Exercises

1	2	3

Surname, First name

1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9
0	0	0	0	0	0	0	0

Computer Vision 1 (52041COV6Y)
CV1 practice exam 1

UNIVERSITEIT VAN AMSTERDAM

Before you start
1. Do not open this booklet until the supervisor signals the start.
2. Fill out your student number and name on this page. Notify the instructors immediately if you made a mistake.
3. The exam needs to be finished within 3 hours unless an extension is given by the exam committee.
4. You are allowed to use a calculator and one A4 memo sheet. Cell phone and laptop are strictly prohibited unless you have special permission from the exam committee.
5. This exam has in total 45 points.

During your exam
1. Keep a good manner and try not to disturb other students.
2. If you find any typo or ambiguity in the questions, raise your hand and let the supervisors know.
3. If you want to use the toilet, raise your hand.
4. Use the scrap paper for drafting.
5. Use a pencil for multiple choice and figure-related questions!

After the exam
1. Before you submit, double check that your name/student ID are filled out correctly and you did not miss any questions.
2. Please fill out the course evaluation.
3. Submit everything at the front desk and sign the attendance sheet.
4. Leave quietly.

Good Luck!

This page is left blank intentionally

Question 1: Low Level Vision

Camera Model

Figure 1.1:

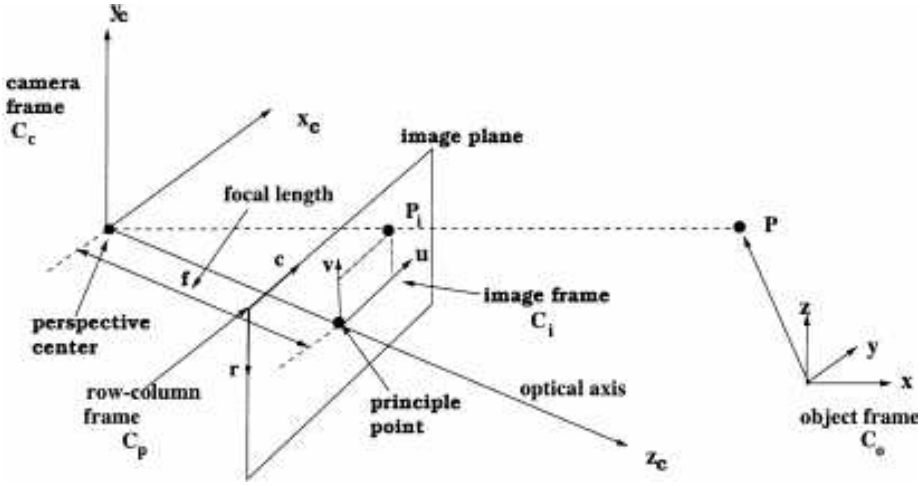


Figure 1.1 illustrates the projection from 3D space to 2D image using a pin-hole camera. The equation can be written as:

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

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

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

0.5p **1a** What is the name of matrix **[R t]**?

- (a) Intrinsic matrix
- (b) Extrinsic matrix
- (c) Rotation matrix
- (d) Translation matrix
- (e) Projection matrix

0.5p **1b** What is the meaning of u_0?

(a) translation on the image plane

(b) rotation of the lens

(c) scaling on the image plane

(d) skew on the image plane

(e) none of them

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

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

0.5p **1d** If we want to flip the image coordinates upside down by changing the intrinsic matrix, what will you change?

0.5p **1e** 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?

Human Vision and Color

0.5p **1f** 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.

(a) only A

(b) only B

(c) only C

(d) A and C

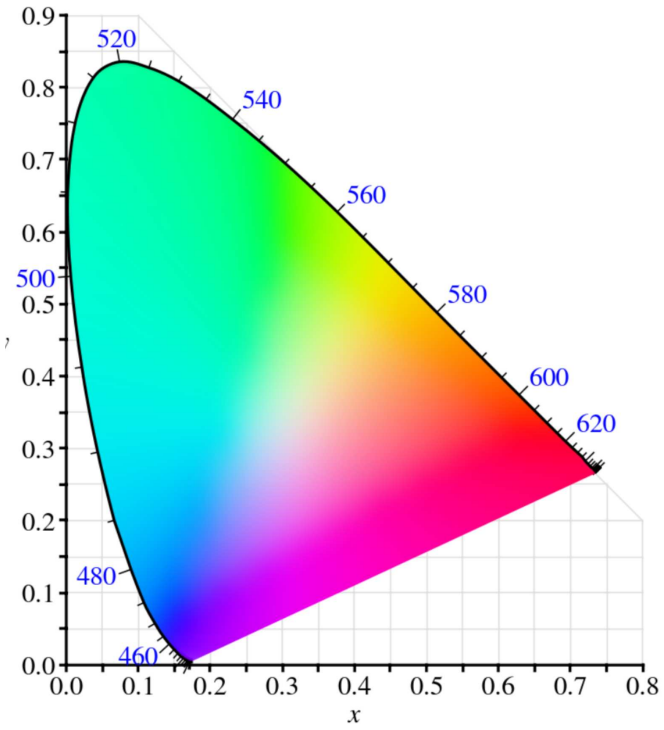
(e) B and C

(f) all of the three

(g) none are correct

1.5p **1g** 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)

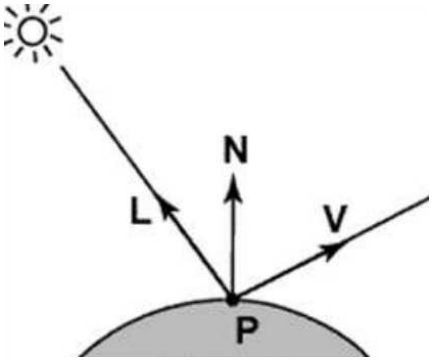
Figure 1.2:



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

Reflection Model

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.

0.5p **1h** 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?

0.5p **1i** 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?

0.5p **1j** 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.

0.5p **1k** Follow the question above. Further assume that when the angle between **N** and **V** is 15^o, and we observe the brightness at **P** is I_P=100. If we change **V** so that angle between **N** and **V** is 45^o, what is the new I_P?

0.5p **1l** 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^o. Angle between **L** and **V** is 30^o. 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^o, and angle between **N** and **V** is now 45^o , everything else remains the same. What is the new observed brightness I_p?

0.5p **1m** 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^o, and the angle between **N** and **V** is 30^o. Can you determine the observed brightness at **P**? If yes, give l_p. if not, why not?

Question 2: Image Processing

In Place Processing and Morphology

1p **2a** 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

y \ x	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

y \ x	0	1	2	3	4	5
0						
1						
2						
3						
4						
5						

0.5p **2b** Follow the above question and consider the foreground. How many connected components are there if using 4-connected neighborhood?

(a) 1 (b) 2 (c) 3 (d) 4 (e) 5 (f) 6

1p **2c** Follow the above question. For the largest component, compute the geometric center of the component. Use the coordinates given on the image.

1p **2d** 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.

1p **2e** Perform a dilation on the eroded image you got in 2d. Use mirror padding.

0.5p **2f** Which of the following statements regarding morphology is *incorrect*?

(a) The output of morphology for boundary pixels depends on the padding method.

(b) There exists duality of morphology, dilation of the foreground is equivalent to erosion of the background.

(c) A combination of erosion and dilation is called closing if we do erosion first.

(d) Erosion can be used for edge detection.

Image Filtering

0.5p **2g** Which of the statements is correct regarding image filtering?

(a) Images are always smoothed after image filtering

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

(c) Gaussian filters are popular because the addition of two Gaussian filters generates another Gaussian filter.

(d) It is possible that the image remains unchanged despite the filtering

1p **2h** This is an equation of 2D Gaussian filter.

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

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

1.5p **2i** For this image coded in float from 0 to 1. (only '1's are visualized, the rest are all '0's)

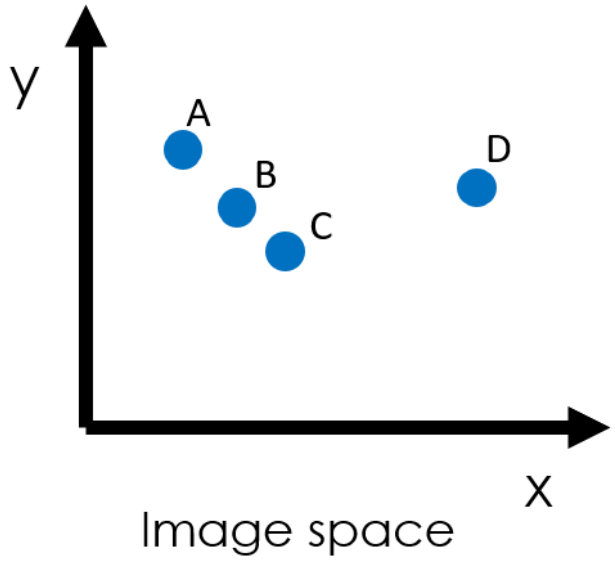
	1					
1	1				1	
	1			1	1	1
	1				1	
1	1	1				

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).

1.5p **2j** Following the above question, apply an 3 x 3 median filter on the original image. This time, use warp padding

Edges and Lines

2p **2k** 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.

0.5p **2l** In the question above, we used the linear version of Hough transform for simplicity. Name a disadvantage of the linear version.

Corners

2p **2m** 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.

1p **2n** What decision criterion is used to identify corners in the Harris corner detector?

(a) If the eigenvectors of the structure tensor are orthogonal.

(b) If the eigenvalues of the structure tensor are both small.

(c) If the eigenvalues of the structure tensor are both large.

(d) If the determinant of the structure tensor is negative.

(e) If the trace of the structure tensor is negative.

Optical Flow

1.5p **2o** Which statements about the Lucas-Kanade optical flow method are correct?

☐ The method assumes color/brightness constancy for corresponding pixel colors.

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

☐ The method assumes that the spatial gradient between neighboring flow vectors is zero.

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

☐ For highly textured regions the structure tensor might not have full rank and the flow vector can be undefined.

☐ The method fails to estimate a flow vector for homogeneously colored image regions.

Linear Transformations

1p **2p** Different types of transformations have a different degree of freedom. What is the degree of freedom of a 2D rigid body transform $T \in SE(2)$?

(a) 2 (b) 3 (c) 4 (d) 6

1p **2q** What is the degree of freedom of a 3D rotation $R \in SO(3)$?

(a) 3 (b) 6 (c) 9 (d) 12

1p **2r** 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?

(a) $R_\theta T_{[a_x,a_y]} B$

(b) $T_{[-a_x,-a_y]} R_\theta T_{[a_x,a_y]} B$

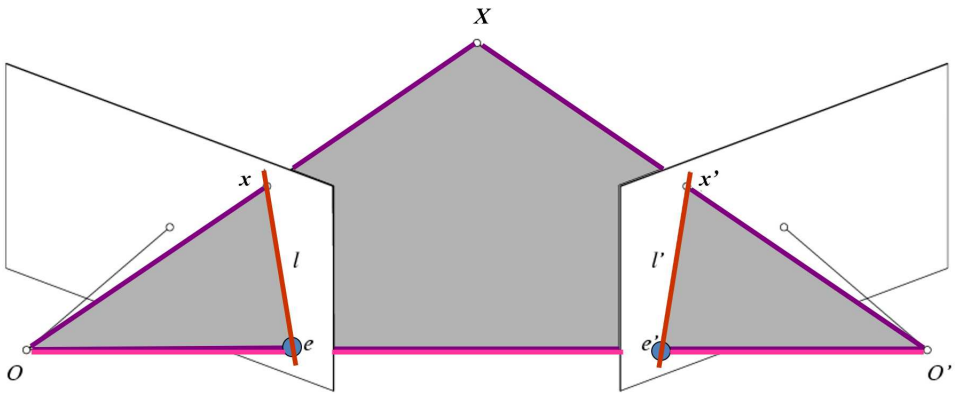
(c) $T_{[a_x,a_y]} R_\theta T_{[-a_x,-a_y]} B$

(d) $T_{[-b_x,-b_y]} R_\theta T_{[a_x,a_y]} B$

(e) $T_{[-b_x,-b_y]} R_\theta T_{[b_x,b_y]} A$

Multiview Geometry and Reconstruction

2p **2s** 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).

- ☐ The epipolar lines lie in the epipolar plane
- ☐ The lines through $\vec{O}x$ and $\vec{O}'x'$ are called the epipolar lines.
- ☐ The points x, x' are called the epipoles.
- ☐ The epipoles lie on the epipolar lines.
- ☐ The epipolar plane is orthogonal to each of the image planes.

1p **2t** Rectified Stereo:

- ☐ In the rectified stereo case the image planes of the two images are orthogonal to each other.
- ☐ In a rectified side-by-side stereo setting all epipolar lines are vertical.
- ☐ In rectified stereo, the dense correspondence problem can be reduced to a 1-dimensional problem only estimating a single disparity value per pixel.
- ☐ In the rectified stereo case the epipoles move to inifinity.

1.5p **2u** Shape-from-X: Name at least 3 particular shape-from-X problems.

0.5p **2v** 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?

(a) 5 (b) 6 (c) 9 (d) 12

0.5p **2w** 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?

(a) 5 (b) 6 (c) 9 (d) 12

Question 3: Image Understanding

Traditional Classification and Retrieval

1p **3a** 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).

- ☐ Recall = $\frac{TP}{TP+FP}$
- ☐ Precision = $\frac{TP}{TN+FP}$
- ☐ Recall = $\frac{TP}{TP+FN}$
- ☐ Precision = $\frac{TP}{TP+FP}$

1p **3b** Mark all statements which are correct.

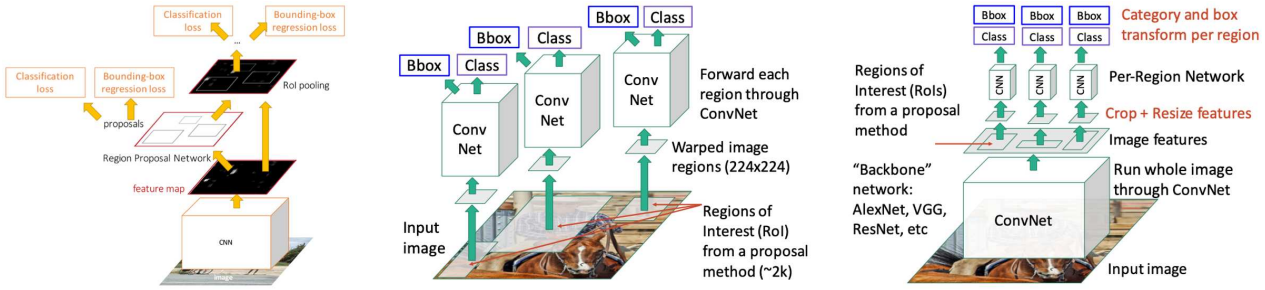
- ☐ The IoU score does not need to be normalized since it is always within the unit interval.
- ☐ F1 = $\frac{1}{2} \times (Precision + Recall)$
- ☐ A bag of visual words representation cannot be used for image retrieval.
- ☐ F1=$2 \times \frac{Precision \times Recall}{Precision+Recall}$
- ☐ The bag of visual words representation is not invariant to permutations of visual words.

Object Detection

1p **3c** Mark all correct statements about object detection methods and their individual steps.

- ☐ The sliding window approach and selective search can be used for generating object proposals.
- ☐ R-CNN predicts absolute bounding box coordinates independent from object proposals.
- ☐ The major difference between Fast R-CNN and (slow) R-CNN are learned vs. non-learned region proposals.
- ☐ Faster R-CNN contains a joint feature extraction stage for the entire image before processing individual regions of interests with locally.

1p **3d** Object detection architectures.



A B C

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

( ) R-CNN

( ) Fast R-CNN

( ) Faster R-CNN

Neural Networks

1p **3e** List the major advantages (>=2) of convolutional layers compared to fully connected layers.

2p **3f** 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 ?

2p **3g** 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 ?

1p **3h** Mark all correct statements about 2-stage and single shot object detectors.

- ☐ Both types of object detectors require non-maximum suppression to filter duplicate detections.
- ☐ Single shot detectors are typically more accurate than 2-stage detectors.
- ☐ 2-stage detectors require substantially more object proposals than single shot detectors.
- ☐ 2-stage detectors are typically faster than single shot detectors.
- ☐ Due to the omitted object proposal stage, single shot detectors require much less training data than 2-stage detectors.

This page is left blank intentionally