Midsem Solutions

Q1

Assigned TA: Sanyam Shah

Sample Answer

a) Similarity, Proximity, Continuity, Closure, Symmetry, Common Fate, Emergence (Dalmatian/Unilever), Multistability (Ambiguous Images), Figure Ground, Invariance, Pragnanz (human tendency to simplify complexity), Common Fate OR describe a scenario in our daily lives where this is seen
AND
Significance of that Gestalt principle in CV task (Proximity $\longrightarrow$ Clustering algos, Similarity $\longrightarrow$ Pattern recognition, Continuity $\longrightarrow$ Edge detection, object tracking, Common Fate $\longrightarrow$ vehicle navigation, object tracking, surveillance, Figure background $\longrightarrow$ Separation of foreground from background)
b) Too big $\longrightarrow$ Reduced sharpness, Increased Intensity, Out of focus (Lower depth of field)
Too small $\longrightarrow$ Diffraction (leads to decrease in resolution/sharpness), Reduced light intensity, More depth of field

Rubric

1. Part A
   - Correct Explanation for any 2 Gestalt Principles [**0.5**]
   - Explanation for significance in CV respectively [**0.5**]
2. Part B
   - Size of pinhole is too big [**0.5**]
   - Size of pinhole is too small [**0.5**]

Q2

Assigned TA: Sreenya Chitluri

Sample Answer

a)

1. Taken the greyscaled image.
2. Pick a value based on the intensity histograms that separates two peaks.
3. Threshold it based on this value.

This would fail when the image has varying illuminations, complex backgrounds, noise, overlapping peaks in the intensity histograms etc.
b)

1. Computational complexity - NP Hard
2. Dependency on user input - need to indicate what is foreground and background in the image.

Rubric

1. Part A

- Any simple thresholding technique with the explanation. **[0.5]**
- Any example or explanation of when it might fail. **[0.5]**

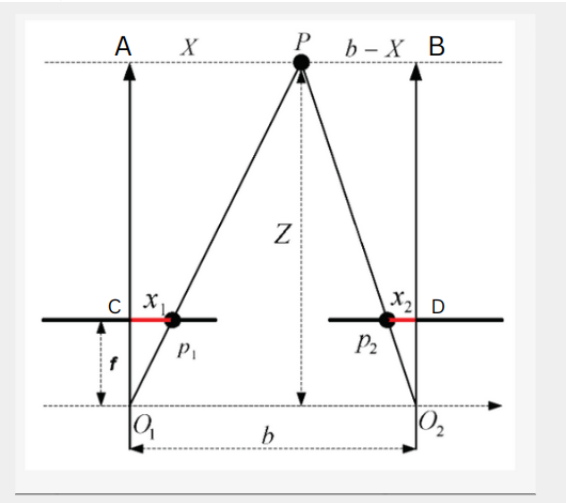
2. Part B
Each valid drawback of graphcut segmentation with minimal explanation **0.5**

Q3

Assigned TA: Mandyam Brunda

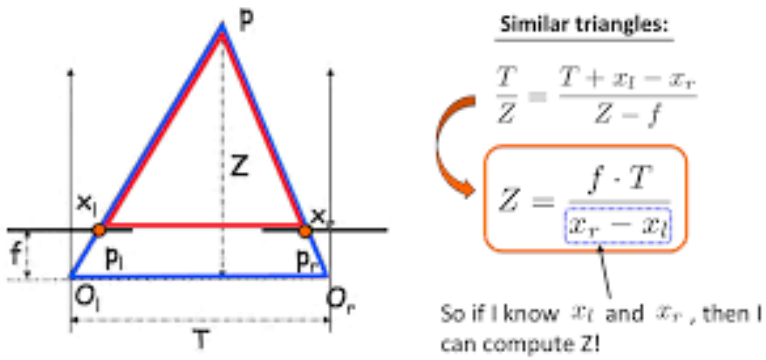
Sample Answer

3.1 , 3.2:



O1Cp1 ,O1AP and O2Dp2 , O2BP are similar By using similar triangles,

$$\frac{x_1}{f} = \frac{X}{Z}$$
$$\frac{x_2}{f} = \frac{b-X}{Z}$$
$$\frac{x_1+x_2}{f} = \frac{b}{Z}$$
$$\text{disparity} = \frac{bf}{Z}$$
$$\text{(or)}$$



3.3:
Epipoles are at infinity or does not lie on the image plane (When the image planes are parallel to each other, then the epipoles will be located at infinity since the baseline joining the centers is parallel to the image planes.

3.4:

1. Cameras very near
2. Object is very far or Object is very near
3. baseline large
4. object is in the field of view of only one camera
5. Getting corresponding points is difficult in case of textureless walls, repititive patterns, foreshortening

Rubric

3.2 If direct final equation is written only one-fourth of the marks were given.
3.3 Marks are not given if the answer written is epipoles lie on the intersection of baseline and image planes.
3.4 Answers that presume baseline, focal length, or camera intrinsics are not given will not be awarded marks and any relevant answer with proper explanation is considered

Q4 [3 Points]

Assigned TA: Mohd Hozaifa Khan

Sample Answer

1. Assumptions: Book visible in the image; We have an image of the book.
2. Rationale for Using SIFT: SIFT is scale, rotation, and illumination invariant. Moreover, features are local.
3. Main Steps of SIFT Detector: Scale Space (DOG, Laplacian), KP Localization, Orientation Assignment.
4. Main Steps of SIFT Descriptor (HoG)
5. Matching Method: Nearest Neighbour, L1 or L2, etc.

Rubric

1. Correct Answer: SIFT
   - Reasonable assumptions stated [**0.25**]
   - SIFT [**0.5**]
   - Detailed explanation of detector and descriptor [**2**]
   - Mentioned a matching/comparing method [**0.25**]
2. Other methods – Evaluated of **2 Points**. You must make reasonable assumptions and then explain the steps clearly. Simply writing "a transformer", "a segmentation model", or "SAM/DINO" wouldn't be considered.

Q5 [3 points]

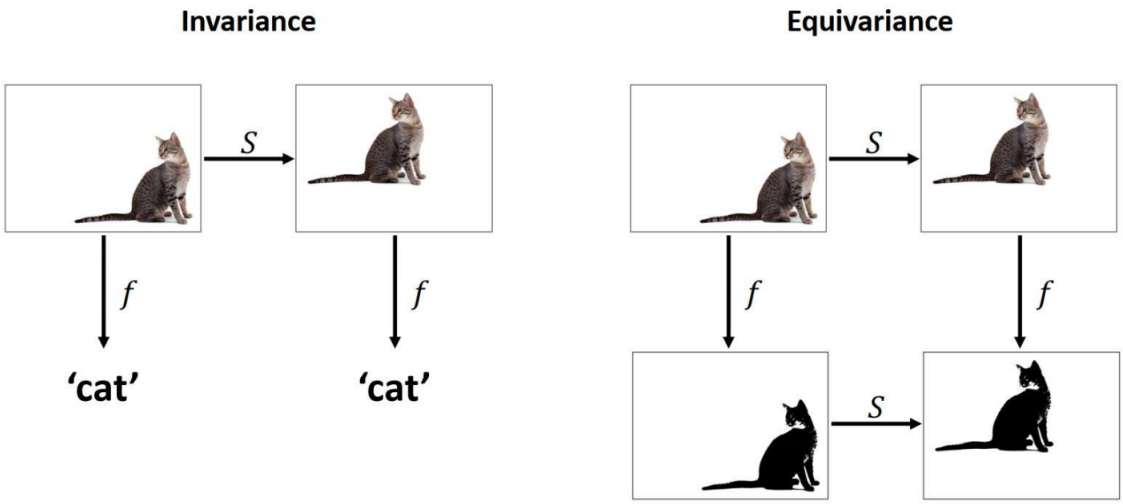
Assigned TA: Shreya

Sample Answer

5.1: In the space of all filters with the same L2 norm, the best filter to detect the pattern would be the pattern itself. i.e. $k \cdot \begin{bmatrix} 0.8 & 1 & 0.8 \\ 0.8 & 1 & 0.8 \\ 0.8 & 1 & 0.8 \end{bmatrix}$
The filter is symmetric, hence it doesn't matter if it's flipped.

5.2: Translation invariance refers to the property of a model where the output remains the same even when the input is shifted or translated spatially. In other words, if the object of interest in an image is moved to a different location, a translation-invariant model should still recognize it. Eg: A cat image classifier should detect the label 'cat' regardless of the position of the cat in an image.
Translation equivariance is a slightly weaker property than translation invariance. It

means that the output of the model changes predictably with the input's translation, maintaining a consistent relationship. i.e. if the input image is translated, the output of a translation-equivariant model will also be translated accordingly. For example, if an object moves a certain distance in the input, the model's output predictions should also shift by the same amount.
Example:



5.3: ResNet introduced residual connections, also known as skip connections, which allowed information from earlier layers to bypass certain layers and be directly fed into deeper layers. This was accomplished by adding the input of a layer to its output, effectively creating shortcuts within the network. In deep networks, as gradients are back-propagated from the output layer to the input layer during training, they may become increasingly small (approaching zero) as they are propagated through many layers. This is called Vanishing gradients and can significantly impede the training process and hinder the convergence of the model to an optimal solution. Skip connections allow for easier gradient propagation to deeper layers mitigating vanishing gradients.

Rubric

5.1 Answers like $\begin{bmatrix}0.8 & 1 & 0.8\end{bmatrix}$, $\begin{bmatrix}0 & 0.8 & 1 & 0.8 & 0\end{bmatrix}$, $\begin{bmatrix}0.8 & 1 & 0.8\\ 0.8 & 1 & 0.8\\ 0.8 & 1 & 0.8\end{bmatrix}$, $\begin{bmatrix}0 & 0.8 & 1 & 0.8 & 0\\ 0 & 0.8 & 1 & 0.8 & 0\\ 0 & 0.8 & 1 & 0.8 & 0\end{bmatrix}$
were given a full score.
0.5 marks were given for writing general edge detection filters like sobel or laplacian.

5.2: 0.5 marks for correct definitions of invariance and equivariance. And 0.5 for explanation with an example.

Q6

Assigned TA: Shreyash Jain

Sample Answer

6a): Loss of information when large sentences data is compressed into a fixed size vector. Network may misinterpret or forget earlier parts of the sequence as the size increases. Since the solution is not generalizable, increasing the size won't fix the problem.

6b): Spiky attention of scores at the start of training indicates a bias towards certain features of the input, while ignoring the other features. This will lead to poor results and more training time.
Fixes:

- Normalizing the attention scores (layer normalization, softmax temperature tweaking).
- Re-initializing weights using better techniques.

6c): Positional Embedding

Rubric

6a):

- Loss of contextual info, forgetting previous parts of the sentence [**0.5**]
- No, increasing size won't help [**0.25**]
- Why increasing the size won't fix the problem [**0.25**]

Some marks have been given for mentioning the idea of attention or the problem of vanishing gradients.
Marks have been cut (0.25) if the reasoning is too vague, just mentioning loss of information without a detailed explanation won't fetch you marks.
No marks have been given if you have just mentioned that increasing the size of vector won't solve the problem without mentioning what the problem is.

6b):

- Why it is a problem?
- Fix 1
- Fix 2

Mentioning both fixes but not why it is a problem [**0.75**]
Mentioning only one fix and why it is a problem [**0.75**]
Mentioning only one fix [**0.5**]
Mentioning only why it is a problem [**0.25**]

6c): Simply mentioning positional encoding/embedding or even explaining the idea of adding positional information along with the image patches in the model has fetched you

1 full mark