EPFL

1

Teacher : Prof. Pascal Fua
CS-442 Computer Vision - MA
20/06/2023
Duration : 90 minutes

Student 1

SCIPER : **999000**

Do not turn the page before the start of the exam. This document is double-sided, has 12 pages, the last ones possibly blank. Do not unstaple.

- Place your student card on your table.
- **A one page two-sided hand-written cheat-sheet** is allowed to be used during the exam.
- Using a **calculator** or any electronic device is not permitted during the exam.
- All questions have one or more correct answers.
- The grading scheme is such that random answering is discouraged:
  - Each answer of a multiple choice question is awarded +1 point if correct and −1 point if incorrect. If the **whole** question is left unanswered no points (positive nor negative) are awarded. Note that "correct" means that a true answer should be ticked and that a false one should be left unticked.

	Correct answers:	Student's answers:	Grading:
a)	■	■	+1
b)	☐	■	-1
c)	■	☐	-1
d)	☐	☐	+1

  - The scores for separate questions are **not clipped to 0**, that is, you can get negative score for a question.
- Use a **black or dark blue ballpen** and clearly erase with **correction fluid** if necessary.
- If a question is wrong, the teacher may decide to nullify it.

Respectez les consignes suivantes | Observe this guidelines | Beachten Sie bitte die unten stehenden Richtlinien

choisir une réponse | select an answer | Antwort auswählen

ne PAS choisir une réponse | NOT select an answer | NICHT Antwort auswählen

Corriger une réponse | Correct an answer | Antwort korrigieren

ce qu'il ne faut **PAS** faire | what should **NOT** be done | was man **NICHT** tun sollte

Pour votre examen, imprimez de préférence les documents compilés à l'aide de auto-multiple-choice.

First part: Multiple choice questions

For each question, mark the box corresponding to the correct answer. Each question has **at least one** correct answer.

Question 1 Which of the following provide(s) an exhaustive but non-redundant description of an edge?

- ☐ Direction, center, strength
- ☐ Normal, direction, center, strength
- ☐ Normal, direction, strength
- ☐ Normal, center, strength

Question 2 An image property $f(I)$ which can be computed pixel-wise is called *equivariant* to a coordinate transformation T if $f(T(I)) = T(f(I))$. As for example, if T is a rotation, we get the same result by

(a) Rotating the image and computing the property f.

(b) Computing the property f and rotating the grid of the outputs.

Now, let T be a 90 degree rotation about the center of an image. Which of the following edge properties are equivariant to T?

- ☐ Edge center
- ☐ Edge strength
- ☐ Edge direction
- ☐ Edge normal

Question 3 Which of the following 2D filters are separable?

- ☐ Derivative of an isotropic Gaussian
- ☐ A 2D Gaussian with $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- ☐ The Sobel filter
- ☐ A 2D Gaussian with $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

Question 4 Consider having already computed pixel-wise edge strength s_{ij} for a square image with side $h = w$. We are now applying hysteresis thresholding with thresholds $T_{low} < T_{high}$. Which of the following are true? Make sure your answers are correct regardless of h, image content, edge filter parameters and T_{low}, T_{high}.

- ☐ The edge strength value for certain pixels can be accessed an arbitrary number of times throughout the algorithm's iteration.
- ☐ The number of the pixel labeling iterations is linear in h, possibly with a large constant.
- ☐ Each pixel's edge strength will be accessed at least once.
- ☐ The number of pixels with $s < T_{low}$ has no impact on the number of the pixel labeling iterations.

Question 5 A digital camera captures images as a grids of pixel. Which of the following statement(s) is (are) correct?

- ☐ Humans vision has a sense of depth because one eye perceives depth, as opposed to digital cameras that only record intensities.
- ☐ A digital camera sensor converts an optical signal (photons) into an electrical signal (electrons).
- ☐ In a gray scale image, 8 bits can store 2*8=16 different gray levels.
- ☐ A color image is usually stored as a stack of 3 gray scale images.

Question 6 Projecting 3D coordinates to 2D pixels is done using a matrix multiplication in projective space (i.e. homogeneous coordinates). Which of the following statement(s) is (are) correct?

- ☐ In 2D homogeneous coordinates, $\mathbf{x} = (a, b, c)$ represents the 2D point (a, b) in cartesian coordinates.
- ☐ The matrix of intrinsic parameters K of a camera stores its rotation with respect to a canonical orientation.
- ☐ In 2D homogeneous coordinates, points are invariant to scaling. For example, $\mathbf{x} = (a, b, c)$ and $3\mathbf{x} = (3a, 3b, 3c)$ represent the same point.
- ☐ The matrix of intrinsic parameters K of a camera converts image coordinates into pixels.

Question 7 The pinhole camera model approximates real physical cameras. Which of the following statement(s) is (are) correct?

- ☐ Estimating the internal parameters of a camera is called the *"camera mensuration"* process.
- ☐ Given one camera used in multiple scenes, one has to estimate the camera's intrinsic parameters for each scene.
- ☐ The pinhole camera model accounts for the perspective (i.e. the further the objects, the smaller they appear).
- ☐ The pinhole camera model accounts for the depth of field (i.e. some objects can be out of focus).

Question 8 Which of the following statement(s) is (are) true about deep learning in the context of computer vision?

- ☐ A neural network with $n > 1$ layers is an affine or linear function.
- ☐ Adding layers to a neural network usually increases its descriptive power.
- ☐ To segment an image, a convolutional neural network is more adapted than a fully connected one (also called a multi layer perceptron).
- ☐ When using a convolution layer, the same convolution kernel is applied repeatedly on the entire image.

Question 9 Which segmentation algorithm(s) can process images without manually or automatically labeled foreground and background pixels?

- ☐ Trained U-net models
- ☐ Histogram Splitting
- ☐ ST Min-Cut
- ☐ SLIC Superpixels

Question 10 Which statement(s) about the Histogram Splitting segmentation algorithm is (are) *correct*?

- ☐ Global Histogram Splitting is sensitive to the initialization.
- ☐ Global Histogram Splitting works well when the images have complex illumination conditions.
- ☐ Histogram Splitting works better when the images have larger contrasts.
- ☐ Histogram Splitting considers the connectivity of pixel's neighborhoods.

Question 11 You can initialize the K-Means segmentation algorithm to obtain meaningful results with...

- ☐ seeds on a regular grid.
- ☐ seeds specified manually by mouse clicks.
- ☐ four seeds at the image center pixel.
- ☐ random seeds but choose the best result after trying multiple times.

Question 12 Which image segmentation algorithm(s) require(s) minimizing a loss function before obtaining segmentation masks?

- ☐ ST Min-Cut
- ☐ K-Means
- ☐ U-net
- ☐ Local Histogram Splitting using the HSV color space

Question 13 You have N pictures of the same object, with known camera positions. On each picture, you have computed the object silhouette. Consider the cones formed by casting rays from the camera origin to the complete silhouette. Denote the intersection of the cones $I \subset \mathbb{R}^3$, and their union $U \subset \mathbb{R}^3$. Finally, consider a point $x \in \mathbb{R}^3$.

- ☐ "x is inside the object" $\implies x \in I$
- ☐ $x \in I \implies$ "x is inside the object"
- ☐ $x \notin I \implies$ "x is outside the object"
- ☐ $x \notin U \implies$ "x is outside the object"

Question 14 You reconstruct a visual hull from multiple silhouettes of an object. However, you soon realize that reprojecting the visual hull onto the cameras does not match the initial silhouettes that you provided. How can this be explained?

- ☐ There are not enough cameras to recover the object's shape.
- ☐ The cameras are not well calibrated.
- ☐ You should have computed a convex hull instead.
- ☐ Some of the object concavities are lost.

Question 15 You are reconstructing a shape using a gradient descent method on the vertex locations of a mesh. The output is too irregular and spiky, and you would like to add a smoothing regularizer. Let us assume the shape is represented as a mesh (V, F) representing vertices and faces, respectively. Consider a vertex v_j, and its m neighboring vertices $(v_i)_{i\in[1,m]}$. Which of the following expression(s) define(s) a reasonable per-vertex smoothness loss $L_{reg}(v_j)$?

- ☐ $L_{reg}(v_j) = (v_j - \sum_{i=1}^{m} \frac{v_i}{m})^2$
- ☐ $L_{reg}(v_j) = (v_j - \sum_{i=1}^{m} v_i)^2$
- ☐ $L_{reg}(v_j) = (v_j + \sum_{i=1}^{m} \frac{v_i}{m})^2$
- ☐ $L_{reg}(v_j) = (v_j + \sum_{i=1}^{m} v_i)^2$

Question 16 Consider an object, as well as its bounding box, its convex hull, and its visual hull as seen from N cameras.

- ☐ The visual hull is included in the real object.
- ☐ Some object concavities can be lost in the convex hull.
- ☐ The visual hull is included in the convex hull.
- ☐ Some object concavities can be lost in the visual hull.

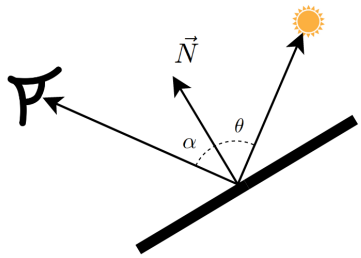
Question 17 We would like to reconstruct the shape of an object using *Shape from Shading*. Which of the following make(s) this task easier?

- ☐ The position and intensity of the light source are known.
- ☐ The object has constant albedo.
- ☐ We have images taken under different lighting conditions.
- ☐ The object has a repeating texture.

Question 18 At an art exhibition, you fall in love with a statue you want to digitize. It has a rather simple shape, and its surface is covered in perfect mirrors. You would like to reconstruct its shape, using shape from shading in a single image. Which of the following statements is (are) true in this case?

- ☐ This object is more diffuse than specular.
- ☐ The mirror surface makes this task more challenging.
- ☐ This object is more specular than diffuse.
- ☐ The object is approximately Lambertian.

Question 19



A camera is looking at a flat surface, which we assume to be perfectly lambertian. The surface's normal vector is $\vec{N}$, and it is lit by an infinitely distant light. Let us denote θ the angle between $\vec{N}$ and the vector that goes from the surface's center to the light source, and α the angle between $\vec{N}$ and the vector from the surface to the camera.

- ☐ If θ is high ($80° < \theta < 90°$), the surface will look very bright.
- ☐ Appearance does not depend on α.
- ☐ If I know $\vec{N}$ and θ, I can estimate α.
- ☐ α and θ can have the same value.

Question 20 In a shop, you see a white dress with a pattern of small black dots, regularly spaced, forming a grid. You'd like to reconstruct the shape of the dress. Which of the following methods will perform best?

- ☐ Shape from texture.
- ☐ None of the other will produce reasonable results.
- ☐ Shape from specularities.
- ☐ Shape from shading.

Question 21 Which statement(s) is(are) *true* about epipolar geometry? We assume that there are overlapped regions in the two images.

- ☐ The epipolar constraint applies when the two cameras are aimed in the same direction but at different places.
- ☐ Given two RGB images, epipolar geometry provides point-to-point correspondences.
- ☐ The depth information is required to reconstruct 3D points in triangulation.
- ☐ The epipolar constraint applies when the two cameras are at the same place but aimed in different directions.

Question 22 Which statement(s) is(are) *true* about camera, world, and pixel coordinates?

- ☐ The origin of the world coordinates is always at the center of the camera.
- ☐ Pixel coordinates are used to specify positions on the image plane, and their origin is always at the center of the image.
- ☐ Changes in lighting conditions can affect the transformation from camera coordinates to pixel coordinates.
- ☐ The relative transformation between camera coordinates and world coordinates is represented by the extrinsic parameters.

Question 23 Which of the following create(s) a challenge(s) for a stereo vision system?

- ☐ Occlusion
- ☐ Depth estimation given a disparity map
- ☐ Ambiguity in matching corresponding points
- ☐ Camera decalibration

Question 24 Which of the following is(are) the goal(s) of rectification in stereo vision?

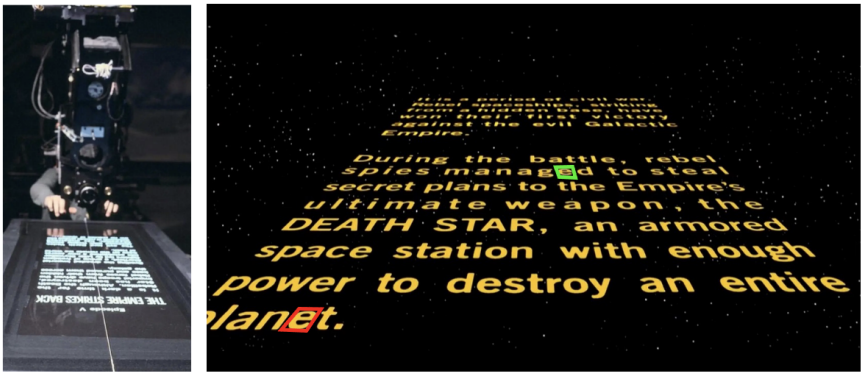
- ☐ To correct the lens distortion which could interfere with accurate stereo correspondence.
- ☐ To align the images so that corresponding points lie on the same horizontal line.
- ☐ To deal with the occlusion problem.
- ☐ To adjust the color balance in the stereo images.

Second part, open questions

Answer in the empty space below each question. Your answer should be carefully justified, and all the steps of your argument should be discussed in details.
Leave the check-boxes empty, they are used for grading.

"War drums echo through heavens as a roll up slowly crawls into infinity". This is how George Lucas described Star Wars opening title in his original screenplay.

Before the days of computer generated images (CGI), everything had to be done with practical effects. Physical models were filmed laid out on the floor (see below picture). The crawl effect was accomplished by the camera moving longitudinally along the model. It was a meticulous and time-consuming process to achieve a flawless smooth scrolling effect.



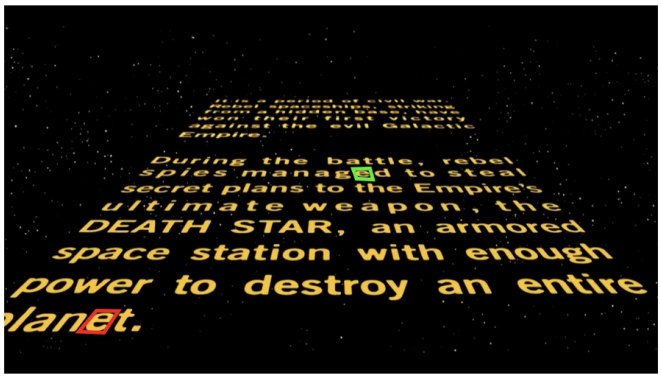
We will try to reconstruct the process of this filming and estimate the speed v of the text relative to the camera. To do that, we will apply techniques we learned in the course.

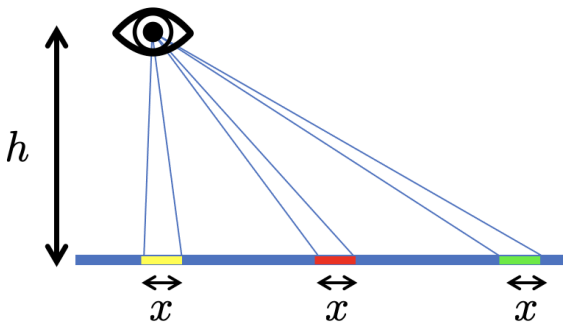
Question 25: *This question is worth 9 points.*

☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9

To estimate the speed of the text, we need to detect common patterns in successive images–a letter in our case. What methods do you remember from the course that are useful for automatically detecting generic patterns or shapes? Name a few.

Let us get back to our task.





We will follow the letters "e" in the video frames (*left*). On one frame, we marked two of them with red and green. We draw the scheme in a convenient way on the *right*. We estimated that the height of the camera over the model was $h = 10$ cm. The vertical sizes of letters "e" in pixels are $a = 25$ px (red) and $b = 5$ px (green). Additionally, we know that the letter "e" right below the camera would occupy $c = 50$ px (yellow). All three letters (yellow, red, green) have the same real size x. The focal length of the camera f and the size of the letter x are unknown.

Question 26: *This question is worth 6 points.*

☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6

We want to use the perspective camera (i.e., perspective projection) model to compute some relations between h, x, f, a, b and c. What assumptions does this model make?

Question 27: *This question is worth 9 points.*

☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9

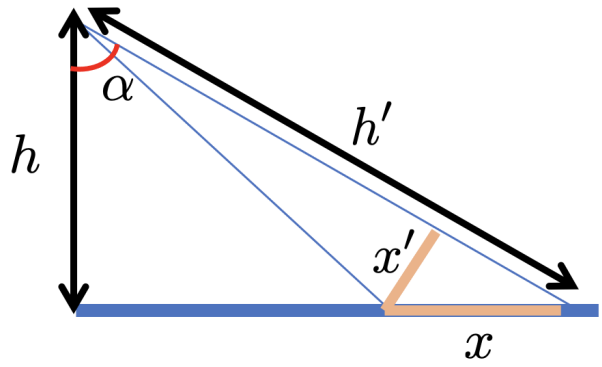
Note that the focal length of the camera f and the size of the physical letter "e" x are unknown. We will not need them explicitly but we will need their product $f \cdot x$.

Let us focus on the left triangle with yellow "e". Applying the perspective projection model to this triangle, can you provide the mathematical formula of the product $f \cdot x$ as a function of known variables? What is its value (in px·cm)?
Note: The correct value will not be counted if the formula is not provided!

We have two more triangles (with red and green letters) that contain additional information. They differ from the one we discussed in the previous question, as they are shifted along x-axis. Let us explore one of these triangles to find out the relation between the variables.

Question 28: *This question is worth 15 points.*

0 1 2 3 4 5 6 7
8 9 10 11 12 13 14 15



In the picture *above* we provide an illustration for one of these triangles. Note that the scale is exaggerated for better visualization.
The base x is shifted along the x-axis by the angle α (see the picture). The "effective" physical size of the letter now is its projection x', while the distance to the camera is h'. Let us assume that the pixel size of x' is equal to a (as for the red letter).
Do the following:

(a) Using trigonometry, derive h' as a function of h and $\cos(\alpha)$ and x' as a function of x and $\cos(\alpha)$.

(b) Apply the perspective projection model to x', and use it with the results of (a) to derive a mathematical formula for $\cos^2(\alpha)$.

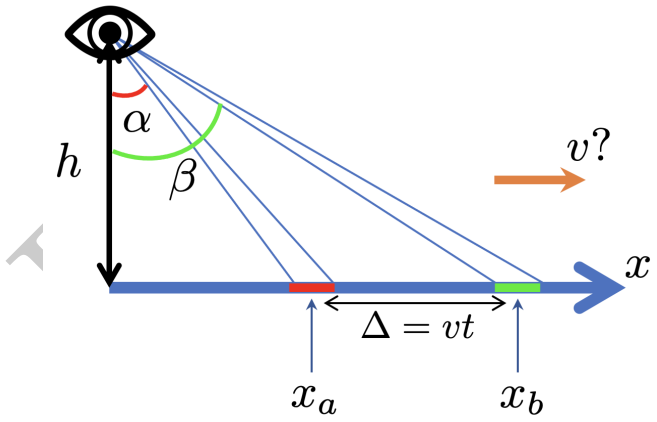
Hint: Since the angular size of the letter is small, we additionally assume that $x' \ll h'$. It implies, for example, that the the distance from the camera to the letter is equal to h' for every point on the letter.

(c) Using the formula $\left[\tan^2\alpha = \frac{1}{\cos^2\alpha} - 1\right]$, derive $\tan^2\alpha$.

(d) Knowing the variables' values, can you say what exactly the angle α is for the red letter "e" (in degrees)?

(e) Apply the same formulas to the green letter (where the pixel size is b) and compute its $\tan^2\beta$.

Almost done! Given a video sequence, we measured the time for the red "e" to reach the line of green "e", it took $t = 5$ sec.



Now the speed could be computed simply by dividing the distance $\Delta = x_b - x_a$ between green and red x-coordinates (see the picture *above*) by time t, but these coordinates are unknown.

Question 29: *This question is worth 12 points.*

☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9 ☐10 ☐11 ☐12

(a) Using trigonometry, derive the coordinates x_a and x_b explicitly through the known variables and the angles α and β (see the picture).

(b) Use the derivations from the previous questions and find the formula and the value of the velocity v in cm/sec.

Note: The formula for speed v must be a function of known h, t, a, b and c.