CV1 practice exam 1

52041COV6Y Computer Vision 1 24/25 (1.1) · 3 exercises · 43.0 points

1 Question 1: Low Level Vision

8.0 points · 13 questions

\vspace{0.5cm}

Camera Model

Text

Figure 1.1:

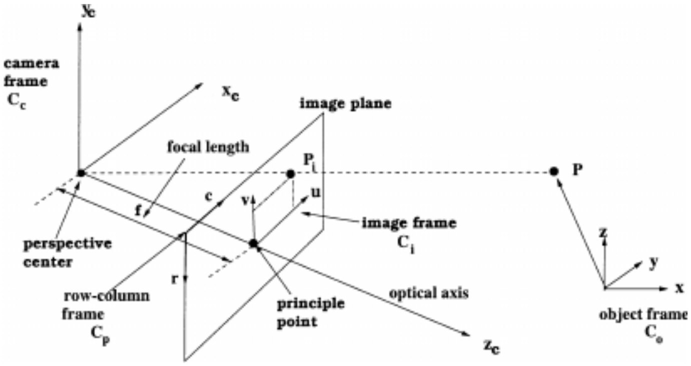


Figure 1.1 illustrates the projection from 3D space to 2D image using a pin-hole camera. The equation can be written as: \vspace{0.5cm}

$$\mathbf{x} = \mathbf{K}\begin{bmatrix}\mathbf{R} & \mathbf{t}\end{bmatrix}\mathbf{X}$$

$$\Downarrow$$

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

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 **[R t]**?

0.5 points · Multiple choice · 5 alternatives

○ Intrinsic matrix	0.0
● **Extrinsic matrix**	0.5
○ 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 What is the meaning of u_0?

0.5 points · Multiple choice · 5 alternatives

●	**translation on the image plane**	0.5
○	rotation of the lens	0.0
○	scaling on the image plane	0.0
○	skew on the image plane	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

c Following the questions above, we know **R** is an identity matrix, **t**=0, $\mathbf{X}_1$=[0, 1000, 1000, 1]$^\top$, $\mathbf{X}_2$=[1000, 1000, 1000, 1]$^\top$, $\alpha = \beta = 0.1$, $s = 0.01$, $u_0 = v_0 = 0$.

Compute the coordinates for $\mathbf{x_1} = w[u_1, v_1, 1]^\top$, $\mathbf{x_2} = w[u_2, v_2, 1]^\top$.

1.0 point · Open · 9/20 Page

+0.5 points

x1 = 1000 [0.01, 0.1, 1]$^\top$

+0.5 points

x2 = 1000 [0.11, 0.1, 1]$^\top$

d If we want to flip the image coordinates upside down by changing the intrinsic matrix, what will you change?

0.5 points · Open · 1/5 Page

+0.5 points

flip the sign of β

e Notice that the image plane is on the same side of the object. Usually this is not possible physically, but it makes the illustration more convenient and intuitive. What is this technique called?

0.5 points · Open · 1/10 Page

+0.5 points

virtual image plan or something with the same meaning.

\vspace{0.5cm}
Human Vision and Color

Text

f Here are three statements regarding human perception. Which are correct?
A. Both cones and rods on the retina are light preceptors.
B. For humans, there are three types of rods which perceive red, green and light respectively.
C. There is a blind-spot on the human retina.

0.5 points · Multiple choice · 7 alternatives

○ only A	0.0
○ only B	0.0
○ only C	0.0
● A and C	0.5
○ B and C	0.0
○ all of the three	0.0
○ none are correct	0.0

Feedback

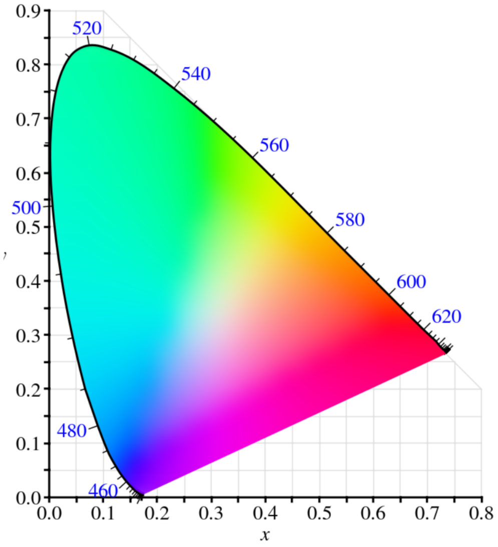
Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

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 $X_A = 200$, $Y_A = 100$ 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 = 0.5, Ay = 0.4

+0.5 points

Hue is approximately 587nm, 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

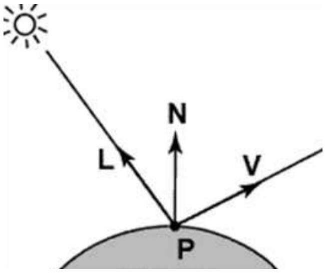
Saturation is approximately 67%

\vspace{0.4cm}

Reflection Model

Text

Figure 1.3:



As illustrated above in Figure 1.3, we observe point **P** on a surface. **N** is the surface normal at **P**. **L** is the direction vector to the light source and **V** is the direction vector to view point.

Text

h Assume the only light source L is duo wavelengths, and the wavelengths are λ_1 = 950nm and λ_2= 350nm. The intensity of two wavelength are equal. The material has a uniform positive wavelength response (albedo) from 500-2000nm.
A camera has a uniform response from 400-1500nm and cannot capture light outside the range. At current setting, the camera captures the light intensity at a point P on the material as I_1=100.

If we change the wavelengths of the light source, now λ_1 = 950nm and λ_2 = 550nm; and keep everything else unchanged. What will be the observed intensity I_2 at point P?

0.5 points · Open · 9/20 Page

+0.5 points

200
Previously only 950nm are reflected and captured, now both are reflected and captured

i Following the above question, we reverse the light source to the original setting, the wavelengths are λ_1 = 950nm and λ_2 = 350nm. This time, instead of using a camera, a human observes P directly. What will they see?

0.5 points · Open · 3/10 Page

+0.5 points

Nothing, both wavelength are out of visible spectrum j Follow the illustration in Figure 1.3. Now we study the direction instead of the wavelength. (Use only the conditions in the block of figure 1.3 and ignore the conditions in other questions.)

Assume the surface is Lambertian (ideal diffuse). Denote the albedo at **P** as ρ, Write down the Lambertian equation that determines the observed brightness I_p, upon the directions using **L, N, V** and ρ.

0.5 points · Open · 3/10 Page

+0.5 points

Lp $\propto \rho$ **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.

k Follow the question above. Further assume that when the angle between **N** and **V** is 15°, and we observe the brightness at **P** is I_P=100. If we change **V** so that angle between **N** and **V** is 45°, what is the new I_P?

0.5 points · Open · 2/5 Page

+0.5 points

Still 100, between for Lambertian surface, Lp is independent from V

l Go back to the original illustration in figure 1.3, We assume Lambertian surface again. This time we assume angle between **L** and **N** is 60°. Angle between **L** and **V** is 30°. We observe the brightness at **P** is I_P=100.

If we change the position of light source and viewpoint, so that angle between **L** and **N** is now 30°, and angle between **N** and **V** is now 45° , everything else remains the same. What is the new observed brightness I_p?

0.5 points · Open · 2/5 Page

+0.5 points

cos 30° / cos 60 ° = $\sqrt{3}$ = 1.732

Give only cos 30° / cos 60° is fine

m Go back to the original setting as illustrated in Fig. 1.3, this time the material is *ideal glossy*. The angle between **L** and **N** is 60°, and the angle between **N** and **V** is 30°. Can you determine the observed brightness at **P**? If yes, give I_p. if not, why not?

0.5 points · Open · 2/5 Page

+0.5 points

Lp = 0, it works as a mirror, the angles between N, L and N, V are not the same

2 Question 2: Image Processing

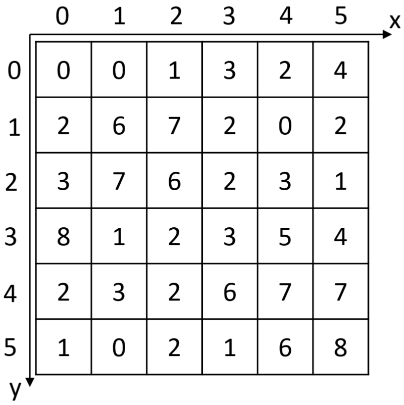
25.0 points · 23 questions

\vspace{0.5cm}

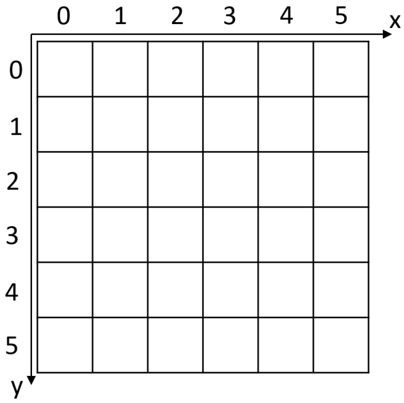
In Place Processing and Morphology

Text

a A binary image is commonly used in editing and analysis. Below is an integer image coded from 0 to 9. Convert it to binary image in the empty image below, using threshold $T_h \geq 5$. Use a pencil in case you need to change your answer



	0	1	2	3	4	5
0	0	0	1	3	2	4
1	2	6	7	2	0	2
2	3	7	6	2	3	1
3	8	1	2	3	5	4
4	2	3	2	6	7	7
5	1	0	2	1	6	8

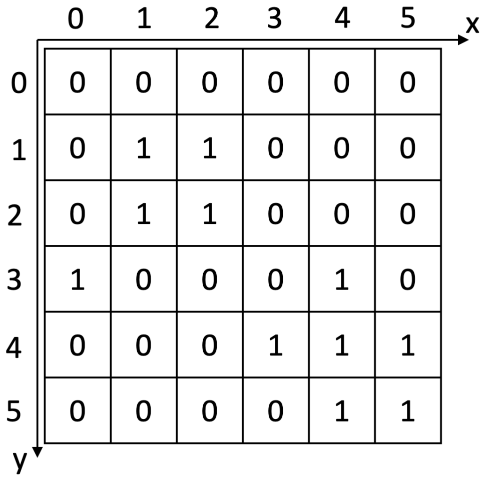


	0	1	2	3	4	5
0						
1						
2						
3						
4						
5						

1.0 point · Free formatted question

+1 point

no mistakes. Give only the foreground is fine



	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	1	0	0	0
2	0	1	1	0	0	0
3	1	0	0	0	1	0
4	0	0	0	1	1	1
5	0	0	0	0	1	1

+0.5 points

up to 2 mistakes

b Follow the above question and consider the foreground. How many connected components are there if using 4-connected neighborhood?

0.5 points · Multiple choice · 6 alternatives

○ 1	0.0
○ 2	0.0
● 3	0.5
○ 4	0.0
○ 5	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

c Follow the above question. For the largest component, compute the geometric center of the component. Use the coordinates given on the image.

1.0 point · Open · 1/2 Page

+0.5 points

mx = (3 * 1 + 4 * 3 + 5 *2) / 6 = 25/6 = 4.17

+0.5 points

my = 25/6 = 4.17

+1 point

d Below is an 5*7 binary image and a 3*1 construction element. The background pixels are '0' and are not visualized in the image.

	1					
1	1				1	
	1			1	1	1
	1				1	
1	1	1				

1
1
1

Perform an erosion on the image using the construction element. Use mirror padding.

1.0 point · Image · 1/100 Page

+1 point

	1					
	1					
	1				1	
	1					
	1					

Only the foreground is fine

+0.5 points

1 mistake

e Perform a dilation on the eroded image you got in 2d. Use mirror padding.

1.0 point · Free formatted question

+1 point

	1					
	1				1	
	1				1	
	1				1	
	1					

+0.5 points

one mistake

f Which of the following statements regarding morphology is *incorrect*?

0.5 points · Multiple choice · 4 alternatives

○ The output of morphology for boundary pixels depends on the padding method. 0.0

○ There exists duality of morphology, dilation of the foreground is equivalent to erosion of the background. 0.0

● **A combination of erosion and dilation is called closing if we do erosion first.** 0.5

○ Erosion can be used for edge detection. 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

\vspace{0.5cm}
Image Filtering

Text

g Which of the statements is correct regarding image filtering?

0.5 points · Multiple choice · 4 alternatives

○ Images are always smoothed after image filtering 0.0

● **A 2D box filter can be separated into two 1D uniform filters in x and y direction. This will speed up the filtering.** 0.5

○ Gaussian filters are popular because the addition of two Gaussian filters generates another Gaussian filter. 0.0

● **It is possible that the image remains unchanged despite the filtering** 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

h This is an equation of 2D Gaussian filter.

$$G_\sigma(x, y) \quad = \quad \frac{1}{2\pi\sigma^2} \exp^{-\frac{x^2 + y^2}{2\sigma^2}}$$

Write down the equation that separates it into two 1D Gaussian filters in x and y direction respectively.

1.0 point · Open · 2/5 Page

+1 point

$$\left(\frac{1}{\sqrt{2\pi\sigma}} \exp^{-\frac{x^2}{2\sigma^2}}\right) \left(\frac{1}{\sqrt{2\pi\sigma}} \exp^{-\frac{y^2}{2\sigma^2}}\right)$$

+0.5 points

one mistake

i For this image coded in float from 0 to 1. (only '1's are visualized, the rest are all '0's)

	1					
1	1				1	
	1			1	1	1
	1				1	
1	1	1				

We want to smooth patch B with a 3 x 3 box averaging filter. Use zero padding. (For your convenience, you can leave the '1/9' outside).

1.5 points · Free formatted question

+1.5 points

1/9

3	3	2	0	1	1	1
4	4	3	1	3	4	3
4	4	3	1	4	5	4
4	5	4	2	3	4	3
3	4	3	1	1	1	1

+1 point

one mistake

+0.5 points

two mistakes

j Following the above question, apply an 3 x 3 median filter on the original image. This time, use warp padding

\vspace{3cm}

1.5 points · Free formatted question

+1.5 points

1	1					
1						
1					1	1
1	1					
	1					

The rest are '0's.

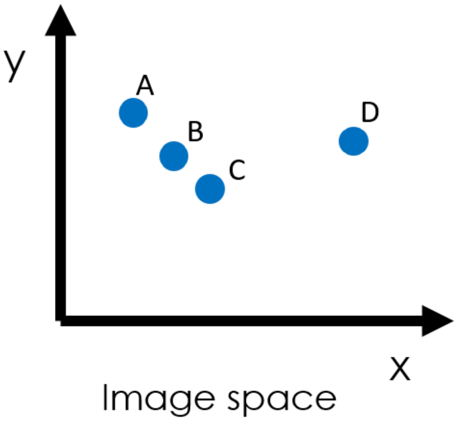
solution1, do a box filtering with wrap padding. Threshold with value >= 5/9
solution2, for any pixel with >= 5 '1's in its neighborhood, output 1, otherwise 0

+1 point
one mistake

+0.5 points
two mistakes

Edges and Lines
Text

k Figure 2.1:



In figure 2.1, we have 4 points in an image. The coordinates are A = (A_x, A_y) = (1, 4), B = (2, 3), C = (3, 2), D = (6, 4).
Show how you can fit a line using Hough transform. For simplicity, use the linear version: y=ax+b.

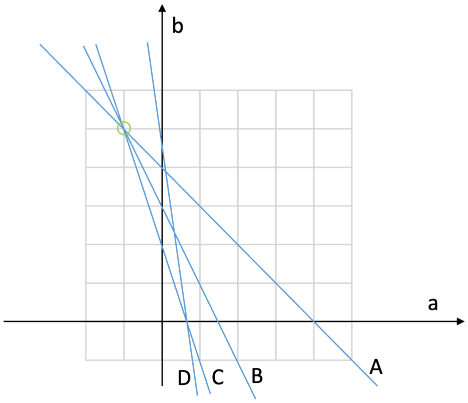
2.0 points · Open · 1/2 Page

+2 points

step 1 Give the four lines in Hough space (1pt)
A: a + b = 4
B: 2a + b = 3
C: 3a + b = 2
D: 6a + b = 4
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 detuction in step 1.

Step 3: find a= -1, b=5 (0.5pt)

+1.5 points

see grading scheme above

+1 point

see grading scheme above

+0.5 points

see grading scheme above

l In the question above, we used the linear version of Hough transform for simplicity. Name a disadvantage of the linear version.

0.5 points · Open · 3/20 Page

+0.5 points

It can not handle vertical lines or anything reasonable.

\vspace{0.5cm}
Corners

Text

m 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 all pixels.

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

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 1
0 0

fy =
1 1
-1 -1

M11 =
0 0
0 1

M12 =
1 1
1 1

M21 =
0 0
0 1

M22 =
0 0
0 1

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

n What decision criterion is used to identify corners in the Harris corner detector?

1.0 point · Multiple choice · 5 alternatives

○ If the eigenvectors of the structure tensor are orthogonal. 0.0

○ If the eigenvalues of the structure tensor are both small. 0.0

● **If the eigenvalues of the structure tensor are both large.** 1.0

○ If the determinant of the structure tensor is negative. 0.0

○ If the trace of the structure tensor is negative. 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

"If the eigenvalues of the structure tensor are both large" is the only correct answer.

\vspace{0.5cm}
Optical Flow

Text

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

1.5 points · Multiple choice · 6 alternatives

☑ **The method assumes color/brightness constancy for corresponding pixel colors.** 0.5

□ Due to the windowed estimation the method can only estimate motion in the direction of constant brightness. -0.5

☑ **The method assumes that the spatial gradient between neighboring flow vectors is zero.** 0.5

□ The method is robust and also works well for large motions (larger than the window size). -0.5

□ For highly textured regions the structure tensor might not have full rank and the flow vector can be undefined. -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{6cm}

Linear Transformations

Text

p Different types of transformations have a different degree of freedom. What is the degree of freedom of a 2D rigid body transform $T \in SE(2)$?

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	1.0
○	4	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

q What is the degree of freedom of a 3D rotation $R \in SO\left(3\right)$?

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.

●	3	1.0
○	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

r Transformation composition: Given two points $A = (a_x, a_y), B = (b_x, b_y) \in \Re^2$, we are looking for a transformation that describes the rotation of point B around point A by a given angle θ. For convenience, we describe the following matrices: $R_\theta \in SO(2)$ is the 2D rotation matrix that rotates the coordinate frame by angle θ, and $T_{[x,y]} = \begin{pmatrix} 1 & 0 & x \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix}$ is a translation matrix, both in homogeneous coordinates.
Which transformation rotates point B around point A?

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.

○ $R_\theta T_{[a_x,a_y]} B$ 0.0

○ $T_{[-a_x,-a_y]} R_\theta T_{[a_x,a_y]} B$ 0.0

● $T_{[a_x,a_y]} R_\theta T_{[-a_{x,} - a_y]} B$ 1.0

○ $T_{[-b_x,-b_y]} R_\theta T_{[a_x,a_y]} B$ 0.0

○ $T_{[-b_x,-b_y]} R_\theta T_{[b_x,b_y]} A$ 0.0

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

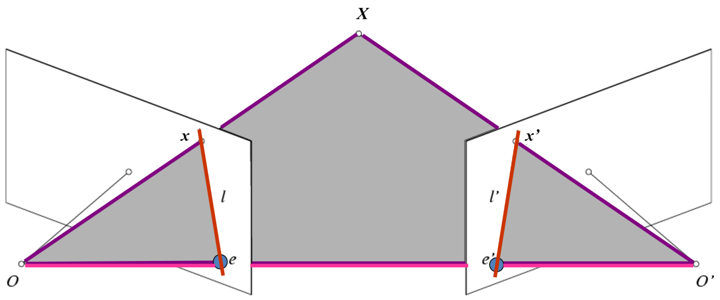
The main idea is to decompose the transformation into three parts:
1) translate everything such that point A is in the new coordinate origin, i.e. $T_{[-a_{x,} - a_y]}$.
2) rotate around the coordinate origin, i.e. apply R_θ
3) translate everything back, i.e. , i.e. $T_{[a_x,a_y]}$.
All transformations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

\vspace{8cm}

Multiview Geometry and Reconstruction

Text

s 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 lines lie in the epipolar plane** 1.0

□ The lines through $\vec{O}x$ and $\vec{O}'x'$ are called the epipolar lines. -1.0

□ The points x, x' are called the epipoles. -1.0

☑ **The epipoles lie on the epipolar lines.** 1.0

□ The epipolar plane is orthogonal to each of the image planes. -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

t Rectified Stereo:

1.0 point · Multiple choice · 4 alternatives

☐ In the rectified stereo case the image planes of the two images are orthogonal to each other. -0.5

☐ In a rectified side-by-side stereo setting all epipolar lines are vertical. -0.5

☑ **In rectified stereo, the dense correspondence problem can be reduced to a 1-dimensional problem only estimating a single disparity value per pixel.** 0.5

☑ **In the rectified stereo case the epipoles move to inifinity.** 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

u Shape-from-X: Name at least 3 particular shape-from-X problems.

1.5 points · Open · 7/20 Page

Model answer

0.5 points for every correctly mentioned problem setting, which is one of the following (max 1.5 points):

- Shape-from-Shading
- Shape-from-Texture
- Shape-from-Focus/Defocus
- Shape-from-Shadows
- Shape-from-Silhouettes
- Shape-from-Motion

+1.5 points

3 correct answers out of the model answers.

+1 point

2 correct answers out of the model answers.

+0.5 points

1 correct answer out of the model answers.

v Intrinsic camera calibration.

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

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

0.5 points · Multiple choice · 4 alternatives

Model answer

5 DOF. The provided intrinsic (/calibration) matrix has 5 parameters (f_x, f_y s, u_0, v_0).

● 5 0.5

○ 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 5 parameters (f_x, f_y s, u_0, v_0).

w 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 Precision and recall are commonly used to evaluate the performance of image classification systems. In the following, please select the correct terms for precision and recall for given numbers of 'false positives' (FP), 'false negatives' (FN), 'true positives' (TP) and 'true negatives' (TN).

1.0 point · Multiple choice · 4 alternatives

□	Recall = $\frac{TP}{TP+FP}$	-0.5
□	Precision = $\frac{TP}{TN+FP}$	-0.5
☑	Recall = $\frac{TP}{TP+FN}$	0.5
☑	Precision = $\frac{TP}{TP+FP}$	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

b Mark all statements which are correct.

1.0 point · Multiple choice · 5 alternatives

☑ The IoU score does not need to be normalized since it is always within the unit interval. 0.5

□ F1 =$\frac{1}{2} \times (Precision + Recall)$ -0.5

□ A bag of visual words representation cannot be used for image retrieval. -0.5

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

□ The bag of visual words representation is not invariant to permutations 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{0.5cm}

Object Detection

Text

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

1.0 point · Multiple choice · 4 alternatives

☑ **The sliding window approach and selective search can be used for generating object proposals.** 0.5

Feedback

Both methods have been repeatedly discussed in the lectures.

□ R-CNN predicts absolute bounding box coordinates independent from object proposals. -0.5

Feedback

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

□ 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.

☑ **Faster R-CNN contains a joint feature extraction stage for the entire image before processing individual regions of interests with locally.** 0.5

Feedback

Both Fast R-CNN and Faster R-CNN have global feature extraction stage which is jointly computed and then used for all region proposals.
Unfortunately, there is a typo in the question. The word "with" needs to be removed.

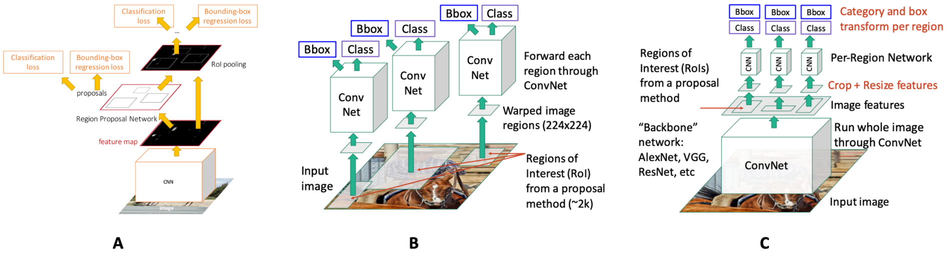
Feedback

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.


Classification loss
Bounding-box regression loss
Classification loss
Bounding-box regression loss
RoI pooling
proposals
Region Proposal Network
feature map
CNN
image
A
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
B
Bbox
Bbox
Bbox
Class
Class
Class
Category and box transform per region
CNN
CNN
CNN
Per-Region Network
Regions of Interest (RoIs) from a proposal method
Crop + Resize features
Image features
"Backbone" network: AlexNet, VGG, ResNet, etc
Run whole image through ConvNet
ConvNet
Input image
C

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

( ) R-CNN \newline

( ) Fast R-CNN \newline

( ) Faster R-CNN

1.0 point · Free formatted question

Model answer

- **B** - R-CNN
- **C** - Fast R-CNN
- **A** - Faster R-CNN

1 pt if all 3 are correct; 0.5 pts if one is correct; 0pts otherwise

+0.5 points

1 method correct.

+1 point

All three methods are correctly assigned.

\vspace{0.5cm}

Neural Networks

Text

e List the major advantages (>=2) of convolutional layers compared to fully connected layers.

1.0 point · Open · 2/5 Page

+1 point

Convolutional layers have multiple advantages over fully connected layers:
- they build in translational invariance by design
- user fewer parameters (if patch size is smaller than images size)
- use fewer memory / less computing operations (if patch size is smaller than images size)
- the weights in the kernel are shared when applied to different locations in the input image

+0.5 points

only 1 advantage mentioned or minor error.

f In a 2D convolutional layer, we have RGB input of size 64 x 64. We apply 4 convolutional filters of size 5 x 5. No padding, and stride = 1.
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 5, s=1 and p=0 we get and output size per dimension of

$$W_{out} = \text{floor}(\frac{64 - 5 + 0}{1}) + 1 = 60$$

thus for 4 conv. filters we get an output size of 60 x 60 x 4.

Part 2)

Number of parameters = kernel height x kernel width x depth input x depth output/nr filters.
Number of parameters = 5*5*3*4 = 300

+1 point

Question part 1. dimensions of output activation: **60 x 60 x 4** (=14400),
being: height (minus kernel size) x width (minus kernel size) x number of filters.

+0.5 points

Question part 1. minor error (like 61 x 61 x 4)

+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 = 5*5*3*4 = 300

+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 = 2.
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 60 and kernel size 3, s=2 and p=0 we get and output size per dimension of

$$W_{out} = \text{floor}(\frac{60-3+0}{2}) + 1 = 29$$

thus for 4 conv. filters we get an **output size of 29 x 29 x 4**.

Alternative answers:
Due to several questions during the exam regarding padding or applying the more meaningful stride of 3 the following alternative answers are also considered correct if the corresponding input conditions/assumptions regarding padding and stride are clearly mentioned:

A1) For input size 60 and kernel size 3, s=2 and p=1 we get and output size per dimension of $W_{out} = \text{floor}(\frac{60-3+2}{2}) + 1 = 30$ and resulting **output size of 30 x 30 x 4**.

A2) For input size 60 and kernel size 3, s=3 and p=0 we get and output size per dimension of $W_{out} = \text{floor}(\frac{60-3+0}{3}) + 1 = 20$ and resulting **output size of 20 x 20 x 4**.

+1 point

Question part 1.
1pt if mentioned input conditions and corresponding output are correct:
s=2,p=0 -> dimensions of output: **29 x 29 x 4** (=3364), or
s=2,p=1 -> dimensions of output: **30 x 30 x 4** (=3600), or
s=3,p=0 -> dimensions of output: **20 x 20 x 4** (=1600)

+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. 30x30 without 4 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

☑ **Both types of object detectors require non-maximum suppression to filter duplicate detections.** 0.5

☐ Single shot detectors are typically more accurate than 2-stage detectors. -0.5

☑ **2-stage detectors require substantially more object proposals than single shot detectors.** 0.5

☐ 2-stage detectors are typically faster than single shot detectors. -0.5

☐ Due to the omitted object proposal stage, single shot detectors require much less training data than 2-stage detectors. -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