Computer Vision 1

March 27th, 2019, 09.00-12.00

Question 1: Reflection Models

To understand the image formation process, a simple reflection model, to define the R, G and B pixel values, is given by:

$$R = \int_\lambda e(\lambda)\rho(\lambda)f_R(\lambda)d\lambda,\ G = \int_\lambda e(\lambda)\rho(\lambda)f_G(\lambda)d\lambda,\ B = \int_\lambda e(\lambda)\rho(\lambda)f_B(\lambda)d\lambda \tag{1}$$

where $e(\lambda)$ is the light source, $\rho(\lambda)$ the surface reflectance and $f_R(\lambda), f_G(\lambda), f_B(\lambda)$ are the R, G, B color filters.

- (a) Considering the reflection model of eq. 1, what is the mathematical extension of the reflection model for a Lambertian surface illuminated by a (distant) point light source? (1 pts)

we could extend the model by taking into account the angle between the surface and the light source

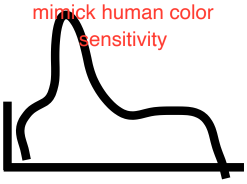
- (b) Which kind of filter responses do you choose to correlate the R, G and B values with a standard human observer (human perception)? Please explain. (1 pts)

to correlate this model more towards the way humans perceive color, you should adjust the color filters such that they more closely mimick human color sensitivity

- (c) For real-world light sources, does $e(\lambda)$ consist of a single wavelength or a combination of wavelengths? Why? (1 pts)

real-world light sources consist of a combination of wavelengths
why: I suppose because such (artificial) lightsources are more useful for illuminating a scene (as opposed to say, a laser)?

- (d) Sketch the spectral power distribution of $e(\lambda)$ for a purple light source. (1 pts)



Assuming a white light source, the simplified dichromatic reflection model, to define the R, G and B pixel values, is given by:

$$R = \cos\theta\ e\ \rho_R + e\ (\cos\phi)^s,\ G = \cos\theta\ e\ \rho_G + e\ (\cos\phi)^s,\ B = \cos\theta\ e\ \rho_B + e\ (\cos\phi)^s \tag{2}$$

where $\cos\theta = \vec{n}\cdot\vec{s}$ is the angle between the surface normal and direction of the light source, and $\cos\phi = \vec{r}\cdot\vec{v}$ depends on ϕ which is the angle between the reflected light $\vec{r}$ and the viewer $\vec{v}$. Further, s is called the specular exponent.

- (e) Explain the mechanism of the term $(\cos\phi)^s$. How is it used to model the glossy appearance of an object? (1 pts)

e: It models the specular highlights in the image;

- (f) What is approximately the shape of $(\cos\phi)^s$ for different values of s and what is the effect on the size of the specular highlights? (2 pts)

From a viewpoint perspective, approximately circular/elliptical. the higher the exponent, the smaller the specular highlight becomes

- (g) Show that the color of the highlights is dependent on the color of the light source. (2 pts)

The reflection model aims to model the pixel values for R, G and B seperately. Each reflected color has its own specular component. The strength of the specular highlight is directly dependent on the intensity of the light source, but

If you write out all the terms the cos \theta terms cancel out, thereby showing that R-G/R-B only depends on the surface reflectance

- (g) Show that $\frac{R-G}{R-B}$ only depends on the surface reflectance (albedo) i.e. ρ_R, ρ_G and ρ_B. (2 pts)

Question 2: Filters and Image Features

Edges and corners are important features from which image descriptors can be extracted. Consider the following image filters (F) and image path (I):

$$F_1 = \begin{array}{|c|c|c|}\hline 1 & 1 & 1 \\ \hline 0 & 0 & 0 \\ \hline -1 & -1 & -1 \\ \hline \end{array} \quad F_2 = \begin{array}{|c|c|c|}\hline 1 & 0 & -1 \\ \hline 1 & 0 & -1 \\ \hline 1 & 0 & -1 \\ \hline \end{array} \quad I_{\mathrm{A}} = \begin{array}{|c|c|c|c|c|}\hline 0 & 5 & 5 & 5 & 0 \\ \hline 0 & 5 & 5 & 5 & 0 \\ \hline 0 & 5 & 0 & 0 & 0 \\ \hline 0 & 5 & 5 & 5 & 0 \\ \hline 0 & 5 & 5 & 5 & 0 \\ \hline \end{array}$$

?

- (a) Compute the cross-correlation of filter F_1 and F_2 on image path I_A. Make clear how you deal with the borders (zero-padding?). (1 pts)
- (b) What do these filters compute? (1 pts) Gradient in y and x direction
- (c) Compute the gradient magnitude and orientation (in degrees). (1 pts) magnitude: sqrt(gx^2 + gy^2) orientation: ...
- (d) For which photometric transformation is the gradient invariant? (1 pts) Translation

Consider the following image filters (F) and image path (I):

$$F_3 = \begin{array}{|c|c|c|}\hline \mathrm{A} & \mathrm{B} & \mathrm{C} \\ \hline \mathrm{D} & \mathrm{E} & \mathrm{F} \\ \hline \mathrm{G} & \mathrm{H} & \mathrm{I} \\ \hline \end{array} \qquad I_{\mathrm{I}} = \begin{array}{|c|c|c|c|c|c|c|}\hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \end{array}$$

- (e) What is the result of applying filter F_3 on the identity image I_{I}? (1 pts) ?
- (f) It is mathematical convenient, that a filter applied on an identity patch results in the filter. How should you transform the procedure above in order to get the filter as output? (1 pts)

transform I_i into the identity matrix?
Or alternatively, reduce i_1 to a 3x3 matrix with a 1 in the center pixel

Consider the following image patches:

$$P = \begin{array}{|c|c|c|c|c|}\hline 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ \hline 10 & 10 & 10 & 10 & 10 \\ \hline 10 & 10 & 10 & 10 & 10 \\ \hline \end{array} \quad Q = \begin{array}{|c|c|c|c|c|}\hline 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ \hline 10 & 10 & 10 & 0 & 0 \\ \hline 10 & 10 & 10 & 0 & 0 \\ \hline 10 & 10 & 10 & 0 & 0 \\ \hline \end{array}$$

Intensity values of two small image patches P and Q.

- (g) Compute the derivatives f_x and f_y of image patches P and Q using a simple derivative filter $h_x = \begin{pmatrix} -1 & \underline{1} \end{pmatrix}$ in the x-direction and $h_y = \begin{pmatrix} \underline{1} \\ -1 \end{pmatrix}$ in the y-direction. All elements exceeding the image patches are mirrored. The elements outside the derivative filters are all zero. (2 pts)

- (h) Compute the autocorrelation matrix $M = \begin{pmatrix} \sum f_x^2 & \sum f_x f_y \\ \sum f_x f_y & \sum f_y^2 \end{pmatrix}$ for image patches P and Q. (1 pts)
- (i) Compute the eigenvalues of M for image patch Q. How can these eigenvalues be used to determine a corner? (3 pts)
- (j) Compute the eigenvectors of M for image patch Q. What do these eigenvectors mean? (3 pts)

Question 3: Object Classification and Performance

Deep learning and ConvNets are very useful for object recognition and detection. Consider the following four (simple) image patches of the letters Y, L, O and X:

$$I = \begin{array}{|c|c|c|}\hline 1 & 0 & 1 \\ \hline 0 & 1 & 0 \\ \hline 0 & 1 & 0 \\ \hline \end{array} \quad L = \begin{array}{|c|c|c|}\hline 0 & 1 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & 1 & 1 \\ \hline \end{array} \quad O = \begin{array}{|c|c|c|}\hline 0 & 1 & 0 \\ \hline 1 & 0 & 1 \\ \hline 0 & 1 & 0 \\ \hline \end{array} \quad X = \begin{array}{|c|c|c|}\hline 1 & 0 & 1 \\ \hline 0 & 1 & 0 \\ \hline 1 & 0 & 1 \\ \hline \end{array}$$

Intensity values of four image patches of the letters Y, L, O and X.

- (a) After training a single layer neural network (multi-class), the following weight matrix $\vec{M} = \begin{bmatrix} 0 & 0.3 & 0 & 0 & 0.6 & 0 & 0 & 1 & 0 \\ 0 & 0.5 & 0 & 0 & 1 & 0 & 0 & 0.5 & 1 \\ 0 & 0 & 0 & 0.2 & 0 & 1 & 0 & 0.5 & 0 \\ 0.9 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0.9 \end{bmatrix}$ is obtained (i.e. final weight parameters). Bias $\vec{b}$ is not considered. Logit z for each class j is given by $z_j = \vec{M} \cdot \vec{x}$, where input $\vec{x}$ is an image (e.g. Y, L, O and X) expressed in vector form (i.e. all pixel values are ordered from top-left to bottom right). Compute output y_i using a softmax layer i.e. $y_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$ for images O and X. What can you conclude about the prediction? (3 pts)

a: Compute prediction z_1 to z_4
for each of the three image patches
then pull them through the softmax layer
then check which letter
has the highest pred prob.
Try to interpret that result

1. conv layer 2. maxpool layer. 3 fc layer.
4. softmax layer. 5. non-linearity?

- (b) Describe the 5 different main layers used in a Convolutional Neural Network. (1 pts)

conv layer convolves over the image via kernel.
FC is just a linear projection matrix

- (c) Describe the differences between a Conv Layer and FC layer. (1 pts)
- (d) Assume the input of a particular Conv Layer is 100x100x25. 100 filters with a receptive field of 5x5 are learned. What is the total number of parameters to be learned in this layer? (1 pts)

kernel_w * kernel_w * channels * filters
= 5 x 5 x 25 x 100, + 100 biases

- (e) The next Conv Layer learns also 100 filters with a receptive field of 5 x 5. How many parameters are learned in this layer? (1 pts)

kernel_w * kernel_w * channels * filters
= 5 x 5 x 100 x 100, + 100 biases

Consider the following input to a max-pooling layer: $\begin{matrix} 1 & 1 & 2 & 4 \\ 5 & 6 & 7 & 8 \\ 3 & 2 & 1 & 0 \\ 1 & 2 & 3 & 4 \end{matrix}$.

- (f) What is the output of the max-pooling layer with a 2x2 window, with stride 2? (1 pts)

[[6, 8],
[3, 4]]

ranking for (h):
3, 5, 4, 6, 7, 2, 1
t, t, f, f, t, t, f

t = true, f = false

1*(1/1) + 1*(2/2) + 0*(2/3) + 0*(2/4) + 1*(3/5) + 1*(4/6) + 0*(4/7)
= 3.26666667

Then AP = 3.26666667 / (no. relevant documents retrieved) = 0.816666668

- (g) What is the gradient from this layer to a previous layer? (1 pts)

For retrieval, consider the following ranking:

Document id	1	2	3	4	5	6	7
Ground truth label	0	1	1	0	1	0	1
Score	0.1	0.3	0.9	0.7	0.8	0.5	0.4

- (h) Compute Average Precision for the scoring of the documents above. (2 pts)
- (i) Describe how a 10 class classification system, either based on SVMs or on DeepNets, could be evaluated using Average Precision. (1 pts)

Question 4: Deep Video

We could compute AP for 10 relevance rankings, one ranking for each class, after which we could take the average over AP scores to obtain a 'global' evaluation metric

Suppose you have a 2D convolutional layer which receives an image input of 224x224x3 and generates feature maps of 224x224x64 using 5x5 filters. You want to extend the architecture to receive a video of size 224x224x3x8 using a 3D filter of 5x5x3.

- (a) How many parameters are added to the model? Suppose all the Conv Layers have a bias. (1 pts)
- (b) Compute the computational cost for both the image and video model. (1 pts)
- (c) Propose and draw an architecture for video generation from a sentence (input:sentence, output: video). Justify your architectural choices. (1 pts)
- (d) Suppose you have a zero-bias 2x2x2 kernel with parameters

$$K_{t=1} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} K_{t=2} = \begin{bmatrix} 0 & -1 \\ 1 & 2 \end{bmatrix}.$$

Compute the output feature maps for the following input data

$$I_{t=1} = \begin{bmatrix} 4 & 4 \\ 3 & 0 \end{bmatrix} I_{t=2} = \begin{bmatrix} 3 & 4 \\ 2 & 3 \end{bmatrix} I_{t=3} = \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix}. \text{ (2 pts)}$$

- (e) What is a drawback of a standard RNN? How does LSTM address this problem? (1 pts)
- (f) What are the pros and cons of weight sharing? (1 pts)
- (g) What is self-supervised learning paradigm? When can it be useful? Explain with one example. (2 pts)