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.

Answer to Exercise 20

The "Paradox of Cognitive Penetrance" refers to the fact that the visual tasks that humans are particularly skilled at, such as face recognition, visual learning, navigation, and solving Correspondence Problems, are performed without our having an understanding of how we do them. In contrast, the tasks for which we have an in-depth theoretical understanding and which we know how to write algorithms for, are often tasks that we humans are rather poor at performing, such as numerical operations and mathematical transformations.

The systematic geometrical illusions which occur in the human visual system suggest that fidelity to image properties is not always a goal of biological visual algorithms. We are aware of the illusions but we don't know why they occur. In machine vision today, it is difficult to imagine trying to design algorithms which would intentionally make systematic errors; and yet arguably the human visual illusions are consequences of valuable adaptive strategies.

The significance of the Paradox of Cognitive Penetrance is that the prospects for "reverse engineering" human visual faculties may be reduced by the difficulty of gaining insight into how we actually do what we do. A further implication is that machine vision algorithms, even if successful, are likely to adopt quite different strategies than the biological ones.

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.

Answer to Exercise 21

The physical photonic properties of a surface determine how it scatters light, e.g. over a broad range of angles (a Lambertian surface) or only over a narrow range of angles obeying Snell's Law that angle of emission equals angle of incidence between an illuminating ray and the local surface normal (a specular, or mirror-like, surface). Any surface can be described as lying somewhere on a continuum between these two extremes. These different surface behaviours make it difficult to interpret image data in terms of surface shape, since the scattering angles for reflected light depend on unknown factors. In the case of face images, the relative wetness or oilyness of the skin at a given moment can transform the face from a Lambertian surface to a specular surface, thus confounding "shape-from-shading" methods for inferring the facial structure.

Answer to Exercise 22

The central issue in pattern recognition is the relation between within-class variability and between-class variability. These are determined by the degrees of freedom spanned by the pattern classes. Ideally the within-class variability should be small and the between-class variability large, so that the classes are well separated. In the case of encoding faces for identity, one would like different faces to generate face "codes" that are as different from each other as possible, while different images of the same face should ideally generate similar codes across conditions. Several recent investigations of how well this goal is achieved have studied the invariances in face coding schemes under changes in illumination, perspective angle or pose, and expression. Their results have tended to show that there is greater variability in the code for a given face across these three types of changes, than there is among the codes for different faces when these three factors are kept constant.

When there is variability across two or more dimensions (let us say both face identity and facial expression), then discriminability can benefit from variability *within* a class of the other dimension, but not *between* classes of the other dimension. For example, facial expressions are more reliably distinguished if there is large variation among the different expressions generated by a given face, but small variation in how a given expression is generated among different faces.

The general principle is to use the observed dimensions of variability to find clusters and create categories in such a way as to minimise the within-class variability while simultaneously maximising the between-class variability for the particular task.

Advantages of wavelets for face coding include the fact that most of the major facial features (lips, eyes, etc.) are well described by just a very small number of suitably-chosen wavelets, particularly the smoothly-varying 2D Gabor wavelets. An advantage of alternative binary Haar wavelets is that their convolutions can be implemented simply by additions instead of multiplications, and a cascade of them as "weak classifiers" fits well with the AdaBoost learning framework, as demonstrated by Viola & Jones. An advantage of 2D Gabor wavelets over more traditional feature descriptors (edges, lines, blobs) is that major facial structure is continuous-tone and differentiable, and undergoes continuous deformation, which Gabor wavelets can well-accommodate but more punctate feature descriptors cannot. A disadvantage of wavelet descriptors is that they do not naturally generate translation-invariant (or size or orientation invariant) codes, and they are 2D (image-based) rather than 3D (volumetric solid based) descriptors. The latter may be more appropriate since faces are surfaces of 3D solids (heads) and they project different 2D images with rotations in 3D. So, exhaustive iterations at multiple size scales and for multiple pose angles are required instead by wavelet methods, at cost of speed.

Answer to Exercise 23

(*i*) Any given presenting face is represented in terms of factors precomputed from the database population of example faces, by a process called Principal Components Analysis (see answer to (*ii*) below). The database represents "knowledge" about what human faces are like, and what some of their main forms of variability are. Thus the population database should have the same kinds of diversity as the images that will be presented for recognition.

(*ii*) A statistical procedure called Principal Components Analysis finds the major forms of variation among the database face images. Using linear algebraic methods (diagonalizing a covariance matrix to find its eigenvectors and their corresponding eigenvalues), a set of "eigenfaces" are precomputed (these are the eigenvectors). They have the mathematical property that the greatest amount of variability in the database is spanned by the smallest number of basis vectors, up to any desired degree of approximation, when these eigenvectors are used as the representation. Any given face is represented simply as a linear combination of these eigenfaces – usually about 20 to 40 – and that short sequence of numbers (the eigenvalues) is the distinguishing code for a particular face.

(*iii*) A strength of the method is that the representation for a given face is very compact, and so searches can be performed at great speed. A second strength is that the basis vectors (eigenfaces) are orthogonal, ordered by importance, and they capture the greatest amount of variability in the smallest number of terms. The weaknesses of the method are that: (1) the representation is image-based, i.e. two-dimensional appearance based, and so it captures no invariances for pose angle or perspective angle; (2) similarly it has no invariance for changes in facial expression; and (3) it is very sensitive to changes in illumination; – so much so that usually the 2nd or 3rd eigenface is just an illumination factor (e.g. above versus below), and so if a person is enrolled under one type of illumination he tends not to be recognized under another. The same problem occurs if size normalization is imperfect.

(*iv*) The algorithm is efficient because the Principal Components Analysis of the database is precomputed off-line; any given presenting face then only needs to be projected onto each of the precomputed eigenfaces (a simple series of inner product operations). The algorithm can learn over time as more faces are encountered, simply by continuing the process of PCA as the database grows. However, its lack of fundamental invariances is its fatal flaw. In trials of the method, error rates of 43% to 50% have been found either when there are large changes in illumination geometry, or for images taken more than one year apart.

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?

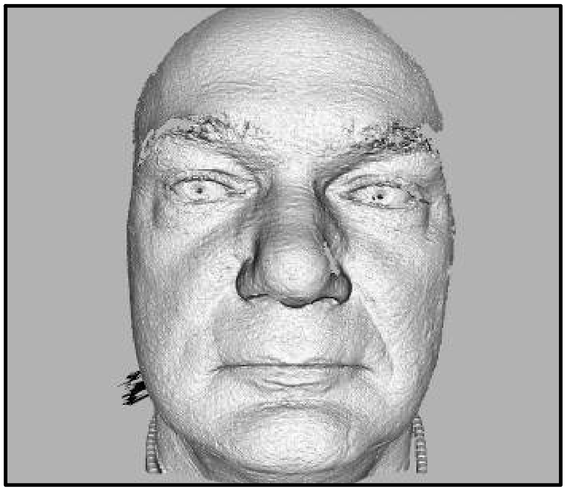
(*iv*) Finally, contrast the two in terms of the use of a data-dependent or a data-independent (universal) expansion basis.

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.



Answer to Exercise 24

Mathematical similarities between eigenfaces and Fourier transforms:

(*i*) They are both linear integral expressions, taking an inner product between an image and some kernel or basis function (an eigenface or a complex exponential). The original data (face or signal) is then represented as a linear combination of those basis functions.

(*ii*) In each case, the projection coefficient is the result of the inner product mentioned above. When those are used as expansion coefficients by re-multiplying them by the same corresponding basis functions, the result would be the original data or face (or an approximation to it, if the complete set was not used).

(*iii*) The orthogonal basis for eigenface computations consist of the principal components that emerged from a Karhunen-Loëve Transform on a database of faces that was used for training. (Each eigenface is an image, that captures some aspects of facial structure.) For a Fourier transform, the orthogonal basis is a set of complex exponentials. Orthogonality means that the inner product of any two different members of the corresponding set is zero.

(*iv*) The eigenface representation does not use a universal and independent expansion basis (like the complex exponentials of the Fourier transform), but rather a data-dependent basis, that must be computed from some training data. Thus, it lets the available data determine the terms of the representation, instead of using always the same universal and pre-determined set of functions.

Answer to Exercise 25

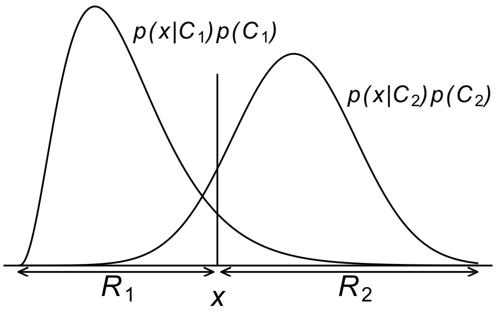
The hardest part of building a system for face recognition is that the image data depends so strongly on the pose angle of the face, or equivalently on the viewing angle, as well as on illumination geometry and expression. Having data that consists of video sequences creates the possibility of incorporating samples with all these forms of variation, for the same person. It is thus intrinsically richer, and spans the dimensions over which we must construct invariants. One way to try to achieve this is to build a 3-dimensional representation of the face, and to perform recognition in terms of that model, rather than using impoverished 2-dimensional still frame "appearance based" representations. But the added difficulty of this approach is that it amounts to "inverse optics:" building a 3-dimensional model from a 2-dimensional image (or a sequence of them). Besides the enormous computational and memory requirements (up to 1 GB for each such full 3D model), this is inherently an ill-posed problem.

Answer to Exercise 26

A 3D shape model is extracted from a face by various means, which may include laser range-finding (with millimetre resolution); stereo cameras; projection of structured light (grid patterns whose distortions reveal shape); or extrapolation from a multitude of images taken from different angles (often a 4×4 matrix). The size of the data structure can be in the gigabyte range, and significant time is required for the computation. The texture model is the photographic appearance itself but expressed in the coordinates of the shape model, so it is possible to project the texture (tone, colour, features, etc) onto the shape and thereby generate models of the face in different poses. Clearly sensors play an important role here for extracting the shape model, but it is also possible to do this even from a single photograph if sufficiently strong Bayesian priors are also marshalled, assuming an illumination geometry and universal aspects of head and face shape. In order to account for variations in illumination, shadows, and specular reflections, the algorithm simulates the process of image formation in 3D space, using computer graphics. The algorithm also relies on *learned* knowledge about the nature of faces as 3D objects, so that the texture mapping is appropriately projected to the 3D shape model. This is one illustration of how computer vision now incorporates methods from machine learning.

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?

Answer to Exercise 27

1. Likelihood of class C_k given an observation x:

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

2. We minimise the probability of misclassification if we assign each observation x to the class with the highest posterior probability. Assign x to class C_k if:

$$p(C_k|x) > p(C_j|x) \quad \forall j \neq k$$

3. Since the denominator in the answer to (i) was independent of k, we can rewrite the minimum misclassification criterion simply as: Assign x to class C_k if

$$p(x|C_k)p(C_k) > p(x|C_j)p(C_j) \qquad \forall j \neq k$$

   This corresponds in the figure to assigning classes based on observations x by imposing a decision boundary where the curves cross each other.

4. Total probability of error:

$$\begin{aligned} P(\text{error}) &= p(x \in R_2, C_1) + p(x \in R_1, C_2) \\ &= p(x \in R_2|C_1)p(C_1) + p(x \in R_1|C_2)p(C_2) \\ &= \int_{R_2} p(x|C_1)p(C_1)dx + \int_{R_1} p(x|C_2)p(C_2)dx \end{aligned}$$

5. Such functions of the observations x are called discriminant functions.

Exercise 28

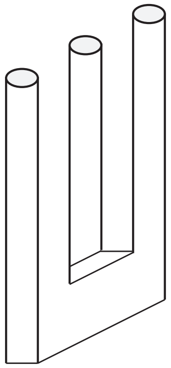
Discuss the significance of the fact that typically in mammalian visual systems, there are almost ten times more corticofugal neural fibres sent back down from the visual cortex to the thalamus, as there are ascending neural fibres bringing visual data from the retina up to the thalamus. Does this massive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

Answer to Exercise 28

When visual data leaves the retina down the million fibres of either optic nerve and reaches its first synapse at the thalamus (or LGN, lateral geniculate nucleus), it is met there by a much larger flood of feedback signals coming back down from the visual cortex. This feedback projection is estimated to contain as many as ten times more fibres than the afferent fibres bringing data up from the retina. One interpretation of this puzzling observation is that vision works by a kind of hypothesis-generation and testing process, in which graphical models are constructed in the brain about the external world and the objects that populate it (and these “graphics” are really what one sees); and the graphics are shaped, constrained, and updated by the 2D image data coming from the retina. Hence we see not image data, but 3D models constructed to be consistent with such data. This is the theory of “vision as [inverse] graphics.”

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.



Answer to Exercise 29

The key concept is that *percepts are hypotheses,* and very few visual processes needed for scene understanding can be accomplished purely in a bottom-up, or data-driven, way. Rather, visual solutions are top-down interpretations that depend greatly on contexts, expectations, and other extraneous factors that go far beyond the data directly available in the image. This reality also underlines the view that *vision is an AI-complete problem*, in the sense that solutions to most or all of the problems in Artificial Intelligence are required before most of the problems in vision can be solved.