Indian Institute of Technology Jodhpur
CSL7360: Computer Vision, Major Exam
Date: May 10, 2024, Max Marks: 60 Max Time: 120 minutes

1. (10 points) A camera is rigidly mounted so that it views a planar table top. A projector is also rigidly mounted above the table and projects a narrow beam of light onto the table, which is visible as a point in the image of the table top. The height of the table top is precisely controllable but otherwise the positions of the camera, projector, and table are unknown. For each of the following table top heights, the point of light on the table is detected at the following image pixel coordinates:

Table Height	Image coordinates of beam of light
50mm	(100,250)
100mm	(140,340)

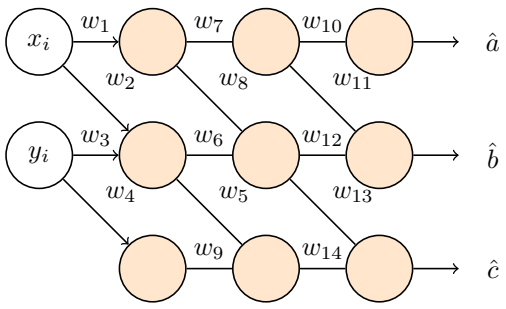
   (a) Using a projective camera model (described at the end of the question) specialized for this particular scenario, write a general formula that describes the relationship between world coordinates (x), specifying the height of the table top, and image coordinates (u, v), specifying the pixel coordinates where the point of light is detected. Give your answer using homogeneous coordinates and a projection matrix containing variables.

   (b) Once the camera is calibrated, given a new unknown height of the table and an associated image, can the height of the table be uniquely solved for? If so, give the equation(s) that is/are used. If not, describe briefly why not.

   A generalized projective camera model is described as follows. Let $\mathbf{X} \in \mathbb{R}^n$ be a scene point in and $\mathbf{x} \in \mathbb{R}^m$ be the projected point in the image plane. Then, the projective camera model in the homogeneous coordinate system is described by the equation $\begin{bmatrix} \lambda\mathbf{x} \\ \lambda \end{bmatrix} = \mathbf{M} \begin{bmatrix} \mathbf{X} \\ 1 \end{bmatrix}$. Here, λ is the depth of the scene points from the camera center and $\mathbf{M} \in \mathbb{R}^{(m+1)\times(n+1)}$ represents the projective camera matrix.

2. (10 points) Consider a set if 10 images of a static scene captured by an orthographic camera from 10 different viewpoints. Consider 20 3D points $\{\mathbf{P}_i\}_{i=1}^{20}$ in the scene are being projected in the 10 images. Let $(r_{a,b}, s_{a,b})$ be the projection of the b-th 3D point $\mathbf{P}_b$ in the a-th image. Assume that the $\{(r_{a,b}, s_{a,b})\}$ are given to you. where, $b = 1, \ldots 20$, and $a = 1, \ldots, 10$. Now, design an algorithm to find the pose of the each camera and the locations of all the 3D points $\{\mathbf{P}_i\}_{i=1}^{20}$ with respect to a world coordinate system whose origin is centered at the 3D point $2\mathbf{P}_3$.

3. (10 points) Consider a set of keypoints $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ detected in an image around a linear edge represented as a line $ax + by + c = 0$. Now in order to find the optimal parameters (a, b, c) of this line, a student has designed a neural network shown as below. Each neuron in the first and the second layers has sigmoid as the activation layer and the last layer neurons do not have any activation function. Design a loss function (call it ℓ) to train this neural network. Also, find $\frac{d\ell}{dw_7}$ and $\frac{d\ell}{dw_4}$.



4. Answer the following questions.

   (a) (3 points) What is the condition in the spectral clustering of n pixels for which

   $$\text{Trace}(\mathbf{H}^\top \mathbf{H}\mathbf{L}) = \text{RatioCut}(\mathcal{C}_1, \ldots, \mathcal{C}_k)$$

   and prove it mathematically.

   (b) (7 points) Perform only a single maximization (not expectation) step and find the updated cluster parameters, given the following status in a GMM clustering: number of clusters $k = 2$, the points being clustered are $p_1, p_2, p_3 = \{(1,2),(4,2),(5,1)\}$, current cluster centers are $c_1, c_2 = \{(2,2),(6,2)\}$ and

	p_1	p_2	p_3
z_{i1}	0.8	0.8	0.4
z_{i2}	0.2	0.2	0.6

   , where z_{nk} denotes the probability of the nth point belonging to the kth cluster.

5. Assume there is a convolutional neural network containing 3 convolutional layers c_1, c_2, c_3 containing 1 filter each of size 3×3, 3×3, and 3×3, respectively, with stride 1 and no padding. The output o_1 produced by c_1 is fed to c_2 and the output o_2 produced by c_2 is fed to c_3 which produces the final output o_3.

   (a) (2 points) If the input is an image of width 7 pixels and height 7 pixels, what is the final output size in pixels

   (b) (2 points) Take any point in the intermediate output o_2 and find the size (width and height) of the exact region in the original 7×7 image that influences the value of that point.

   (c) (2 points) What is the total number of learnable weights and biases in this convolutional neural network?

   (d) (2 points) Design a fully connected layer to take the original 7×7 image as input and produce an output o_{fc} of the same size as o_3. Fully connected layer does not directly support a 2 dimensional input. So you will have to modify the input for passing it to a fully connected layer. How many neurons does this layer have?

   (e) (2 points) What is the number of learnable weights and biases in the fully connected layer designed in part (d)?

6. Answer the following questions

   (a) (2 points) Create a separable 2D square filter that can convert an integral image into a normal image through the correlation operation.

   (b) (2 points) What are the constituent 1D filters for the above filter?

   (c) (3 points) Apply this filter to the following integral image $I_{\text{integral}} = \begin{bmatrix} 1 & 6 & 13 & 16 \\ 3 & 13 & 24 & 29 \\ 6 & 22 & 35 & 41 \end{bmatrix}$ and find the original image which has the same size. Also, mention the amount of padding, if needed.

   (d) (3 points) Arrange 3 Gaussian filters with standard deviations σ, 2σ, and 3σ in the decreasing order of the range of high-frequency components that get filtered out by them. Explain why?