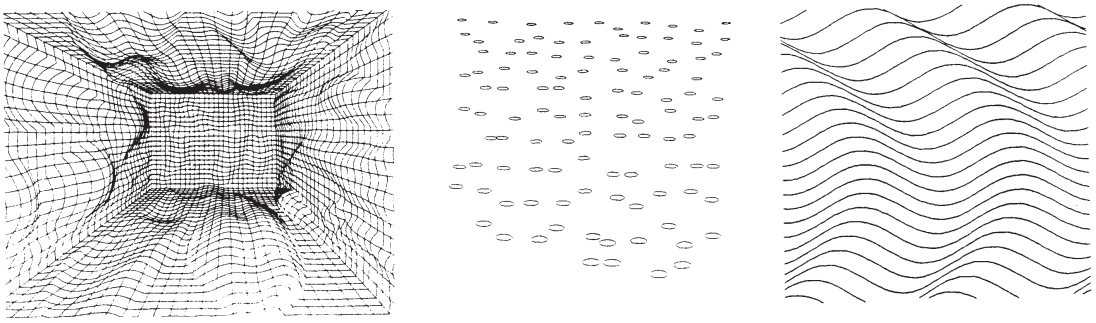
Exercise 10

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

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.

Answer to Exercise 10

Any three from the following list would do:

1. Convolution of an image with some operator, for example an edge detection operator or feature detecting operator, is ubiquitous in computer vision. Convolution is computationally costly and slow if done "literally," but it is very efficient if done instead in the Fourier domain. One merely needs to multiply the Fourier transform of the image by the Fourier transform of the operator in question, and then take the inverse Fourier transform to get the desired result. For kernels larger than about (5×5), the benefit is that the Fourier approach is vastly more efficient.

2. The Fourier perspective on edge detection shows that it is really just a kind of frequency-selective filtering, usually high-pass or bandpass filtering. For example, applying the ∇^2 second-derivative operator to an image is equivalent to multiplying its Fourier transform by a paraboloid, $\mu^2 + \nu^2$, which discards low frequencies but emphasises high frequencies, in proportion to their square.

3. Texture detection, and texture segmentation, can be accomplished by 2D spectral (Fourier) analysis. Textures are well-defined by their spatial frequency and orientation characteristics, and these indeed are the polar coordinates of the Fourier plane.

4. Motion can be detected, and its parameters estimated, by exploiting the "Spectral co-planarity theorem" of the 3-D spatio-temporal Fourier transform.

5. Active contours as flexible boundary descriptors ("snakes") can be implemented through truncated Fourier series expansions of the boundary data.

Answer to Exercise 11

Most surfaces are covered with texture of one sort or another, which can serve as a cue to surface shape because of the foreshortening it undergoes as it follows the shape of the object *if* one can assume that it has some uniform statistics along the surface itself. The examples given illustrate the inference of surface slant and of 3D surface shape from texture cues when they are combined with the assumption of texture uniformity on the surface itself. Texture information (especially texture gradients) can be used to infer 3D surface shape and orientation of objects, as well as contributing to object classification and identity.

A natural way both to detect the quasi-periodicity of texture and to estimate its pitch (and thereby the pitch gradients) is by Fourier analysis, and Fourier-related methods such as wavelet techniques. However, these must be localised, in order to detect variation across space (across the surface) of the quasi-periodicity and directionality that are central metrics returned by Fourier analysis. Thus we can generate textural statistics of pitch variation within windows, or patches, across the surface. When combined with the ancillary "metaphysical" assumption that the texture is actually uniform on the surface itself, we can directly infer surface slant and local shape variation from these spectral texture metrics.

Answer to Exercise 12

The aligned stereo pair of cameras with parameters as specified would compute a depth of:

$$d = fb/(\alpha + \beta)$$

Camera calibration is critically important for stereo vision because all inferences depend directly on the geometric parameters of the system. Each camera has 6 degrees-of-freedom describing its disposition (3 spatial coordinates X,Y,Z and 3 Euler rotation angles), together with a focal length. The most important relative parameters are: (1) the base of separation b between the two cameras; (2) their actual alignment, if in fact their optical axes are not parallel; (3) their focal length f (normally fixed); and any rotation around each camera's optical axis, even if the optical axes are strictly parallel, as this affects the solution to the Correspondence Problem.

Answer to Exercise 13

1. Stereo vision requires that corresponding object points, in two images acquired from slightly different vantage points, be associated to each other in order to make possible the measurement of their relative disparity in the image plane and thereby a computation of their depth in space relative to the focal plane.

   Motion vision requires that corresponding object points, in two images acquired from the same vantage point but at slighly different moments in time, be associated to each other in order to make possible a measurement of their relative displacement and thereby a calculation of the motion vector over this interval of time.

2. These two cases of the Correspondence Problem are symmetrical, with the roles of space and time simply interchanged. In stereo vision, the two image frames are simultaneous in time but displaced in space. In motion vision, the two image frames are from the same vantage point (coincident in space) but displaced in time.

3. The complexity of the computation depends on the number of possible pairings of individual features, which varies quadratically with the number of features present in each of the two image frames constituting the data. This is because in principle, every feature in one image frame could be associated with every possible feature in the other image frame. Hence N features in each generate up to NxN individual pairing hypotheses about which feature from Frame 1 goes with which feature from Frame 2. (N could even be as large as the number of pixels in either image, but in practice it would be more common to select the sorts of sparse features used in SIFT, such as oriented edges, corners, etc.)

4. One way to make this computation more efficient is by stochastic relaxation, in which large-deviation (large displacement) correspondence hypotheses no longer need to be considered once enough evidence has accumulated for a more conservative solution. The amplitude of the deviations may be slowly diminished in a way that corresponds to decline of temperature in "annealing" algorithms. It is also helpful to approach the problem in a "course to fine" pyramid, first using large coarse features (e.g. from a highly blurred version of the image) so that only a few matches need to be evaluated, with coarse alignment; then when the winning hypothesis at that resolution is selected, proceed to higher resolution to match a larger number of (finer) features; evenutually converging on finer correspondences between the full set.

Answer to Exercise 14

When there is motion of objects in a scene, there is a relationship between the spatial derivatives (e.g. gradient) of image structures such as edges, and the temporal derivatives of those same points, over successive image frames. Estimating both the local spatial and temporal derivatives allows the velocity vector $\vec{v}$ to be inferred, through the following relationship over an image sequence $I(x, y, t)$:

$$-\frac{\partial I(x,y,t)}{\partial t} = \vec{v} \cdot \vec{\nabla} I(x,y,t)$$

Thus the ratio of the local image time-derivative to the spatial gradient gives an estimate of the local image velocity (in the direction of the gradient).

An alternative way to exploit such measured derivatives for motion estimation is used in "Dynamic zero-crossing models" by finding the edges and contours of objects and then taking the time-derivative of the Laplacian-Gaussian-convolved image $I(x, y, t)$

$$-\frac{\partial}{\partial t}\left[\nabla^2 G_\sigma(x,y) * I(x,y,t)\right]$$

in the vicinity of a Laplacian zero-crossing. The amplitude of the result is an estimate of speed, and the sign of this quantity determines the direction of motion relative to the normal to the contour.

Answer to Exercise 15

The Spectral Co-Planarity Theorem asserts that rigid translational visual motion has the consequence that in the 3D spatio-temporal frequency domain, all the spectral energy (which would normally form a cloud, a 3D distribution in this 3D Fourier domain) collapses into 2D, on an inclined plane going through the origin of Fourier space. The elevation of that spectral plane specifies the speed of motion, and its azimuth corresponds to the direction of motion. More specifically, the spherical coordinates $(\theta, \phi, 1)$ of the inclined spectral plane's unit normal correspond to the speed (ϕ) and direction (θ) of motion. Thus, detecting the "geometry" of this energy distribution (its regression plane coordinates) amounts to detecting the speed and direction of visual motion.

Answer to Exercise 16

Shape descriptors such as "codons" or Fourier boundary descriptors encode the properties of a closed 2D shape's boundary over a domain from 0 to 2π (the total perimeter of a closed curve, in radians). This means that the same shape in different sizes would always produce the same code, and so this achieves size invariance. Likewise, the description is always relative to the center of the closed curve, and so it achieves invariance to position in two dimensions. Finally, a rotation of the 2D shape in the plane amounts to merely a scrolling in angle of the code for the shape in terms of the boundary descriptors; and in the case of codons, the lexicon entry for the shape (defined by a grammar of zeroes of curvature and inflexion points) is completely unaffected by rotations.

The achievement of these sorts of invariances, and indeed of invariance to some non-affine distortions and deformations of the shape, are important steps in pattern recognition and classification. They serve to diminish the unimportant elements of "within-class variability," and to give a compact description that sharpens the "between-class variability." These are central goals for pattern classification.

Answer to Exercise 17

Codon strings can constitute the elements of a *shape grammar*, based on the combinatorics of curvature polarity and inflexions. Constraints on codon strings for closed curves are very strong. For example, while sequences of (say) 6 codons have $5^6 = 15,625$ possible combinations, these make only 33 generic shapes. The ordinal relations among singular points of curvature (maxima, minima, and zeroes) remain invariant under translations, rotations, and dilations. The inflexion (a zero of curvature) of a 3D curve is preserved under 2D projection, thereby guaranteeing that ordinal relations among the extrema of curvature will also be preserved when projected to an image.

Exercise 18

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 19

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.

Answer to Exercise 18

In classification problems, visual features should be chosen which minimise within-class variability and maximise between-class variability. This allows the diameters of clusters in feature space to be small compared with the spacings amongst the clusters, thereby minimising overlap and thus classification errors. One can select (or define) features for the classifier with the goal of maximising the resulting between-class variability whilst minimising the resulting within-class variability. In the case of face recognition, the problem is that different faces in the same frontal pose and neutral expression may resemble each other more closely than do different images of the same face when seen in different pose angles, illumination geometries, or different expressions. Defining a feature such as "how many eyes does this face possess" would not be a good discriminator, since there is little between-class variability in that particular dimension, but choosing some secondary facial structure may be a good source of such discriminating variability.

Answer to Exercise 19

Bayesian pattern classification asserts that:

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

The terms have the following meanings:

- $P(C_k|x)$ is the *posterior probability* of object class C_k, given observation of features x. This is the outcome of the Bayesian inference calculation: it assigns a probability to each of the possible classification hypotheses $\{C_k\}$.
- $P(x|C_k)$ is the *class-conditional likelihood* that the features x would be observed, if the object belonged to class C_k. This is one of the ways in which Bayesian inference exploits expert knowledge about the problem domain.
- $P(C_k)$ is the *prior,* or *unconditional likelihood* of class C_k. In the absence of data this is the plausibility of the hypothesis C_k, and it is another way in which Bayesian inference incorporates prior expert knowledge. When Bayesian inference is applied in an iterative way for learning over time as more data arrives, the last calculated posterior probability $P(C_k|x)$ can be used as the new prior $P(C_k)$.
- $P(x)$ is the simple probability of observing features x. It can be calculated by summing $P(x|C_k)P(C_k)$ over all the k classes C_k.

It is obvious from the rôle of $P(x)$ in the denominator of Bayes' Rule above that if $P(x)$ is small since unusual feature x is rarely seen except with C_k, we will infer the class C_k with higher probability $P(C_k|x)$ as a result.