Exercises for Computer Vision

Exercise 1

Explain why inferring object surface properties from image properties is, in general, an ill-posed problem: some of Hadamard's criteria for well-posed problems are not satisfied. In the case of inferring the colours of objects from images, how does knowledge of the properties of the illuminant affect the status of the problem and its solubility? More generally, illustrate how addition of ancillary constraints or assumptions, even metaphysical assumptions, allows an ill-posed problem to be converted into a well-posed problem.

Exercise 2

In human vision, photoreceptors (cones) responsible for colour are numerous only near the fovea, mainly in the central ± 10 degrees. High spatial resolution likewise exists only there. So then why does the visual world appear to contain colour information everywhere in the field of view? Why does it also seem to have uniform spatial resolution? Why does the world appear stable despite all our eye movements? Discuss some implications for computer vision principles that might be drawn from these observations.

Exercise 3

Present five experimental observations about human vision that support the thesis that "vision is graphics:" what we see is explicable only partly by the optical image itself, but is more strongly determined by top-down knowledge, model-building and inference processes.

Exercise 4

The binary image pixel array on the left below is convolved ($*$) with what operator $\boxed{?}$ to give the result on the right? Specify the operator by numbers within an array, state its relationship to finite difference operators of specific orders, and identify what task this convolution accomplishes in computer vision.

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	**1**	**1**	**1**	**1**	0	0	0
0	0	0	**1**	**1**	**1**	**1**	0	0	0
0	0	0	**1**	**1**	**1**	**1**	0	0	0
0	0	0	**1**	**1**	**1**	**1**	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

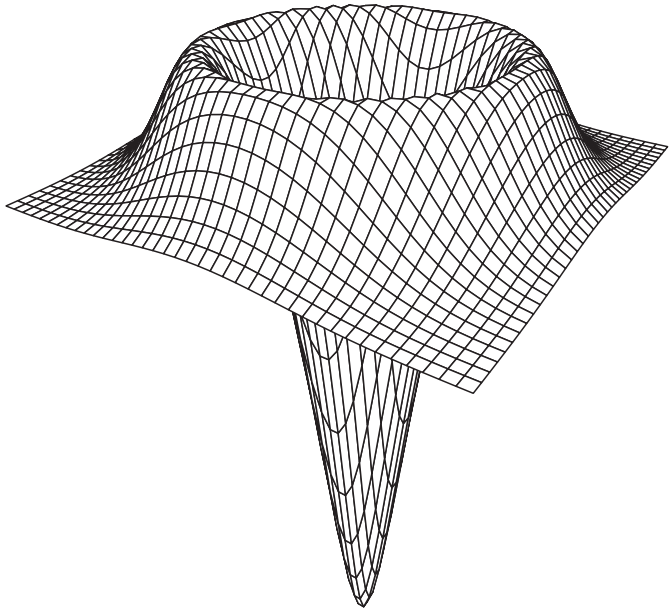
$*$ $\boxed{?}$ $\Rightarrow$

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	**-1**	**1**	0	0	**1**	**-1**	0
0	**-1**	**1**	0	0	**1**	**-1**	0
0	**-1**	**1**	0	0	**1**	**-1**	0
0	**-1**	**1**	0	0	**1**	**-1**	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Exercise 5

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 6 (extension of Exercise 5)

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.

(e) Provide a 3x3 discrete filter kernel array that approximates the Laplacian operator. Explain what the Laplacian might be used for, and what is the significance of the sum of all of the taps in the filter.

Exercise 7

(a) 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.

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

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

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

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

Exercise 8

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 9

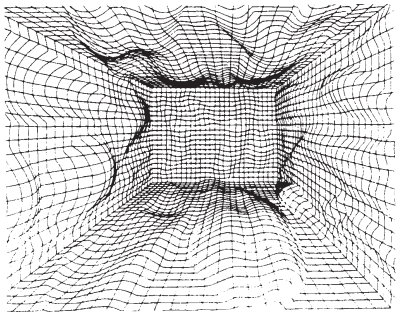
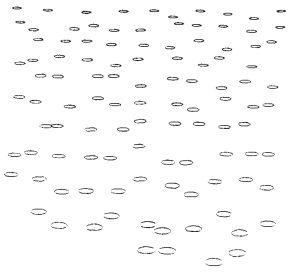
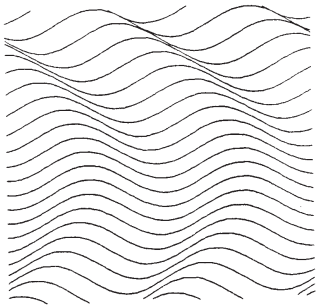
Explain the method of *Active Contours*. What are they used for, and how do they work? What underlying trade-off governs the solutions they generate? How is that trade-off controlled? What mathematical methods are deployed in the computational implementation of Active Contours?

Exercise 10

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.

Exercise 11

Discuss the use of texture gradients as a depth cue in computer vision. How can texture gradients be measured? What role can Fourier analysis play in this? What ancillary "metaphysical" assumptions must be invoked by a vision algorithm in order to make the inference task well-posed and thereby make such computations possible? You may find it helpful to refer to the following figures:

Exercise 12

For a stereo pair of cameras whose optical axes are parallel, separated by base distance b, both having focal length f, suppose a target point projects onto points in the two image planes which are outside the optical axes oppositely by amounts α and β:

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

Exercise 13

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 vision 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 14

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.

Exercise 15

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

Exercise 16

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 classification?

Exercise 17

When defining and selecting which features to extract in a pattern classification problem, what is the goal for the statistical clustering behaviour of the data in terms of the variances within and amongst the different classes? What roles are played by within-class variability and between-class variability?

Exercise 18

Show how Bayesian inference exploits the distinctiveness, or improbability, of observed features to make stronger classification decisions. We have a data set of observed features x, and we have a set of object classes $\{C_k\}$, for each of which we have some prior knowledge about feature likelihood of the form $P(x|C_k)$. Express Bayes' Rule and explain the meaning of terms $P(C_k|x)$, $P(C_k)$, and $P(x)$.

Now explain how Bayesian inference enhances face recognition when a face contains highly distinctive features, as is exploited by caricature (for example a politician's face). Suppose some facial feature x is unusual so its probability $P(x)$ is small, and that for each k^{th} face described as class C_k we know the class-conditional likelihood $P(x|C_k)$ of observing this unusual feature x, but *a priori* all the classes are equiprobable. Use Bayes' Rule to show how correct classification of face C_k given its unusual feature x acquires higher probability $P(C_k|x)$ than if the feature were more common.

Exercise 19

Sketch out an algorithm for shape classification and the construction of shape grammars, involving active contours, codon strings, and indexing. Explain how codon constraints enable a shape grammar to define broad equivalence classes such as "cashew shaped" objects, with invariance to irrelevant transformations such as planar rotations or dilations.

Exercise 20

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 21

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?

Exercise 22

Detecting, classifying, recognising, and interpreting human faces is 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 & viewing angle, and illumination geometry, many forms of variability are all confounded together, and the performance of algorithms on these problems remains rather disappointing. 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 (Haar or Gabor) for facial structure and its variability?

Exercise 23

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

1. What is the rôle of the database population of example faces upon which this algorithm depends?

2. 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?

3. 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?

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

Exercise 24

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?

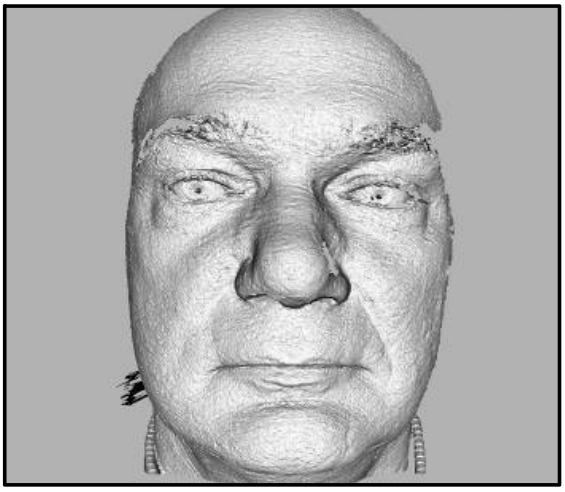
(*iv*) Finally, contrast the two in terms of the use of a data-dependent or a data-independent (universal) expansion basis.

Exercise 25

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?

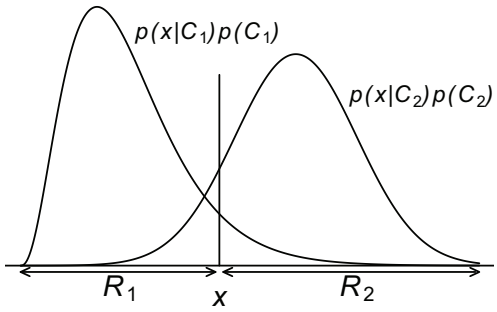
Exercise 26

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.



Exercise 27

A Bayesian classifier assigns visual objects to either one of two classes, C_1 or C_2, by observing x. Prior baseline probabilities are $p(C_1)$ and $p(C_2)$, with sum $p(C_1)+p(C_2)=1$. Observations x have unconditional probability $p(x)$, and class-conditional probabilities of a given observation x are $p(x|C_1)$ and $p(x|C_2)$.



1. Using the above quantities provide an expression for $p(C_k|x)$, the likelihood of class C_k given an observation x.
2. Provide a decision rule using $p(C_k|x)$ and $p(C_j|x)$ for assigning classes based on observations, that will minimise misclassification.
3. Now express your decision rule instead using only the quantities $p(C_k)$, $p(C_j)$, $p(x|C_k)$, $p(x|C_j)$, and relate it to the diagram above.
4. If the classifier decision rule assigns class C_1 if $x \in R_1$, and C_2 if $x \in R_2$ as shown in the figure, what is the total probability of error?
5. If classifier decisions are made by computing functions $y_k(x)$, $y_j(x)$ of the observations x and assigning class C_k if $y_k(x) > y_j(x) \quad \forall j \neq k$, for example $y_k(x) = p(C_k|x)$, what are such functions $y_k(x)$ called?

Exercise 28

Discuss the significance of the fact that typically in mammalian visual systems, there are almost ten times more corticofugal neural fibres sent 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 29

Discuss the theory of vision as model building, hypothesis generation and testing, and knowledge-based processing, in light of the paradoxical figure on the right. What do we learn from bistable or rivalrous percepts? Discuss how top-down context information should drive the integration of low-level data into meaningful visual wholes.