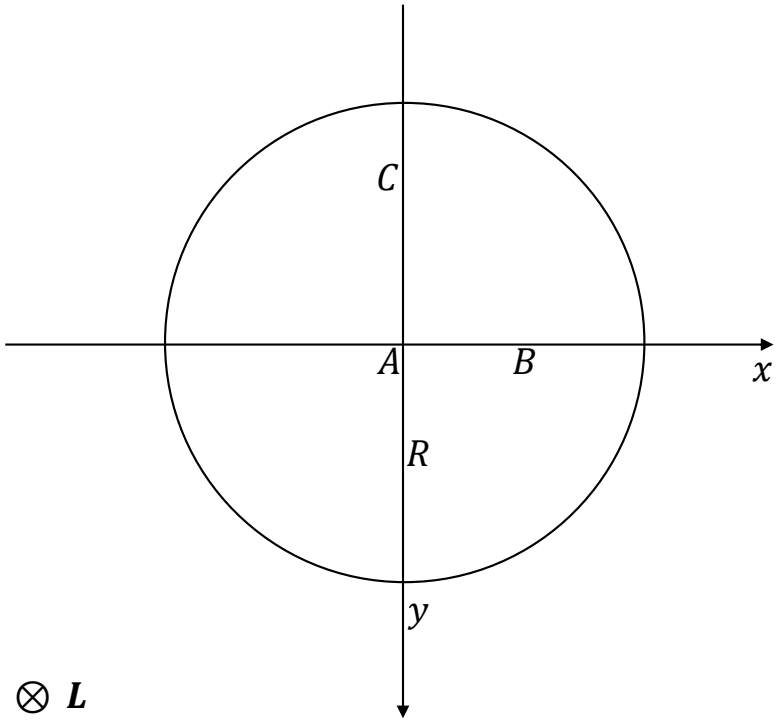

Figure 1: Image of a sphere

Figure 1 shows a captured image of a sphere. The sphere has a radius R. The sphere center is at point A. The camera center is pointing to the sphere center A. The camera is placed sufficiently far away from the sphere so that perspective distortion can be ignored. The camera is assumed to be an ideal

pinhole camera. $\mathbf{L}$ is the only light source, which is a uniform and parallel light pointing to the image plane. There is no other light source or object in the setting.

- (Q.a) (2pt) We know point A is at the center of the sphere. We also know that point B is on x-axis and the distance between A and B is $\frac{1}{2}R$. We know that point C is on y-axis and is $\frac{\sqrt{3}}{2}R$ away from A. Is the information provided so far sufficient to calculate the surface normal at point A, B and C, denoted as $\mathbf{N}_A$, $\mathbf{N}_B$ and $\mathbf{N}_C$ respectively? If so, provide the angle between $\mathbf{L}$ and $\mathbf{N}_A$, denoted as θ_A. Similarly, provide θ_B and θ_C. If the information is not sufficiently, explain why? Use illustrations if necessary.

- (Q.b) (1pt) Assume the image is gray scale, and we know the intensity observed by the camera at point A is $I_A = 100$. Is the information provided so far sufficient to derive the intensity at point B and C, denoted as I_B and I_C respectively. If so, provide the value with derivation. If not, explain why. Use illustrations if necessary.

- (Q.c) (1pt) Assume the sphere material is Lambertian (ideally diffusing). The gray scale image has $I_A = 100$. Is the information provided so far sufficient to derive I_B and I_C? If so, provide the value with derivation. If not, explain why. Use illustrations if necessary.

- (Q.d) (1pt) Assume the sphere material is ideally glossy, which works like a mirror. The gray scale image has $I_A = 100$. Is the information provided so far sufficient to derive I_B and I_C? If so, provide the value with derivation. If not, explain why. Use illustrations if necessary.

- (Q.e) (1pt) Assume the sphere material is Lambertian (ideally diffusing). And the sphere material is uniform, say the albedo is uniform across the sphere. The gray scale image has $I_A = 100$. Is the information provided so far sufficient to derive I_B and I_C? If so, provide the value with derivation. If not, explain why. Use illustrations if necessary.

- (Q.f) (2pt) Assume the image is a 8bits gray scale image, which mean the integer value ranges from 0-255. Originally $I_A = 100$, $I_B = 80$, $I_C = 60$. We want to compress the image to 4bits for storage, and later recover it to 8bits images for display. The compression and decompression are linear. What are the compressed value of I_A, I_B, I_C and what are the decompressed value of them? Provide the value with derivation.

- (Q.g) (1pt) Assume the camera captured image has $I_A = 100$, $I_B = 80$, $I_C = 40$. Assume the light source $\mathbf{L}$ is a single wavelength light, and the wavelength is $1100nm$. Assume the sphere is not florescent, instead of using a camera, and keep the other settings identical, can we human see the sphere using our eyes directly? Why?

- (Q.h) (2pt) Assume the camera captured image has $I_A = 100$, $I_B = 80$, $I_C = 40$. Assume the light source $\mathbf{L}$ is a single wavelength light, and the wavelength is $400nm$. Instead of using a camera, and keep the other settings identical, can we human see the sphere using our eyes directly?

Why? Assume the sphere is not florescent, can we tell the color of the sphere that we see? Why?

- (Q.i) (1pt) We now want to do some geometry editing of the captured image. First, we want to move the entire sphere to the right along the x-axis by 10pixels. Using homogeneous coordinates, we denote the original coordinates of A as $\mathbf{P}_A = [x_A, y_A, 1]^T$. Provide the transform matrix using only translation.

- (Q.j) (2pt) We now want to flip the image upside-down with the flipping axis on the line determined by points A and B. We know originally $\mathbf{P}_A = [100, 100, 1]^T$, $\mathbf{P}_C = [100, 80, 1]^T$. Provide the transform matrix.

- (Q.k) (1pt) We now want to do some affine transformation. We know the coordinates of points A, B and C before and after the transformation. Can we determine the transformation matrix? Why?

Computer Vision 1, Final Exam

23 Oct 2020

2 Image Filtering and Features (20pt)

Edges and corners are important features from which image descriptors can be extracted.

- (Q.a) (1pt) A simple filter that approximates the derivative in x direction f_x is given by $h = [-1, 1]$. Show how higher order derivative f_{xx} can be approximated using combination of h.
- (Q.b) (2pt) Given the following filter:

$$g = \alpha \begin{array}{|c|c|c|} \hline 1 & 2 & 1 \\ \hline \end{array}$$

  Find a simple filter that can achieve the same effect by applying multiple convolutions. We want to keep the mean value of the entire image before and after the filtering, what should the real number coefficient α be.

- (Q.c) (2pt) When convoluted with an image f, which of the filters, h, g or $g * h$ correspond to a low-pass, high-pass or band-pass filter? Please provide the explanations.

Consider the following image patches:

$$P = \begin{array}{|c|c|c|c|} \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 2 & 2 & 2 & 2 \\ \hline 2 & 2 & 2 & 2 \\ \hline \end{array} \qquad Q = \begin{array}{|c|c|c|c|} \hline 0 & 0 & 2 & 2 \\ \hline 0 & 0 & 2 & 2 \\ \hline 0 & 0 & 2 & 2 \\ \hline 0 & 0 & 0 & 0 \\ \hline \end{array}$$

- (Q.d) (2pt) Compute the results after performing cross-correlation on image P by a 3*3 filter U, where all elements at the image boundaries (outside image P) are mirrored. The elements outside filter U are all zeros.

$$U = \begin{array}{|c|c|c|} \hline 1 & -2 & 1 \\ \hline -2 & 4 & -2 \\ \hline 1 & -2 & 1 \\ \hline \end{array}$$

- (Q.e) (1pt) What kind of features are detected by the above filtering?
- (Q.f) (2pt) Show that U can be separated into a combination of two 1 dimensional filters.

- (Q.g) (2pt) Compute the second moment matrix

$$M = \begin{pmatrix} f_x^2 & f_x f_y \\ f_x f_y & f_y^2 \end{pmatrix}$$

for image patch Q using a simple derivative filter $h_x = [-1, 1]$ in the x-direction and $h_y = [-1, 1]^T$ in the y-direction. The center of h_x is at the first element, so as h_y. Use cross-correlation for simplicity. Handle the out-of-boundary pixels using mirror.

- (Q.h) (1pt) Explain how can you use the eigenvalues of M to detect a corner?

- (Q.i) (2pt) Computer the histogram of oriented gradients for patch Q in the same way as SIFT descriptor.

- (Q.j) (2pt) Name two photometric image transformations that the SIFT descriptor is invariant to.

- (Q.k) (1pt) Can we compute a deterministic optical flow between P and Q? If so, provide the optical flow. If not, explain why.

Consider the following image patches:

$$S = \begin{array}{|c|c|c|c|} \hline 0 & 1 & 1 & 0 \\ \hline 0 & 1 & 1 & 0 \\ \hline 2 & 3 & 12 & 3 \\ \hline 2 & 2 & 2 & 3 \\ \hline \end{array}$$

- (Q.l) (2pt) Compute the filtered image using a 2*2 averaging box filter and a 2*2 median filter. Using mirror padding for pixels outside the patch.

Computer Vision 1, Final Exam

23 Oct 2020

3 Object Detection and Classification (15pt)

In general, a sliding-window approach can be used to detect objects.

- (Q.a) (1pt) How does the sliding-window approach work?
- (Q.b) (2pt) Assuming we want to find an object in a 1000*1000 image. The sliding window size is 100*100. Assuming the sliding step is 5 pixels in both x and y direction, how many windows we need in total? If we increase the step to 10 or 20 pixels, how many windows we need in total respectively?
- (Q.c) (1pt) Mention at least one pros and one cons of increasing the searching step.
- (Q.d) (1pt) Do you know another approach to limit the number of window candidates?

Given an image database of 10.000 images. A programmer has designed and implemented two different image retrieval systems S_1 and S_2. A company is interested in the performance of the two systems. To this end, the company has formulated two different search queries Q_1 and Q_2. The number of relevant images with respect to all search queries is 10 composed of the following set $(A, B, C, D, E, F, G, H, I, J)$. The number of images shown to the user is 20 (Answer set). Further, the order of the 20 highest ranked images of the two different retrieval systems S_1 and S_2 for query Q_1 is is follows:

$$S_1: \quad A\ L\ N\ O\ S \quad P\ Q\ B\ T\ C \quad X\ V\ W\ D\ Y \quad E\ Z\ K\ M\ R$$

$$S_2: \quad A\ M\ B\ O\ K \quad C\ D\ R\ S\ E \quad N\ V\ T\ X\ Q \quad L\ P\ U\ Y\ Z$$

And for query Q_2:

$$S_1: \quad A\ B\ L\ C\ P \quad D\ E\ F\ O\ T \quad S\ V\ W\ X\ Y \quad K\ P\ Z\ Q\ N$$

$$S_2: \quad K\ M\ Q\ N\ O \quad S\ A\ R\ B\ T \quad C\ D\ E\ X\ F \quad X\ L\ Y\ W\ P$$

- (Q.e) (1pt) Compute the precision and recall for the systems S_1 and S_2 for Q_1 respectively.
- (Q.f) (3pt) Generate the precision-recall graph for the systems for Q_2.

- (Q.g) (1pt) What conclusion can you draw from the precision-recall graph?

Convolutional Neural Network is widely used for object detection nowadays.

- (Q.h) (1pt) Describe the difference between a Convolutional layer and a fully connected layer.
- (Q.i) (1pt) Assume the input of a particular Conv Layer is 500 * 500 * 25. 100 filters with a receptive field of 7*7 are learned. What is the total number of parameters to be learned in this layer. Provide your derivation.
- (Q.j) (1pt) Assume the next Conv Layer learns also 100 filters with a receptive field of 7*7. How many parameters are learned in this layer?
- (Q.k) (2pt) Consider the following input of a max-pooling layer:

0	1	1	0
0	11	1	0
2	3	12	3
2	2	2	3

(1)

What is the output of the max-pooling layer with a 3*3 window, with stride 1? What is the output of the max-pooling layer with a 2*2 window, with stride 2?