Exercise 1

Give three examples of problems in computer vision which are formally ill-posed. In each case explain how one or more of Hadamard's criteria for well-posed problems has failed to be satisfied. Illustrate how addition of ancillary constraints or assumptions, even meta-physical assumptions about how the world behaves, enable one to convert the ill-posed problem into a well-posed problem.

Exercise 2

What is accomplished by the lateral signal flows within both plexiform layers of the mammalian retina, in terms of spatial and temporal image processing and coding?

Exercise 3

Why can't any computer vision operations be performed directly on `.jpeg` image data formats?

Exercise 4

In human vision, the photoreceptors responsible for colour (cones) are numerous only near the fovea, mainly in the central ± 10 degrees. Likewise high spatial resolution is only found there. So then why does the visual world appear to us uniformly coloured? Why does it also seem to have uniform spatial resolution? What implications and design principles for computer vision might be drawn from these observations?

Exercise 5

Discuss the significance of the fact that mammalian visual systems send perhaps ten times as many corticofugal neural fibres back down from the visual cortex to the thalamus, as there are ascending neural fibres bringing visual data from the retina up to the thalamus. Does this massive neural feedback projection support the thesis of "vision as graphics," and if so how?

Exercise 6

Give finite difference operators that could be applied to 1-dimensional discrete data (such as a row of pixels) in order to approximate the 1st and 2nd derivatives, $\frac{d}{dx}$ and $\frac{d^2}{dx^2}$.
How would your finite difference operators actually be applied to the row of pixels? What is the benefit of using a 2nd finite difference (or derivative) instead of a 1st finite difference (or derivative) for purposes of edge detection?

Exercise 7

Consider the following 2D filter function $f(x, y)$ incorporating the Laplacian operator that is often used in computer vision:

$$f(x,y) = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) e^{-\left(x^2+y^2\right)/\sigma^2}$$

(a) In 2D Fourier terms, what type of filter is this? (*E.g.* is it a lowpass, a highpass, or a bandpass filter?)

(b) Are different orientations of image structure treated differently by this filter, and if so, how? Which term better describes this filter: *isotropic*, or *anisotropic*?

(c) Approximately what is the spatial frequency bandwidth of this filter, in octaves? [Hint: the answer is independent of σ.]

(d) What is meant by image operations "at a certain scale of analysis?" In this context, define a scale-space fingerprint, and explain the role of the scale parameter.

Exercise 8

The following very useful operator is often applied to an image $I(x, y)$ in computer vision algorithms, to generate a related "image" $g(x, y)$:

$$g(x,y) = \int_\alpha \int_\beta \nabla^2 e^{-\left((x-\alpha)^2+(y-\beta)^2\right)/\sigma^2} I(\alpha,\beta)\ d\alpha\ d\beta$$

where

$$\nabla^2 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)$$

(a) Give the general name for the type of mathematical operator that $g(x, y)$ represents, and the chief purpose that it serves in computer vision.

(b) What image properties should correspond to the zero-crossings of the equation, *i.e.* those isolated points (x, y) in the image $I(x, y)$ where the above result $g(x, y) = 0$?

(c) What is the significance of the parameter σ? If you increased its value, would there be more or fewer points (x, y) at which $g(x, y) = 0$?

(d) Describe the effect of the above operator in terms of the two-dimensional Fourier domain. What is the Fourier terminology for this image-domain operator? What are its general effects as a function of frequency, and as a function of orientation?

(e) If the computation of $g(x, y)$ above were to be implemented entirely by Fourier methods, would the complexity of this computation be greater or less than the image-domain operation expressed above, and why? What would be the trade-offs involved?

(f) If the image $I(x, y)$ has 2D Fourier Transform $F(u, v)$, provide an expression for $G(u, v)$, the 2D Fourier Transform of the desired result $g(x, y)$ in terms of only the Fourier plane variables $u, v, F(u, v)$, some constants, and the above parameter σ.

Exercise 9

Extraction of visual features from images often involves convolution with filters that are themselves constructed from combinations of differential operators. One example is the Laplacian $\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ of a Gaussian $G_\sigma(x, y)$ having scale parameter σ, generating the filter $\nabla^2 G_\sigma(x, y)$ for convolution with the image $I(x, y)$. Explain in detail each of the following three operator sequences, where $*$ signifies two-dimensional convolution.

(a) $\nabla^2 [G_\sigma(x, y) * I(x, y)]$

(b) $G_\sigma(x, y) * \nabla^2 I(x, y)$

(c) $[\nabla^2 G_\sigma(x, y)] * I(x, y)$

(d) What are the differences amongst them in their effects on the image?

Exercise 10

Consider the following pair of filter kernels:

-1	-1	-1	-1	-1	-1
-1	-3	-4	-4	-3	-1
2	4	5	5	4	2
2	4	5	5	4	2
-1	-3	-4	-4	-3	-1
-1	-1	-1	-1	-1	-1

1	1	1	1	1	1
-1	-2	-3	-3	- 2	-1
-1	-3	-4	-4	-3	-1
1	3	4	4	3	1
1	2	3	3	2	1
-1	-1	-1	-1	- 1	-1

1. Why do these kernels form approximately a quadrature pair?
2. What is the "DC" response of each of the kernels, and what is the significance of this?
3. To which orientations and to what kinds of image structure are these filters most sensitive?
4. Mechanically how would these kernels be applied directly to an image for filtering or feature extraction?
5. How could their respective Fourier Transforms alternatively be applied to an image, to achieve the same effect as in (4) above?
6. How could these kernels be combined to locate facial features?

Exercise 11

Define the "Correspondence Problem," detailing the different forms that it takes in stereo vision and in motion vision.

1. In each case, explain why the computation is necessary.
2. What are the roles of space and time in the two cases, and what symmetries exist between the stereo and motion vision versions of the Correspondence Problem?
3. How does the complexity of the computation depend on the number of underlying features that constitute the data?
4. Briefly describe at least one general approach to an efficient algorithm for solving the Correspondence Problem.

Exercise 12

(a) For what size of filter kernel does it become more efficient to perform convolutions by instead computing Fourier Transforms, and why?

(b) For an aligned stereo pair of cameras separated by base distance b, each with focal length f, when a target point projects outside the central axis of the two cameras by amounts α and β:

- What is the computed target depth d?
- Why is camera calibration so important for stereo vision computations?
- Identify 4 relevant camera degrees-of-freedom and briefly explain their importance for stereo vision algorithms.

(c) When trying to detect and estimate visual motion in a scene, why is it useful to relate spatial derivatives to temporal derivatives of the image data? Briefly describe how one motion model works by these principles.

(d) What does the Spectral Co-Planarity Theorem assert about translational visual motion, and how the parameters of such motion can be extracted?

(e) What information about the shape and orientation of an object can be inferred, and how, from the extraction of texture descriptors; and what is the role of prior assumptions in making such inferences?

Exercise 13

1. Contrast the use of linear versus non-linear operators in computer vision, giving at least one example of each. What can linear operators accomplish, and what are their fundamental limitations? With non-linear operators, what heavy price must be paid and what are the potential benefits?

2. When shape descriptors such as "codons" or Fourier boundary descriptors are used to encode the closed 2D shape of an object in an image, how can invariances for size, position, and orientation be achieved? Why are these goals important for pattern recognition and classification?

3. Define the general form of "superquadrics" used as volumetric primitives for describing 3D objects. What are their strengths and their limitations?

Exercise 14

(a) Explain why infering object surface properties from image properties is, in general, an ill-posed problem. In the case of infering the colours of objects from images of the objects, how does knowledge of the properties of the illuminant affect the status of the problem and its solubility?

(b) What surface properties can cause a human face to form either a Lambertian image or a specular image, or an image lying anywhere on a continuum between those two extremes? In terms of geometry and angles, what defines these two extremes of image formation? What difficulties do these factors create for efforts to extract facial structure from facial images using "shape-from-shading" inference techniques?

(c) Explain and illustrate the "Paradox of Cognitive Penetrance" as it relates to computer vision algorithms that we know how to construct, compared with the algorithms underlying human visual competence. Discuss how human visual illusions may relate to this paradox. Comment on the significance of this paradox for computer vision research.

Exercise 15

(a) In a visual inference problem, we have some data set of observed features x, and we have a set of object classes $\{C_k\}$ about which we have some prior knowledge. Bayesian pattern classification asserts that:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$$

Explain the meaning of, and give the name for, each of these three terms:

$P(C_k|x)$
$P(x|C_k)$
$P(C_k)$

(b) Understanding, classifying, and identifying human faces has been a longstanding goal in computer vision. Yet because the face is an expressive social organ, as well as an object whose image depends on identity, age, pose and viewing angle, and illumination geometry, many forms of variability are all confounded together, and the performance of algorithms on these problems remains very poor. Discuss how the different kinds and states of variability (e.g. same face, different expressions; or same identity and expression but different lighting geometry) might best be handled in a statistical framework for generating categories, making classification decisions, and recognising identity. In such a framework, what are some of the advantages and disadvantages of wavelet codes for facial structure and its variability?

(c) Consider the "eigenfaces" approach to face recognition in computer vision.

(i) What is the rôle of the database population of example faces upon which this algorithm depends?

(ii) What are the features that the algorithm extracts, and how does it compute them? How is any given face represented in terms of the existing population of faces?

(iii) What are the strengths and the weaknesses of this type of representation for human faces? What invariances, if any, does this algorithm capture over the factors of perspective angle (or pose), illumination geometry, and facial expression?

(iv) Describe the relative computational complexity of this algorithm, its ability to learn over time, and its typical performance in face recognition trials.

Exercise 16

(*a*) Give three examples of methodologies or tools used in Computer Vision in which Fourier analysis plays a role, either to solve a problem, or to make a computation more efficient, or to elucidate how and why a procedure works. For each of your examples, clarify the benefit offered by the Fourier perspective or implementation.

(*b*) Explain the formal mathematical similarity between the "eigenface" representation for face recognition, and an ordinary Fourier transform, in the following respects:

(*i*) Why are they both called linear transforms, and what is the "inner product" operation in each case?

(*ii*) What is a projection coefficient and an expansion coefficient in each case?

(*iii*) What is the orthogonal basis in each case, and what is meant by orthogonality?

(*iv*) Finally, contrast the two in terms of the use of a data-dependent or a universal (data-independent) expansion basis.

(*c*) How can dynamic information about facial appearance and pose in video sequences (as opposed to mere still-frame image information), be used in a face recognition system? Which core difficult aspects of face recognition with still frames become more tractable with dynamic sequences? Are some aspects just made more difficult?

Exercise 17

(*a*) In pattern classification with two classes, explain how an ROC curve is derived from the underlying distributions. Define a threshold-independent performance metric based on the distributions' moments.

(*b*) When visually inferring a 3D representation of a face, it is useful to extract separately both a shape model, and a texture model. Explain the purposes of these steps, their use in morphable models for pose-invariant face recognition, and how the shape and texture models are extracted and later re-combined.