Counsel Of Despair

Vision is inverse graphics in that it tries to invert the 3D to 2D projection. Unfortunately this is, strictly, mathematically impossible. Most computer vision problems are not well posed in that:

- No solution necessarily exists
- Solutions are not necessarily unique
- Solutions may not depend continuously on the data

Technology

Spatial resolution is determined by density of CCD array elements and lens properties. Luminance resolution, the number of distinguishable grey levels, is determined by the number of bits per pixel resolved by the digitizer and the SNR of the CCD array.

Frame-grabbers discretize video signals into byte streams.

Biological Visual Mechanisms

Typically neurobiological visual principles inform approaches to machine vision.

Neural activity is fundamentally asynchronous and it is rarely possible to distinguish processing from communication.

The eye consists of 120 million photo-receptors, of which 6 million are cones, arranged in regular hexagonal lattices. Signal flows in the eye occur both longitudinally and laterally. Despite the number of inputs, there are only 1 million “output channels” via the optic nerve, so considerable preprocessing occurs before the brain, which may be summarized as:

- Image sampling by photo-receptors
- Center-surround comparisons implemented by bipolar cells
- Temporal differentiation by amacrine cells
- Separate coding of sustained versus transient image information by different ganglion cells
- Initial colour separation by opponent processing channels

Neurons in the retina can be considered as linear operators or filters, and their behaviour fully understood. The signal flow travels along the optic nerve, splits at the optic chiasm and go via the thalmus. This “relay station” receives 3 times as many efferent fibres from the cortex as it emits afferent fibres from the eyes. Ocular dominance columns attempt to integrate the signals from the two eyes in a way suitable for stereoscopic vision, while simultaneously using orientation columns to detect structures with preferred orientations.

The retina-based receptive fields of neurons are determined experimentally, and enjoy 5 degrees of freedom:

- Position of the field, horizontally and vertically
- Size of the field
- Orientation of excitatory/inhibitory boundaries
- Phase of the receptive field

The fields may be closely described as Gabor wavelets.

The representation of the retina in the brain is retinatopic (adjacent points in the retina project to adjacent points in a cortical map) but there is a distortion to magnification of the fovea by the cortical magnification factor. It has been proposed that this accomplishes a log-polar projection for scale and rotation invariance.

Mathematical Operations

Any image can be represented by a linear combination of basis functions by $f(x,y) = \sum_k a_k \Psi_k(x,y)$. In the case of Fourier, $\Psi_k(x,y) = e^{i(\mu_k x + \nu_k y)}$ where ν andμ are vector spatial frequencies that may be resolved into polar coordinates as $\omega = \sqrt{\mu^2 + \nu^2}$ and $\phi = tan^{-1}(\frac{\nu}{\mu})$. The coefficients a_k are computed as the orthonormal projection of the entire image into the conjugate Fourier component: $a_k = \int_X \int_Y e^{-i(\mu_k x + \nu_k y)} f(x,y) dxdy$.

Shift theorem: $f(x-\alpha, y-\beta) \leftrightarrow F(\mu,\nu)e^{-i(\alpha\mu+\beta\nu)}$, giving translation invariance for the power spectrum of isolated patterns

Similarity theorem: $f(\alpha x, \beta y) \leftrightarrow \frac{1}{|\alpha\beta|}F(\frac{\mu}{\alpha}, \frac{\nu}{\beta})$

Rotation theorem: $f(xcos(\theta) + ysin(\theta), -xsin(\theta) + ycos(\theta)) \leftrightarrow F(\mu cos(\theta) + \nu sin(\theta), -\mu sin(\theta) + \nu cos(\theta))$, so if we work with our Fourier domain (μ,ν) in log-polar space $(r = log(\sqrt{\mu^2+\nu^2}), \theta = tan^{-1}(\frac{\nu}{\mu}))$ then size change becomes translation along r and rotation becomes translation along θ, and we can make these immaterial by considering the power spectrum.

Convolution theorem: if $h(x,y) = \int_\alpha \int_\beta f(\alpha,\beta) g(x-\alpha, y-\beta) d\alpha d\beta$ then $H(\mu,\nu) = F(\mu,\nu)G(\mu,\nu)$, giving an efficient way to compute Fourier expressions after the application of filtering

Differentiation theorem: $\left(\frac{d}{dx}\right)^m \left(\frac{d}{dy}\right)^n f(x,y) \leftrightarrow (i\mu)^m (i\nu)^n F(\mu,\nu)$ so in particular $\nabla^2 f(x,y) \leftrightarrow -(\mu^2+\nu^2)F(\mu,\nu)$ - notice that this emphasises high frequencies and discards the DC part

Edge Detection

This information is useful as:

- Edges demarcate boundaries and parts of objects
- Occlusion edges reveal the geometry of the scene
- Edges may appear in more abstract domains than luminance
- Velocity fields may be understood as the movement of edges
- Aligning edges can be used to solve the correspondence problem effectively

You can find this information computationally by convolving with $\begin{bmatrix} -1 & 1 \end{bmatrix}$ and finding large amplitude or $\begin{bmatrix} 1 & -2 & 1 \end{bmatrix}$ and looking for zero crossings. In two dimensions either directional or non-directional derivatives may be employed. An example discrete isotropic operator is the Laplacian:

-1	-2	-1
-2	12	-2
-1	-2	-1

Operators which sum to 0 are known as filters as they are insensitive to the overall brightness of a scene.

Logan's theorem says that for 1D signals that are band-limited to at most one octave and have no complex zeroes in common with their Hilbert transforms you are able to recover the signal from just its zero-crossings.

Multi-scale Analysis

Multi-scale analysis may be used with edge detection as non-redundant structure typically exists in images at all scales. Marr proposed that the image be convolved with a multi-scale family of isotropic blurred second derivative filters, retaining only their zero-crossings. This can be concretely implemented by the operator $\nabla^2 \left[G_\sigma(x,y) \star I(x,y)\right] = G_\sigma(x,y) \star \nabla^2 I(x,y) = \left[\nabla^2 G_\sigma(x,y)\right] \star I(x,y)$ (with the last being the preferred version).

The Gaussian-Laplacian approach tends to be very noise-sensitive, and more sophisticated non-linear detectors have been developed. Furthermore, it is not clear how to generalize the constraint of one-octave band-limiting to 2D signals, and the zeroes of a 2D signal are not countable.

Causality is the property that edges at lower resolutions must be caused by edges in the underlying data, and are not artifacts of the blurring process. Fingerprint theorems show that the Gaussian blurring operator uniquely possesses this property.

A plot showing the evolution of zero-crossings in the image after convolution with a linear operator as a function of the scale of that operator is called scale-space. A mapping of the edges in an image is called a scale-space fingerprint.

Models

Active contours are one expression of a model-fitting approach that relies jointly on a data term (model-input similarity) and a cost term (model complexity). Iterative numerical methods (regularization methods) exist that optimize a functional that is a linear combination of the two terms: $argmin_m \int ((M-I)^2 + \lambda (M_{xx})^2) dx$.

The family of filters that uniquely achieve the lowest possible conjoin uncertainly in both space and Fourier domains are Gabor wavelets: $f(x) = e^{-i\mu_0(x-x_0)} e^{-\frac{(x-x_0)^2}{\alpha^2}}$, $F(x) = e^{-ix_0(\mu-\mu_0)} e^{-(\mu-\mu_0)^2 \alpha^2}$. Such functions are non-orthogonal and hence the coefficients are hard to obtain. When they are parametrized to be self-similar (dilates and translates of each other) they constitute a wavelet basis, e.g. $\Psi_{mpq\theta}(x,y) = 2^{-2m} \Psi(2^{-m}(x\cos(\theta) + y\sin(\theta)) - p, 2^{-m}(-x\sin(\theta) + y\cos(\theta)) - q)$.

By taking the modulus of a facial image after convolution with complex-valued 2D Gabor wavelets key features may be detected: this is known as a quadrature demodulator network.

Texture

Texture is a cue to surface shape and image segmentation. It is defined by the existence of certain statistical correlations across the image, with an underlying notion of quasi-periodicity.

The detection of periodicity is best done by Fourier methods. However, the usual exponential eigenfunctions are globally defined so in order to recover local information you typically "window" the sinusoids. The optimal set of windowing functions are Gaussians due to their optimal spatial/spectral localization. Hence the final basis used is 2D Gabor wavelets. Edge detection on the modulus of the Gabor coefficients can detect textured regions.

Colour is difficult to recover because wavelengths received depend as much on the illuminant as upon the spectral reflectances of the surface. Since $R(\lambda) = I(\lambda)O(\lambda)$, some have proposed searching for specular regions in the image where reflected light would be a faithful estimate for $I(\lambda)$. A more robust approach is Retinex, which works on the basis that the colors of object or areas in a scene are determined by their surrounding spatial context. A sequence of ratios computed across object boundaries enables the illuminant to be algebraically discounted.

Correspondence And Motion

Stereoscope disparity results in the images from the left and right eyes differing. Making use of this disparity to infer depth is called the correspondence problem.

Current algorithms for determining correspondence require large searches for matching features under a large number of permutations. A multi-scale image pyramid can be used to guide this search at successively finer scales to improve efficiency. Once feature correlation has been found, $d = \frac{fb}{\alpha+\beta}$ where f is the camera focal length, b is the base of triangulation and α and β are the disparities of the projections of the object in the two images relative to the respective optical axis.

For motion vision we need to solve the correspondence problem for two images coincident in space but acquired with a temporal displacement. Requirements include the need to infer 3D trajectories, make local velocity estimations, disambiguate object from contour motion and assign more than one velocity vector to a given region!

Intensity gradient models assume time derivative is related to local spatial gradients due to velocity $\bar{v}$: $-\frac{\delta I(x,y,t)}{\delta t} = \bar{v}\vec{\nabla}I(x,y,t)$

Dynamic zero-crossing models measure velocity by finding edges and contours and then applying the time derivative in the vicinity of a zero crossing: $-\frac{\delta}{\delta t}(\nabla^2 G_\sigma(x,y) \star I(x,y,t))$

Spatio-temporal correlation models detect motion by observing the most likely correlation between the time-separated images, realized as a pair of coordinates from which the velocity can be calculated. This has been supported somewhat by biological investigation of the visual system of the fly.

Spatio-temporal spectral models detect and measure motion purely by Fourier means, exploiting the fact that motion creates a covariance in the spatial and temporal spectra of the image $I(x,y,t)$ where $F(\omega_x,\omega_y,\omega_z) = \int_X \int_Y \int_T I(x,y,t)e^{-i(\omega_x x+\omega_y y+\omega_t t)}dxdydt$. Motion detection occurs by filtering the image sequence in space and time and observing that tuned spatio-temporal filters whose center frequencies are co-planar in this space are activated together. This is a consequence of the spectral co-planarity theorem, which says that since $I(x,y,t) = I(x - v_x t_o, y - v_y t_0, t - t_0)$, $F(\omega_x,\omega_y,\omega_t) \neq 0$ iff $\omega_x v_x + \omega_y v_y + \omega_t = 0$. The spherical coordinates of the normal of the plane correspond to the speed and direction of motion.

Surfaces

Albedo of a surface is the fraction of the illuminant that is re-emitted from the surface in all directions.

Lambertian surfaces are pure matte, i.e. have no specular component.

Specular surfaces are locally mirror-like and obey Snell's law.

The reflectance map is a function $\phi(i,e,g)$ where i is the illuminant angle, e is the reflected angle and g is the angle between the two that specifies the fraction of incident light reflected per unit surface area, per unit solid angle in the direction of the camera. For Lambertian surfaces, $\phi(i,e,g) = cos(i)$. For Lunar surfaces, $\phi(i,e,g) = \frac{cos(i)}{cos(e)}$. For specular surfaces $\phi(i,e.g) = \begin{cases}1 & g = i+e \\ 0 & g \neq i+e\end{cases}$. Typical surfaces are a blend and are governed by $\phi(i.e.g) = \frac{s(n+1)(2\cos(i)\cos(e)-\cos(g))^n)}{2} + (1-s)\cos(i)$, where s is the fraction of light emitted specularly and n is the sharpness of the specular peak.

Shape Description

Cues to surface shape are texture, colour, stereo, motion and shading information. However, it is an inherently ill-posed problem as many ambiguous factors have to be resolved, such as surface reflectance, geometry, material and illuminant geometry.

Closed boundary contours can be represented by their curvature map: $\theta(s) = \lim_{\Delta s \to 0} \frac{1}{r(s)}$ where $r(s)$ is the limiting radius of a circle that best fits the contour at position s and Δs is the arc length. This is position and orientation independent, scales easily and represents mirror symmetry by a sign change. Additionally, these maps can be expanded with basis functions.to generate a description which is rotation, translation and dilation invariant. Grammars of such invariant shapes are called codon libraries.

The 2.5-dimensional sketch is a 2-dimensional image with surface normals assigned to each point in the image domain.

Solids can also be represented as the unions and intersections of generalized superquadric objects which are defined by equations of the form $Ax^\alpha + By^\beta + Cz^\gamma = R$. This allows volumetric descriptions of the objects in a scene by just giving a list of 3D parameters and relations.

Deformable parametric models fit human recognisable parameters to the models for the purposes of lossy coding or customization of an avatar.

Perceptual Psychology

Recent developments include the idea of a process grammar which models objects and shapes in terms of their morphogenesis.

Percepts can be considered as hypotheses: top-down interpretations that depend greatly on contexts, expectations and other extraneous factors beyond the stimulus.

Agnosias are failures of recognition that result from brain injury. They include things such as the loss of ability to recognise faces but no other objects, loss of colour vision, loss of ability to see in 3D and the inability to simultaneously see more than one thing.

Bayesian Analysis

Bayesian statistics provide a means for integrating prior information with empirical information gathered from incoming data. This is especially relevant in computer vision, where there are many sources of uncertainty. The governing equation is $p(H|D) = \frac{p(D|H)p(H)}{p(D)}$, with the old posterior iteratively becoming the new prior.

Statistical decision theory describes a decision environment that recognises similarity between "different" patterns and differences between "similar" patterns:

	Actually Same	Decision "Same"
Hit	$\surd$	$\surd$
Miss	$\surd$	$\times$
False Alarm	$\times$	$\surd$
Correct Reject	$\times$	$\times$

The criterion for similarity should be set so as to minimize the expected cost of errors. If both types of errors have the same cost then this will be where this causes the area under the probability density curves to be equal. You can derive a Receiver Operating Characteristic which plots the hit rate against the false alarm rate for a range of thresholds.

The decidability of the signal detection task is defined as $d' = \frac{|\mu_2 - \mu_1|}{\sqrt{\frac{1}{2}(\sigma_2^2 + \sigma_1^2)}}$ where μ_i and σ_i are the characteristics of the respective distributions.

Bayesian classifiers take into account the prior probabilities of the possible classifications. The minimum misclassification criterion is that $\forall j \neq k . P(x|C_k)P(C_k) > P(x|C_j)P(C_j)$ where C_i is class i. This can be satisfied by assigning an x to the class with the highest posterior probability. However, in situations where error costs differ this misclassification criterion may not be appropriate.

Discriminant functions are functions $y_k(x)$ associated with each class C_k such that an observation x is assigned to that class iff $\forall j \neq k . y_k(x) > y_j(x)$. Decision boundaries between regions are defined by those loci where $y_k(x) = y_j(x)$.

Face Detection

The central issue in pattern recognition is the relation between within-class and between-class variability. Often there is greater variability in the code for a given face across changes in the illuminant, angle or expression than for different faces with these factors constant, which leads to real-world error rates approaching 50%.

For face detection, identification and expression interpretation for the problem of identifying distinct expressions, within-class variability is desirable and between-class variability undesirable. Conversely, for interpreting expressions in the classes of same/different faces, within class variability is desirable and between class variability undesirable.

Face detection is a harder problem than face recognition and current leading approaches rely just on skin hue!

Template-matching face recognition algorithms store an array of size-invariant pictures of faces in a number of pose angles and match on a pixel-by-pixel basis.

Eigenfaces work with a Karhunen-Loeve Transform of a large database of faces to define all faces as linear combinations of the "most likely" face basis functions. It is limited since many of the principle components just extract shading variations and lack invariance to illumination, pose angle and size!

Wavelets can be used for face recognition as due to their localization they can track changes in facial expression in a local way: faces are a kind of texture! To allow for deformations associated with changes in pose angle or expression, these "Gabor jets" are placed on a deformable graph that tolerates distortions relative to fiducial points, but performance is comparable to that of Eigenfaces.

Focus today is on modelling faces as three-dimensional objects and fitting these models to the percepts.

Motion energy models can be used to extract motion signatures from parts of faces and classify these as expressions.