CV1 practice exam 2

52041COV6Y Computer Vision 1 24/25 (1.1) · 4 exercises · 43.5 points

1 Question 1: Low Level Vision

10.0 points · 11 questions

\vspace{0.5cm}
Camera Model

Text

Figure 1.1:

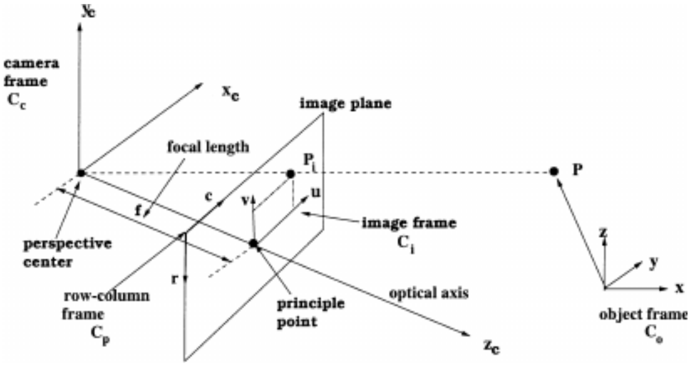


Figure 1.1 illustrates the projection from 3D space to 2D image using a pin-hole camera. The equation can be written as: \vspace{0.5cm}

$$\mathbf{x} = \mathbf{K}[\mathbf{R} \quad \mathbf{t}]\mathbf{X}$$

where **X** is the coordinates of the 3D point in homogeneous coordinates and **x** is the homogeneous coordinates on the 2D image.

Text

a What is the name of matrix **K**?

0.5 points · Multiple choice · 5 alternatives

●	**Intrinsic matrix**	0.5
○	Extrinsic matrix	0.0
○	Rotation matrix	0.0
○	Translation matrix	0.0
○	Projection matrix	0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

b Provide the expanded form of **x** and **X** :

1.0 point · Open question with final answer · 1/4 Page · 0 answers

+0.5 points

$x = w[u, v, 1]^T$

Use other symbols are fine, as long as indicates homogeneous coordinates

+0.5 points

$\mathbf{X} = [x, y, z, 1]^T$

Use other symbols are fine, as long as indicates homogeneous coordinates c The matrix K takes the form of a 3 by 3 matrix. Four elements have been provided. Can you provide the missing 5 elements for a pinhole camera model supporting non-square sensor pixels and a skew parameter:

$$K = \begin{bmatrix} \square & \square & \square \\ 0 & \square & \square \\ 0 & 0 & 1 \end{bmatrix}$$

Also, provide the meaning of the 3 elements in the first row in their correct order:

2.0 points · Open question with final answer · 1/100 Page · 0 answers

+0.5 points

fill in the symbols correctly. Use other symbols are acceptable if correctly explaining the meaning.

K =

$$\begin{bmatrix} f_x & s & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

+0.5 points

f_x is the focal length in x direction
or
f_x is the scaling in x direction

+0.5 points

s is the skew

+0.5 points

u_0 is the displacement/translation in x direction d Given the current form of K, can we perform an arbitrary rotation using K? If yes explain how, if not explain why?

0.5 points · Open · 1/5 Page

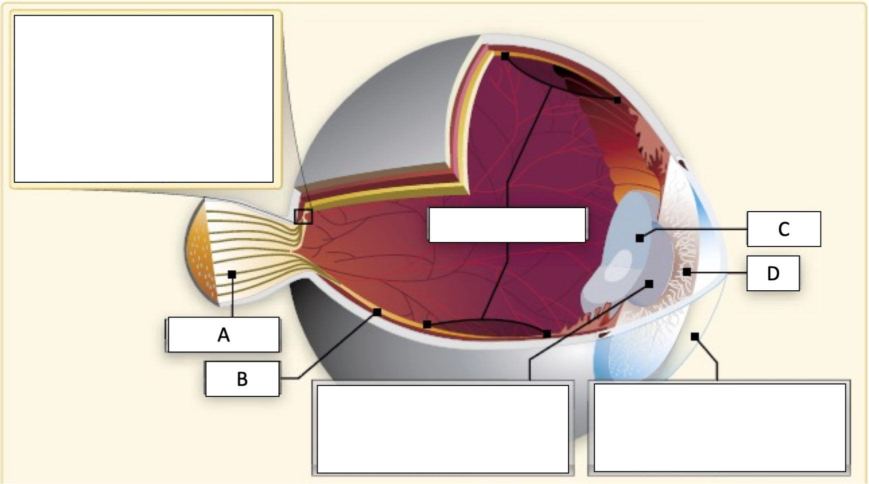
+0.5 points

the 2nd row, 1st column of K has been fixed to be zero, and therefore cannot perform arbitrary rotation

\vspace{4.5cm}
Human Vision and Color

Text

e This is an anatomy of human eye.



Which one is the retina?

0.5 points · Multiple choice · 5 alternatives

○	A	0.0
●	B	0.5
○	C	0.0
○	D	0.0
○	None of them	0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

f Here are two statements regarding human perception. Which are correct?
A. There are more rods than cones on the human retina.
B. Rods are more sensitive in low light (darkness) than cones.

0.5 points · Multiple choice · 4 alternatives

○	only A	0.0
○	only B	0.0
●	**Both of them**	0.5
○	None of them	0.0

Feedback

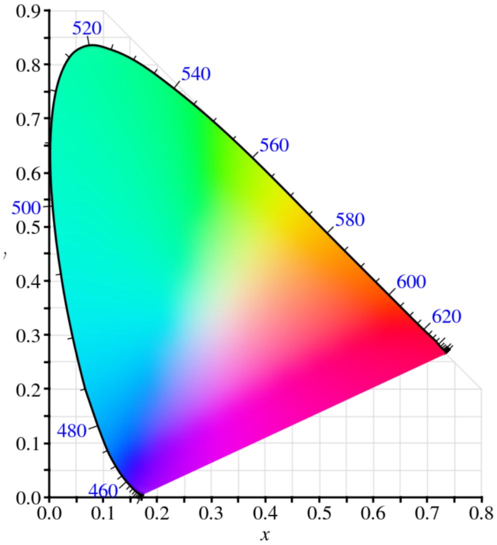
Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

g CIE systems are commonly used to study color. Assume the sunlight (ideal white) has CIE values $X_S = Y_S = Z_S = 100$. Further, let $X_A = 100$, $Y_A = 300$ and $Z_A = 100$ be the values for a given artificial lamp A. Calculate the chromaticity values *x,y* for both S and A and plot them on the CIE-xy chart in figure 1.2 (0.5pt). (use pencil in case of correction)

Figure 1.2:



Use your plot and calculate the hue (0.5pt) and saturation (0.5pt) of A

1.5 points · Open · 7/20 Page

+0.5 points

Sx = 1/3 = 0.33, Sy = 1/3 = 0.33, Ax = 0.2, Ay = 0.6

+0.5 points

Hue is approximately 517nm, because the result is hand plotted, it doesn't need to be very accurate, but need to show the way of doing it.

+0.5 points

Saturation is approximately 60%

\vspace{0.4cm}

Reflection Model

Text

Figure 1.3:

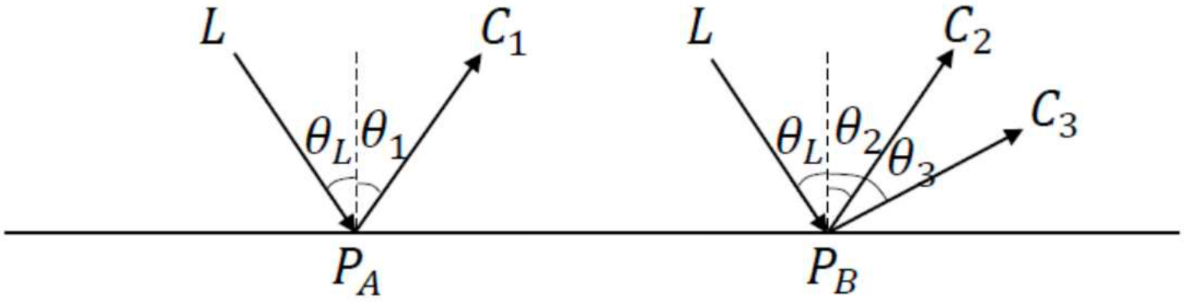


Figure 1.3 shows an image of a flat plane. L is the only light source, which is a uniform and parallel light pointing to the image plane. There is no other light source or object in the setting. There are two points P_A and P_B. And they are observed by three cameras C_1, C_2 and C_3. The angle between light source and surface normal is θ_L, and the angle between camera and surface normal are θ_1, θ_2 and θ_3. The three cameras are identical except there positions.

Text

h Denote the reflected light intensity to C_1 from point P_A as I_1 provide the simplified *Lambertian* reflection model. Please provide an equation to explain how I_1 can be determined. Give extra symbols and explain them if necessary.

0.5 points · Open · 3/10 Page

+0.5 points

$L = a\rho LN$, a is a coefficient, missing it is fine. ρ is the albedo, use other symbols with proper explanation is fine. N is the surface normal and L is the light source i Using the conditions provided in Fig. 1.3 and its descriptions but not the conditions in other questions.

In this question, assume the plane material is *Lambertian* (ideally diffusing).
We know the intensity observed by camera 1 is $I_1 = 300$.

Assume $\theta_1 = \frac{\pi}{6}, \theta_2 = \frac{\pi}{6}$ and $\theta_3 = \frac{\pi}{3}$.
Is the information provided so far sufficient to derive the intensity I_2 and I_3 captured by C_2 and C_3. If so, provide the value with derivation. If not, explain why. Use illustrations if necessary.

1.0 point · Open · 3/10 Page

+0.5 points

Lp $\propto$ ρ **LN ,** use '=' instead of proportional is fine. Normal is assumed to be normalized, add extra normalization is fine. **LN** is inner product between to vectors and is commutable. Use $LNcos\theta$ where θ is the angle between L, N is also fine.

+1 point

not possible, because the material is not uniform, the albedo is not determined.

j Using the conditions provided in Fig. 1.3 and its descriptions but not the conditions in other questions.

In this question, assume the plane material is *glossy*, which works like a mirror (ideally diffusing).
We know the intensity observed by the camera 1 is $I_1 = 200$.

Assume $\theta_1 = \frac{\pi}{6}, \theta_2 = \frac{\pi}{6}$ and $\theta_3 = \frac{\pi}{3}$.
Is the information provided so far sufficient to derive the intensity I_2 and I_3 captured by C_2 and C_3. If so, provide the value with derivation. If not, explain why. Use illustrations if necessary.

1.0 point · Open · 2/5 Page

+0.5 points

If assume the material is perfectly reflective then C2 has the same situation as C1 and therefore I2 = I1 = 200.

+0.5 points

I3 = 0, because of the law of reflection.

k Using the conditions provided in Fig. 1.3 and its descriptions but not the conditions in other questions.
In this question, assume the plane material is *Lambertian* (ideally diffusing) and *uniform*,

We know the intensity observed by the camera 1 is $I_1 = 100$.

Assume $\theta_1 = \frac{\pi}{6}, \theta_2 = \frac{\pi}{6}$ and $\theta_3 = \frac{\pi}{3}$.
Is the information provided so far sufficient to derive the intensity I_2 and I_3 captured by C_2 and C_3. If so, provide the value with derivation. If not, explain why. Use illustrations if necessary.

1.0 point · Open · 2/5 Page

+0.5 points

I2 = I1 = 100

+1 point

+0.5 points

I3 = I1 = 100

2 Question 2: Image Processing

23.5 points · 18 questions

\vspace{0.5cm}

In Place Processing and Morphology

Consider the following image patches:

$$P = \begin{array}{|c|c|c|c|}\hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 \\ \hline \end{array} \qquad Q = \begin{array}{|c|c|c|c|}\hline 0 & 0 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 \\ \hline 0 & 0 & 1 & 1 \\ \hline 0 & 0 & 0 & 0 \\ \hline \end{array}$$

Text

a Lets do some binary morphology. Given a construction element

$$U = \begin{array}{|c|c|c|}\hline 0 & 1 & 0 \\ \hline 1 & 1 & 1 \\ \hline 0 & 1 & 0 \\ \hline \end{array}$$

Compute the results after performing *dilation* and *corrosion* separately on image *P* by using template *U*.
All elements at the image boundaries (i.e. outside image *P*) are mirrored. The elements outside filter *U* are all zeros.
P after dilation

P after corrosion

2.0 points · Free formatted question

+1 point

no mistakes. Give only the foreground is fine
dilation:

0	0	0	0
1	1	1	1
1	1	1	1
1	1	1	1

1p for correct answer, each mistake -0.5p

+1 point

erosion
0 0 0 0
0 0 0 0
0 0 0 0
1 1 1 1
1p for correct answer, each mistake -0.5p

b Show the filtering result on *Q* by first do a *corrosion* and then do a *dilation* using *U*. Use zero-padding for both *U* and *Q*.

1.0 point · Open · 0/1 Page

+1 point

0	0	1	0
0	1	1	1
0	0	1	0
0	0	0	0

each mistake -0.5p till 0

\vspace{0.5cm} Image Filtering

Text

c We have a sequence

s = [0, 0, -2, 0, 0, 0, 0, 1, 1, 1].

Show the self correlation result by first applying a difference filter h = [1, 0, -1]. Then apply the Gaussian filter g = [1, 2, 1] as defined in the previous question. Use zero padding.

2.0 points · Open question with final answer · 1/2 Page · 0 answers

+2 points

first step: h * s = [0, 2, 0, -2, 0, 0, -1, -1, 0, 1];
correct answer got 1pt, each mistake -0.5pt.

second step:
g*h*s = [2, 4, 0, -4, -2, -1, -3, -3, 0, 2]
correct answer got 1pt, each mistake -0.5pt.

d Following the above question, explain what do these operations do? Why do we need to combine a Gaussian filter with a differential filter?

1.0 point · Open question with final answer · 1/5 Page · 0 answers

+0.5 points

this is an edge detector / band pass filter.

+0.5 points

use h only is sensitive to noise, and therefore, a smoother g is applied.

e Explain the concept of a Bilateral filter:

0.5 points · Open question with final answer · 1/5 Page · 0 answers

+0.5 points

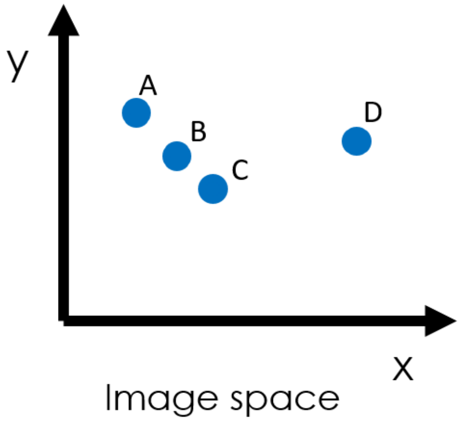
A bilateral filter is **a non-linear, edge-preserving, and noise-reducing smoothing filter for images**. It replaces the intensity of each pixel with a weighted average of intensity values from nearby pixels.

Point will be granted if mention edge/structure preserving.

Edges and Lines

Text

f Figure 2.1:



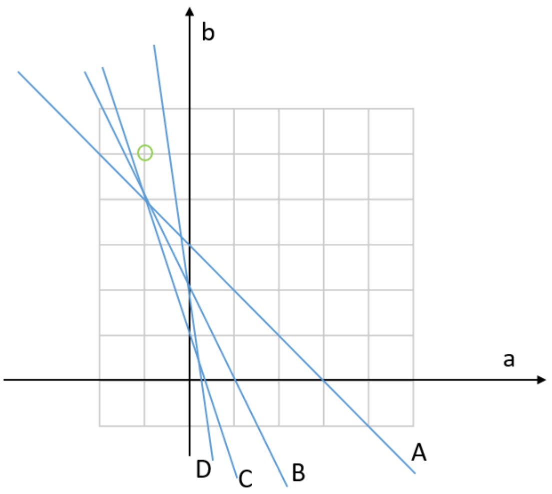
In figure 2.1, we have 4 points in an image. The coordinates are A = (A_x, A_y) = (1, 3), B = (2, 2), C = (3, 1), D = (5, 2).
Show how you can fit a line using Hough transform. For simplicity, use the linear version: y=ax+b.
To get the full points, all steps of Hough transform need to be provided.

2.0 points · Open · 1/2 Page

+2 points
step 1 Give the four lines in Hough space (1pt)
A: a + b = 3
B: 2a + b = 2
C: 3a + b = 1
D: 5a + b = 2
or equivalent form

Each wrong equation deduct 0.5pt, up to 1pt. It is acceptable if lines are plotted correctly without equations.

Step 2:Plot them in hough space (0.5pt)
flip axis a and b is fine
mistake other than wrong equation will deduct point. wrong equation will result in deduction in step 1.



Step 3: find a= -1, b=4 (0.5pt) based on the voting in hough space, must mention voting in hough space.

+1.5 points

see grading scheme above

+1 point

see grading scheme above

+0.5 points

see grading scheme above

g Following the above question, this time, show how you can find the line using RANSAC.

1.5 points · Open · 2/5 Page

+0.5 points

mention random selecting two points.

+0.5 points

mention fitting a light based on two randomly selected points and calculate the confidence of this line.

+0.5 points

mention the line is find based on ranking the confidence

\vspace{0.5cm}
Corners

Text

h Harris corner detection is based on analyzing the second order differentials.
For the patch below:

0	0	1	1
1	1	1	1
0	0	1	1
0	0	0	0

compute its M matrix for the 2 * 2 pixels in the lower right.

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

To compute f_x use 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, idem for h_y. Use cross-correlation for simplicity. Handle the out-of-boundary pixels with mirroring. To save time, assume the window size for summation over the neighborhood Σ is 1x1, i.e. you can ignore the summation.

2.0 points · Open · 1/2 Page

+2 points

fx =
0 0
0 0

fy =
-1 -1
0 0

M33 =
0 0
0 1

M34 =
0 0
0 1

M43 =
0 0
0 0

M44 =
0 0
0 0

Each M matrix is worth 0.5pt

+1.5 points

see grading scheme above

+1 point

see grading scheme above

+0.5 points

see grading scheme above

i Name two **photometric** image transformations that the SIFT descriptor is invariant to.

1.0 point · Open question with final answer · 1/20 Page · 0 answers

+1 point

add/substract constant on brightness
scaling
any other reasonable answer. each worth 0.5p.

Rotation/translation are not photometric properties and are not considered correct .

+0.5 points

only one correct answer.

Optical Flow

Text

j Which statements about the Lucas-Kanade optical flow method are correct?

1.5 points · Multiple choice · 6 alternatives

☑	**The method is not robust to handle large motions (larger than the window size).**	0.5
☑	**The color or brightness of pixels is assumed to remain unchanged during motion.**	0.5
☑	**The method favors neighboring pixels to have the same flow vectors.**	0.5
□	The structure tensor is used to enforce local smoothness of the estimated flow field.	-0.5
□	The structure tensor is required to contain gradient information in at least one direction to uniquely estimate motion.	-0.5
☑	**The method fails to estimate a flow vector for homogeneously colored image regions.**	0.5

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

\vspace{0.5cm}
Linear Transformations

Text

k Different types of transformations have a different degree of freedom. What is the degree of freedom of a 3D rigid body transform $T \in SE(3)$?

1.0 point · Multiple choice · 4 alternatives

Model answer

Rigid body transformations are combinations of rotations and translations. These are described by the special Euclidean group.
In 2D space, that is [R|t] ∈ SE(2), there is 1 degree of freedom for rotation and 2 degrees of freedom for translation.
Hence the correct answer is 3 DOF.

○	2	0.0
○	3	0.0
○	4	0.0
●	6	1.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

l What is the degree of freedom of a 2D rotation $R \in SO\left(2\right)$?

1.0 point · Multiple choice · 4 alternatives

Model answer

Rotations in 3D have 3 degrees of freedom: one angle for the rotation around each axis in 3 dimensions.

●	1	1.0
○	2	0.0
○	3	0.0
○	6	0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

m **Transformation composition:** Consider a kinematic chain of a human skeletal model where joint locations are described relative to each other. In particular, the joint for the hand is described by point $H = (h_x, h_y, h_z) \in \Re^3$ in local coordinates relative to the elbow $E = (e_x, e_y, e_z) \in \Re^3$, while the elbow is defined relative to the shoulder $S = (s_x, s_y, s_z) \in \Re^3$, which in turn is relative to the pelvis $P = (p_x, p_y, p_z) \in \Re^3$ that acts as the main anchor point for the entire skeleton.
All body joints also have locally defined relative rotations $R_J \in SO(3)$ associated with them, where $J \in \{H, E, S, P\}$ is a place holder for any body joint. Therefore, coordinate transformations from a local joint coordinate frame into the corresponding parent coordinate are described by a rigid body transformation $T_J = [R_J | J] \in SE(3)$.
Select all correct statements from the following options (multiple correct answers possible):

1.0 point · Multiple choice · 5 alternatives

Model answer

The correct answer is:

$$T_{[a_x,a_y]} R_\theta T_{[-a_{x,}-a_y]} B$$

The main idea is to decompose the transformation into three parts:
1) translate everything such that point A is in the new coordinate origin, i.e. $T_{[-a_{x,}-a_y]}$.
2) rotate around the coordinate origin, i.e. apply R_θ
3) translate everything back, i.e. , i.e. $T_{[a_x,a_y]}$.
All transformations are multiplied together in right-to-left order.

☐ To obtain the position of the hand in the global coordinate frame one only reads out point H. -0.5

☑ **To obtain the position of the hand in the global coordinate frame one has to compute $T_P T_S T_E T_H \vec{0}$** 0.5

☐ To obtain the position of the hand in the global coordinate frame one has to compute $T_H T_E T_S T_P \vec{0}$ -0.5

☑ **To move the entire skeleton by a given displacement vector $D \in \Re^3$ one has to update the relative translation only for joint P.** 0.5

☐ To move the entire skeleton by a given displacement vector $D \in \Re^3$ one has to update the relative translation for all joints J. -0.5

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

A partially correct answer is not applicable here.

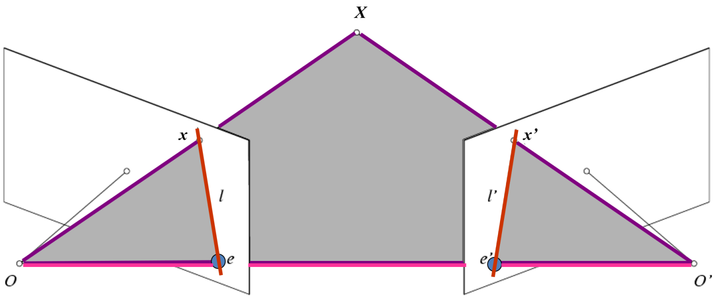
Feedback when the question is answered incorrectly

The main idea is to decompose the transformation into three parts:
1) translate everything such that point A is in the new coordinate origin, i.e. $T_{[-a_{x,}-a_y]}$.

2) rotate around the coordinate origin, i.e. apply R_θ
3) translate everything back, i.e. , i.e. $T_{[a_x,a_y]}$.
All transformations are multiplied together in right-to-left order.

\vspace{12.5cm}
Multiview Geometry and Reconstruction
Text

n The **epipolar geometry** is a key concept in multi-view stereo.



The figure above shows the camera centers O and O' and a 3D point X which projections on the image planes are x and x'.

Please mark correct all correct statements about epipolar geometry (multiple correct answers are possible).

2.0 points · Multiple choice · 5 alternatives

☑ The epipolar plane always contains the epipolar lines. 0.7

☑ The line through the points x and e is called epipolar line. 0.7

☑ The epipoles are always part of the the epipolar lines. 0.7

□ The epipolar lines are always orthogonal to the line defined by the points o and o'. -0.7

□ The line through the camera center o and the point X defines the principal axis. -0.7

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

Rectified Stereo:

1.0 point · Multiple choice · 4 alternatives

☑ Image rectification reduces the correspondence search between corresponding pixels in the two input images from a 2D to a 1D search problem. 0.5

☑ In a rectified stereo setting all epipolar lines are parallel. 0.5

□ The principal axes of two cameras are orthogonal after stereo rectification. -0.5

□ Stereo rectification can be achieved by transforming both images with a rigid-body transform (i.e. SE(3)). -0.5

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

p The following groups are generalizations of each other. Please order them from the most to the least general by entering numbers from 1 to 4.

(   ) General linear group \newline
(   ) Special orthogonal group \newline
(   ) Orthogonal group \newline
(   ) Set of square matrices

2.0 points · Free formatted question

Model answer

Solution:
2
4
3
1
because: SO(n) ⊂ O(n) ⊂ GL(n) ⊂ Set of square n x n matrices.
Grading:
2p: all 4 correct
1p: 2 correct
0.5: 1 correct
alternatively, extra points are given if the order is partially correct.

+1 point

Intrinsic camera calibration. Consider the following camera model:

$$w\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}f&0&u_0\\0&f&v_0\\0&0&1\end{bmatrix}\begin{bmatrix}r_{11}&r_{12}&r_{13}&t_x\\r_{21}&r_{22}&r_{23}&t_y\\r_{31}&r_{32}&r_{33}&t_z\end{bmatrix}\begin{bmatrix}x\\y\\y\\1\end{bmatrix}$$

How many parameters (DOF) need to be estimated for intrinsic camera calibration for the shown camera model?

0.5 points · Multiple choice · 5 alternatives

Model answer

3 DOF. The provided intrinsic (/calibration) matrix has 3 parameters (f, u_0, v_0). The model assumes square pixels (because f=f_x=f_y) and no screw parameter (because s=0)

● 3 0.5

○ 4 0.0

○ 6 0.0

○ 9 0.0

○ 12 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

The provided intrinsic (/calibration) matrix has 3 parameters (f, u_0, v_0).

r **Extrinsic camera calibration.** Using the same camera model as in the previous question:
How many parameters (DOF) need to be estimated for extrinsic camera calibration?

0.5 points · Multiple choice · 4 alternatives

Model answer

6 DOF

○	5	0.0
●	6	0.5
○	9	0.0
○	12	0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

Although the provided matrix has 9 parameters for rotation and 3 parameters for translation (12 in total), the rotation parameters are constrained to be rotation matrices, that is, members of the special orthogonal group SO(3). This means that the rows and column of rotation matrix need to be mutually orthogonal and the its determinant is +1. Such rotation matrices can be for example be generated from just 3 parameter, the angular rotation angles around each spatial axis. Hence, the degree of freedom (DOF) of the rotation
matrix is just 3. Together with the translation parameters, the extrinsic matrix has 6 DOF.

3 Question 3: Image Understanding

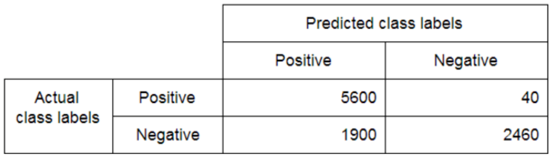
10.0 points · 8 questions

\vspace{0.5cm}

Traditional Classification and Retrieval

Text

a Consider a binary classifier with the following classification results.



		Predicted class labels	
		Positive	Negative
Actual class labels	Positive	5600	40
	Negative	1900	2460

Please compute the precision and recall values for this classifier using the provided confusion matrix.

1.0 point · Open · 1/4 Page

Model answer

[0.5p] Recall = 5600 / (5600 + 40) = 0.99

[0.5p] Precision = 5600 / (5600 + 1900) = 0.75

+1 point

Recall = 5600 / (5600 + 40) = 0.99

Precision = 5600 / (5600 + 1900) = 0.75

+0.5 points

Only one of the two is correct.

b Mark all correct statements.

1.0 point · Multiple choice · 5 alternatives

☑ **The maximum IoU score is 1.** 0.5

□ F1 score is not normalized and requires additional normalization to ensure a value range of [0,1]. -0.5

☑ **A bag of words representation does not account for the frequency a word occurs.** 0.5

□ F1=$2 \times \frac{Precision+Recall}{Precision \times Recall}$ -0.5

□ The bag of visual words representation describes an ordered set of visual words. -0.5

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

\vspace{4.5cm}

Object Detection

Text

c Mark all correct statements about object detection methods and their individual steps.

1.0 point · Multiple choice · 4 alternatives

□ For given object proposals, Fast R-CNN predicts absolute bounding box coordinates. -0.5

Feedback

This is incorrect, since the method predicts coordinate updates relative to the initial proposal.

☑ **All R-CNN methods require object proposals as input.** 0.5

Feedback

True. All these methods require them.

□ The major difference between Fast R-CNN and (slow) R-CNN are learned vs. non-learned region proposals. -0.5

Feedback

This is incorrect. Both methods use non-learned region proposals. In the lecture we also discussed Faster R-CNN which uses learned region proposals, but the question refers to other methods.

☑ **Both Fast R-CNN and Faster R-CNN have a global feature extraction stage which is jointly computed and then used for all region proposals.** 0.5

Feedback

Correct.

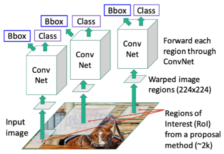
Feedback

Feedback when the question is answered correctly

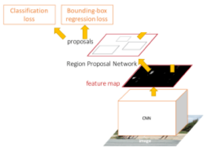
Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

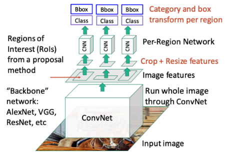
d Object detection architectures.


Bbox
Class
Bbox
Class
Bbox
Class
Conv Net
Conv Net
Conv Net
Forward each region through ConvNet
Warped image regions (224x224)
Regions of Interest (RoI) from a proposal method (~2k)
Input image

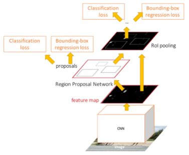
A



B


Bbox
Bbox
Bbox
Class
Class
Class
Category and box transform per region
Regions of Interest (RoIs) from a proposal method
CNN
CNN
CNN
Per-Region Network
Crop + Resize features
Image features
"Backbone" network: AlexNet, VGG, ResNet, etc
ConvNet
Run whole image through ConvNet
Input image

C



D

Match the correct name with the architectures **A, B, C, D** depicted above:

( ) R-CNN \newline

( ) Fast R-CNN \newline

( ) Faster R-CNN \newline

( ) Single Stage Detector

1.0 point · Free formatted question

Model answer

- **A** - R-CNN
- **C** - Fast R-CNN
- **D** - Faster R-CNN
- **B** - Single Stage Detector

1 pt if all 4 are correct; 0.5 pts if two are correct; 0pts otherwise

+1 point

All four methods are correctly assigned.

+0.5 points

At least 2 methods correctly associated.

\vspace{0.5cm}
Neural Networks

Text

e Select all correct statements about neural network architectures.

1.0 point · Multiple choice · 4 alternatives

☑ **Convolutional layers have fewer parameters than fully connected layers if the kernel size is smaller than the image size.** 0.5

□ Fully connected layers are naturally invariant to object translations in the image. -0.5

□ Convolutional layers are naturally invariant to object rotations in the image. -0.5

☑ **For patch sizes smaller than the image, convolutional layers require less computing operations and occupy less memory.** 0.5

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

f Consider a 2D convolutional layer with RGB-D input of size 64 x 64. We apply 3 convolutional filters of size 4 x 4. All input channels are convolved together, not separetely, no padding, and stride = 2.

1. What are the dimensions of the output activation layer ?
2. How many weight parameters have to be trained in this layer ?

2.0 points · Open · 1/4 Page

Model answer

Part 1)

The output size W_{out} of a convolutional layer for **each** dimension can be computed for a given input size W_{inp}, a filter/kernel size k, padding p and stride s as follows:

$$W_{out} = \text{floor}(\frac{W_{inp} - k + 2p}{s}) + 1$$

that is, for input size 64 and kernel size k=4, s=1 and p=0 we get and output size per dimension of

$$W_{out} = \text{floor}(\frac{64 - 4 + 0}{2}) + 1 = 31$$

thus for 4 conv. filters we get an output size of 31 x 31 x 3.

Part 2)

Number of parameters = kernel height x kernel width x depth input x depth output/nr filters.

Number of parameters = 4*4*3*3 = 144

+1 point

Question part 1. dimensions of output activation: **31 x 31 x 3** (=2883),
being: height (minus kernel size) x width (minus kernel size) x number of filters.

+0.5 points

Question part 1. minor error (like 30 x 30 x 3)

+1 point

Question part 2. Number of learnable parameters = dimensions of the kernel:
5 x 5 x 3 x 4,
being:
kernel height x kernel width x depth input x depth output/nr filters.
Number of parameters = 4*4*3*3 = 144

+0.5 points

Question part 2. Minor error.

g On the result of the convolutional layer of the previous question we apply a 3 x 3 max pooling layer using stride = 1.

1. What are the dimensions of the output layer ?
2. How many weight parameters have to be learned in this layer ?

2.0 points · Open · 1/4 Page

Model answer

The output size W_{out} of a pooling layer can be similarly computed as for convolution layer in the previous question.
That is, for a given input size W_{inp}, a filter/kernel size k, padding p and stride s as follows:

$$W_{out} = \text{floor}(\frac{W_{inp} - k + 2p}{s}) + 1$$

that is, for input size 31 and kernel size 3, s=1 and p=0 we get and output size per dimension of

$$W_{out} = \text{floor}(\frac{31 - 3 + 0}{1}) + 1 = 29$$

thus for 3 conv. filters we get an **output size of 29 x 29 x 3**.

+1 point

Question part 1.
1pt if mentioned input conditions and corresponding output are correct:
s=1,p=0 -> dimensions of output: **29 x 29 x 3** (=2523)

+0.5 points

Question part 1.
if the answer is partially correct, e.g. only one number of the output dimensions is correct or given, e.g. 29x29 without 3 channel dimensions or a similar minor error.

+1 point

Question part 2.
No learnable parameters exist in (vanilla) pooling.

+0.5 points

Question part 2.
Minor error.

h Mark all correct statements about 2-stage and single shot object detectors.

1.0 point · Multiple choice · 5 alternatives

☐	2-stage detectors require substantially less object proposals than single shot detectors.	-0.5
☐	Non-maximum suppression is only required for 2-stage detectors.	-0.5
☑	**Single shot detectors are typically faster than 2-stage detectors.**	0.5
☐	Single shot detectors require much less training data than 2-stage detectors.	-0.5
☑	**Single shot detectors use a small set of fixed regions as object proposals.**	0.5

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly