CS 558 Final Exam
Date: December 17th, 2020
Duration: 3:00 hours

NAME:

Problem 1. (25 points) Bag of Words/Spatial Pyramid. Consider the following pixel grids

E			A				A
		B			A		
D			C				D
			B			C	
	A	B					B
			C	D			E
D		A			B		A

Image A

A							
		B			D		
							C
				E			
	B		C				
						D	
	E			D			

Image B

a) (10 Points) Write the bag of words (BOW) representation for each image, assuming the set of visual words is formed by the letters A through E. Instead of using raw counts, normalize each histogram so that the sum of all its bins is equal to 1. Compute the similarity of both BOW representations using the histogram intersection distance

b) (5 Points) How many TOTAL bins are needed to construct a spatial pyramid with ONE level of recursive partitioning (i.e. four quadrants) for image A?.

c) (10 Points) Write the 1D representation (e.g. histogram) of the spatial pyramid with ONE level of recursive partitioning (i.e. four quadrants) for image A (No normalization needed).

Problem 2. (10 points) Stereo. Consider a rectified stereo pair with baseline of 10 cm and a common focal length of 2.5 pixels. Additionally, a pixel A at column 30 in the left image has two candidate correspondences in the right image; the first (B1) at pixel at column 25, the second (B2) at pixel column 28. Given the known relationship $Z = bf/d$, answer the following

a) (5 Points) What is the depth difference between the two possible correspondences B1 and B2?

b) (5 Points) Is the 3-D Euclidean distance between the 3-D points associated with B1 and B2 equal to their depth difference? Justify and make a sketch.

Problem 3. (15 points) Integral images. Consider the following grid of pixel intensity values

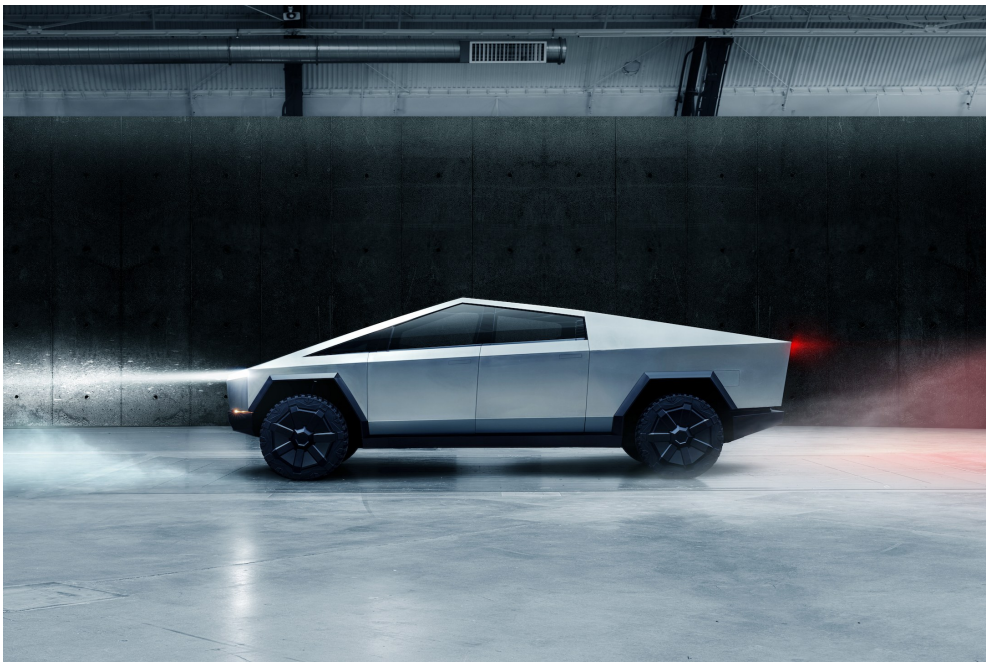
1	0	2	2
2	1	2	1
1	3	2	1
2	2	1	3

a) (10 Points) Compute the integral image

b) (5 Points) Assume you have computed the pixel-wise absolute difference between images I_a and I_b, such that $I_D(u, v) = abs(I_a(u, v) - I_b(u, v))$. If you want to aggregate absolute differences over all 9×9 local windows in I_D, how is the computational burden reduced by using integral images vs. naively scanning and adding up over each local window centered at every image pixel?

Problem 4. (20 points) Robust Estimation Assume you want to track 2D the position of a moving car across a short video sequence. Assume the video is captured in "sideview" (e.g orthogonally observing the drivers door but with a wide enough field of view to capture the entire vehicle) similar to the image below. Describe how you would address the problem. In order to make the method invariant to car types, you may optionally choose to detect the outline of the car's tires. Your description should include:

- what are your input and output
- what features/representation are to be used.
- what estimation framework would you use and how to determine it parameters
- How would you handle/leverage temporal dependencies in the video sequence



Problem 5. (15 points) Clustering Assume you are clustering N points using a naive mean shift clustering where 1) the kernel is a spherical one (i.e. fixed radius without weighting) 2) each data point is explicitly evaluated as a starting point and mean shift executed until convergence and 3) all points sharing a common final mean are assigned the same cluster. Further, assume you have a $D_{N\times N}$ symmetric pairwise Euclidean's distance matrix.

a) (7.5 Points) How many clusters do you expect to find using mean shift if you set the kernel radius r to be slightly smaller than the minimum value stored in $D_{N\times N}$, w/o considering the values stored in the diagonal . That is, $r = (1-\epsilon) \times min\left(D_{N\times N} \setminus diag(D_{N\times N})\right)$.

b) (7.5 Points) How many clusters do you expect to find using mean shift if you set the kernel radius equal to the largest value stored in $D_{N\times N}$. Justify your answer

Problem 6. (15 points) Attentional Cascade Consider an attentional cascade (such as the one used in the Viola-Jones face detector) consisting of N stages, each of which has perfect 100% detection rate and a false positive rate of 50%. Furthermore, each stage increases the number of features used by a factor of $10\times$.

n	2^n	n	2^n	n	2^n
0	1	11	2,048	22	4,194,304
1	2	12	4,096	23	8,388,608
2	4	13	8,192	24	16,777,216
3	8	14	16,384	25	33,554,432
4	16	15	32,768	26	67,108,864
5	32	16	65,536	27	134,217,728
6	64	17	131,072	28	268,435,456
7	128	18	262,144	29	536,870,912
8	256	19	524,288	30	1,073,741,824
9	512	20	1,048,576	31	2,147,483,648
10	1,024	21	2,097,152	32	4,254,967,296

a) (7.5 Points) What is the minimum number of stages N that are needed to achieve a false positive rate less than 1 in 1000? Justify your answer.

b) (7.5 Points) Assuming the first classifier in the cascade A uses 2 features and the number of stages is $N = 4$. And that the growth between stages is the same as before. Compared to a single (i.e. non-cascaded) strong classifier B using 1000 a total of features and trained to achieve the same false positive rate, which is more efficient? Justify your answer.

Problem 7. (15 points) Consider three different models for shape recognition

1. a star shaped model in which each part is a visual word,
2. a constellation model with the same parts, and
3. a bag of words model.

The edge lengths and orientations in both graph models are fixed. Which model do you expect to lead to more false positives and which do you expect to lead to more missed detections? Why?

Problem 8. (15 points) In a conventional image recognition approach, one would design a feature descriptor (e.g. SIFT), compute a dictionary of visual words, and then train classifiers using the bag of words technique for multiple tasks, such as distinguishing between horizontal and vertical surfaces or between indoor and outdoor scenes. Explain what is the disadvantage of this approach compared to deep learning.