(2nd Supplementary Slides: Computer Vision)

Professor John Daugman

University of Cambridge

Computer Science Tripos, Part II
Lent Term 2015/16

Active Contours

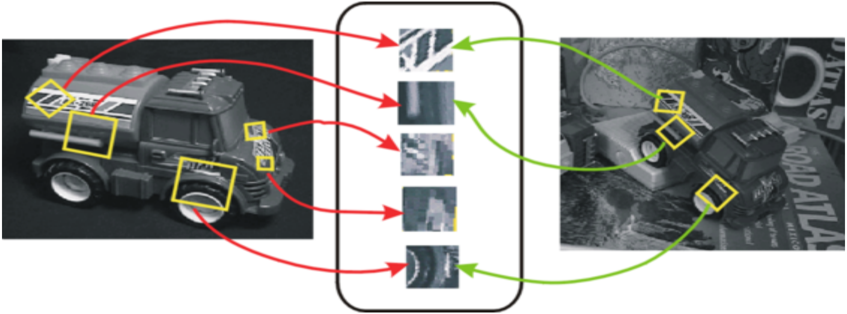
- Match a deformable model to an image, by "energy minimisation"
- Used for shape recognition, object tracking, and image segmentation
- A deformable spline (or "snake") changes its shape under competing forces: image forces that pull it towards certain object contours; and internal forces ("stiffness") that resist excessive deformations
- The trade-off between these forces is adjustable, and adaptable
- External energy reflects how poorly the snake is fitting a contour
- Internal energy reflects how much the snake is bent or stretched
- This sum of energies is minimised by methods like gradient descent, simulated annealing, and partial differential equations (PDEs)
- Problems: numerical instability, and getting stuck in local minima
- With geodesic active contours (used in medical image computing), contours may split and merge, depending on the detection of objects in the image

Demonstration: `https://www.youtube.com/watch?v=ceIddPk78yA`

Scale-Invariant Feature Transform (SIFT)

Goals and uses of SIFT:

- Object recognition with geometric invariance to transformations in perspective, size (distance), position, and pose angle
- Object recognition with photometric invariance to changes in imaging conditions like brightness, exposure, quality, wavelengths
- Matching corresponding parts of different images or objects
- "Stitching" overlapping images into a seamless panorama
- 3D scene understanding (despite clutter)
- Action recognition (what transformation has happened...)

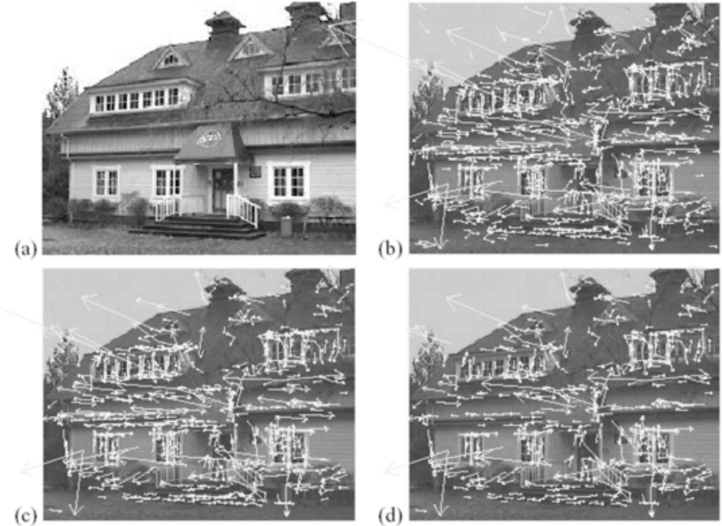


(Scale-Invariant Feature Transform, con't)

Key idea: identifying keypoints that correspond in different images, and discovering transformations that map them to each other.

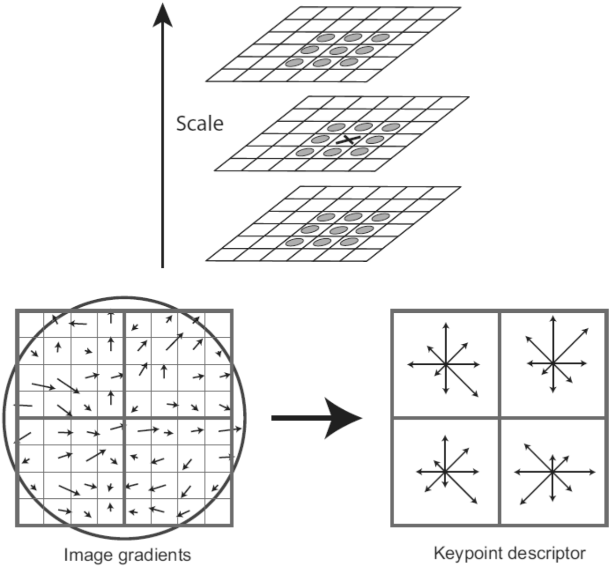
- Various kinds of feature detectors can be used, but they should have an orientation index and a scale index
- Classic approach of Lowe used extrema (maxima and minima) of difference-of-Gaussian functions in scale space
- Build a Gaussian image pyramid in scale space by successively smoothing (at octave blurring scales $\sigma_i = \sigma_0 2^i$) and resampling
- Dominant orientations of features, at various scales, are detected and indexed by oriented edge detectors (*e.g.* gradient direction)
- Low contrast candidate points and edges are discarded
- The most stable keypoints are kept, indexed, and stored for "learning" a library of objects or classes

(Scale-Invariant Feature Transform, con't)



Examples of keypoints (difference-of-Gaussian extrema) detected in an original image, of which 35% are discarded as low contrast or unstable.

(Scale-Invariant Feature Transform, con't)



For each local region (four are highlighted here), an orientation histogram is constructed from the gradient directions as a keypoint descriptor.

(Scale-Invariant Feature Transform, con't)

- The bins of the orientation histogram are normalised relative to the dominant gradient direction in the region of each keypoint, so that rotation-invariance is achieved
- Matching process resembles identification of fingerprints: compare relative configurations of groups of minutiae (ridge terminations, spurs, etc), but search across many relative scales as well
- The best candidate match for each keypoint is determined as its nearest neighbour in a database of extracted keypoints, using the Euclidean distance metric
- Algorithm: best-bin-first; heap-based priority queue for search order
- The probability of a match is computed as the ratio of that nearest neighbour distance, to the second nearest (required ratio > 0.8)
- Searching for keys that agree on a particular model pose is based on Hough Transform voting, to find clusters of features that vote for a consistent pose
- SIFT does not account for any non-rigid deformations
- Matches are sought across a wide range of scales and positions; 30 degree orientation bin sizes; octave (factor of 2) changes in scale

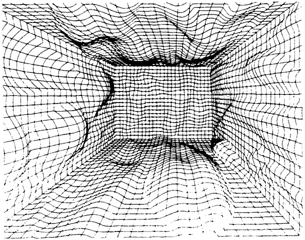
Summary: philosophy and theology of the SIFT

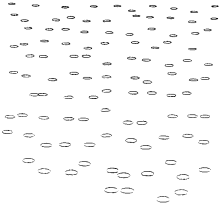
The Doctrine of Suspicious Coincidences

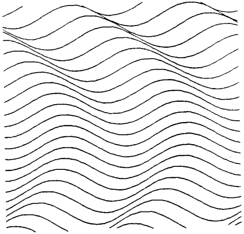
When the recurrence of patterns just by chance is a highly improbable explanation, it is unlikely to be a coincidence.

Structure from Texture

- Most surfaces are covered with texture, of one sort or another
- Texture is both an identifying feature, and a cue to surface shape
- If one can assume uniform statistics along the surface itself, then textural foreshortening or stretching reveals 3D surface shape
- As implied by its root, linking it with (woven) textiles, texture is defined by the existence of statistical correlations across the image
- From grasslands to textiles, the unifying notion is quasi-periodicity
- Variations from uniform periodicity reveal 3D shape, slant, distance



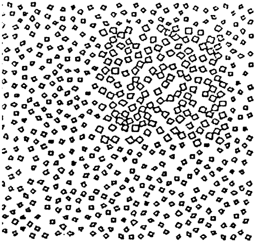
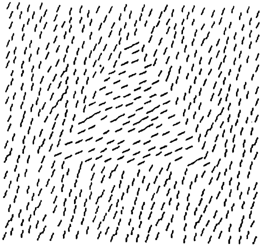
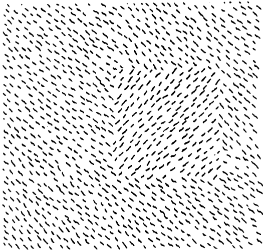




(Structure from Texture, con't)

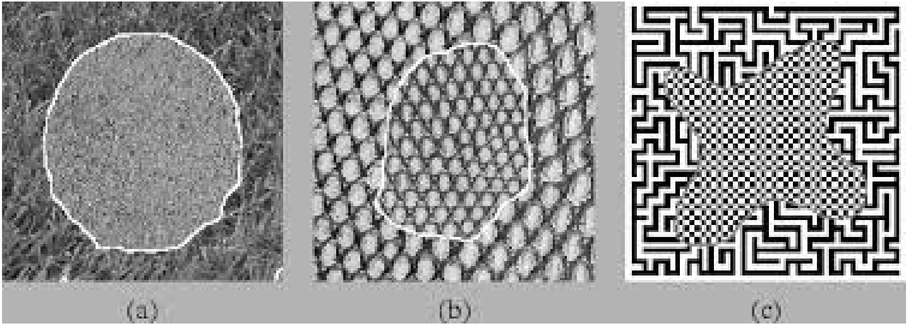
- Quasi-periodicity can be detected best by Fourier-related methods
- The eigenfunctions of Fourier analysis (complex exponentials) are periodic, with a specific scale (frequency) and wavefront orientation
- Therefore they excel at detecting a correlation distance and direction
- They can estimate the "energy" within various quasi-periodicities

- Texture also supports figure/ground segmentation by dipole statistics
- The examples below can be segmented (into figure vs ground) either by their first-order statistics (size of the texture elements), or by their second-order statistics (dipole orientation)

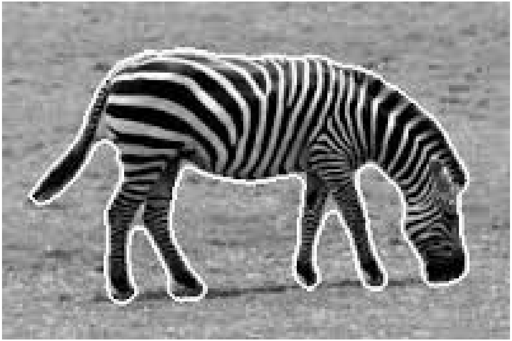
(Structure from Texture, con't)

- Images can be segmented into "figure" vs "ground" regions using Gabor wavelets of varying frequencies and orientations
- The modulus of Gabor wavelet coefficients reveals texture energy variation in those frequencies and orientations across the image
- This can be a strong basis for image segmentation (outlined regions)



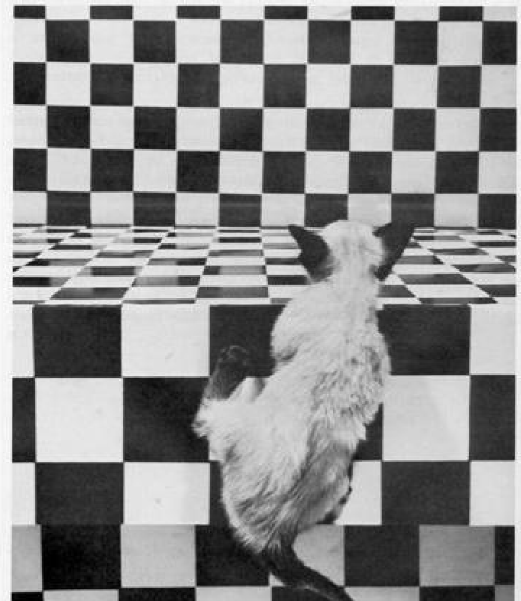
(Structure from Texture, con't)

- Resolving textural spectra simultaneously with location information is limited by the Heisenberg Uncertainty Principle, and this trade-off is optimised by Gabor wavelets
- Texture segmentation using Gabor wavelets can be a basis for extracting the shape of an object to recognise it. (Left image)
- Phase analysis of iris texture using Gabor wavelets is a powerful basis for person identification. (Right image)



(Structure from Texture, con't)

Inferring depth from texture gradients can have real survival value...

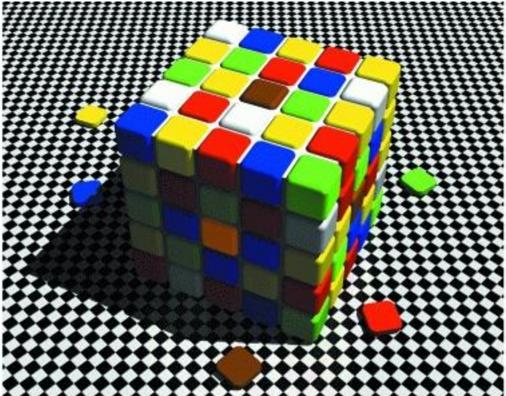


Colour Information

Two compelling paradoxes are apparent in how humans process colour:

1. Perceived colours hardly depend on the wavelengths of illumination (colour constancy), even with dramatic changes in the wavelengths
2. But the perceived colours depend greatly on the local context

The brown tile at the centre of the illuminated upper face of the cube, and the orange tile at the centre of the shadowed front face, are actually returning the same light to the eye (as is the tan tile lying in front)

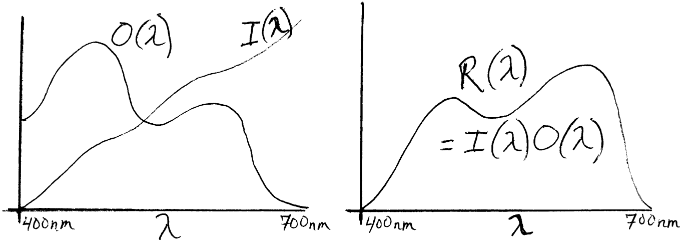


(Colour Information, con't)

Colour is a nearly ubiquitous property of surfaces, and it is useful both for object identification and for segmentation. But inferring colour properties ("spectral reflectances") of object surfaces from images seems impossible, because generally we don't know the spectrum of the illuminant.

- Let $I(\lambda)$ be the wavelength composition of the illuminant
- Let $O(\lambda)$ be the spectral reflectance of the object at some point (the fraction of light scattered back as a function of wavelength λ)
- Let $R(\lambda)$ be the actual wavelength mixture received by the camera at the corresponding point in the image, say for ($400\text{nm} < \lambda < 700\text{nm}$)

Clearly, $R(\lambda) = I(\lambda)O(\lambda)$. The problem is that we wish to infer the "object colour" $O(\lambda)$, but we only know $R(\lambda)$, the mixture received.



(Colour Information, con't)

An algorithm for computing $O(\lambda)$ from $R(\lambda)$ was proposed by Dr E Land (founder of Polaroid Corporation). He named it the *Retinex Algorithm* because he regarded it as based on biological vision (RETINa + cortEX).

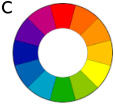
It is a ratiometric algorithm:

1. Obtain the red/green/blue value (r, g, b) of each pixel in the image
2. Find the maximal values $(r_{max},\ g_{max},\ b_{max})$ across all the pixels
3. Assume that the scene contains some objects that reflect "all" the red light, others that reflect "all" the green, and others "all" the blue
4. Assume that those are the origins of the values $(r_{max},\ g_{max},\ b_{max})$, thereby providing an estimate of $I(\lambda)$
5. For each pixel, the measured values (r, g, b) are assumed to arise from actual object spectral reflectance $(r/r_{max},\ g/g_{max},\ b/b_{max})$
6. With this renormalisation, we have discounted the illuminant
7. Alternative variants of the Retinex exist which estimate $O(\lambda)$ using only local comparisons across colour boundaries, assuming only local constancy of the illuminant spectral composition $I(\lambda)$, rather than relying on a global detection of $(r_{max},\ g_{max},\ b_{max})$

(Colour Information, con't)

Colour assignments are very much a matter of calibration, and of making assumptions. Many aspects of colour are "mental fictions".

For example, why does perceptual colour space have a seamless, cyclic topology (the "colour wheel"), with red fading into violet fading into blue, when in wavelength terms that is moving in *opposite* directions along a line ($\lambda \rightarrow$ 700nm red) versus (blue 400nm $\leftarrow \lambda$)?



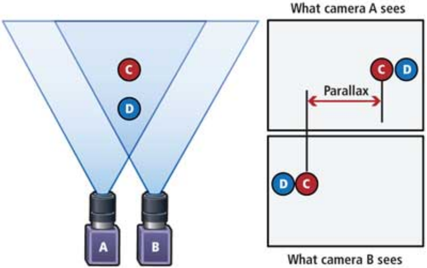
The next slide is a purely monochromatic (black-and-white) picture. But you can cause it to explode into compelling colours by re-calibrating your brain, using the *subsequent* false colour image (2 slides ahead):

1. Stare at the blue disk in the false colour image for about 10 seconds, without moving your eyes. (Finger on key, ready to "flip back")
2. Flip back to the monochromatic image, while continuing to fixate on that same central point
3. As long as you don't move your eyes, you should see very rich and compelling and appropriate colours in the monochromatic image
4. The spell will be broken, your brain's original calibration restored, once you move your eyes

Structure from Stereo Vision

An important source of information about the 3D structure of the surrounding (near) visual world is stereo vision, using stereo algorithms

- Having 2 (or more) cameras, or 2 eyes, with a base of separation, allows the capture of simultaneous images from different positions
- Such images have differences called stereoscopic disparity, which depend on the 3D geometry of the scene, and on camera properties
- 3D depth information can be inferred by detecting those differences, which requires solving the correspondence problem



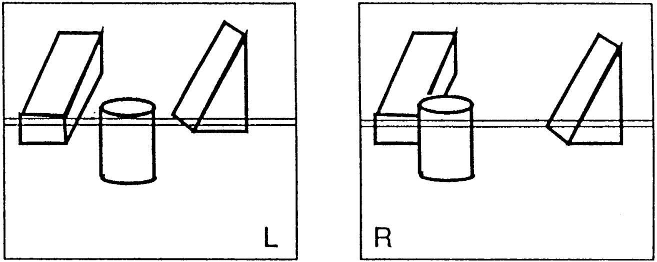
(Structure from Stereo Vision, con't)

Of course, alternative methods exist for estimating depth. For example, the *"Kinect"* gaming device projects an infrared (IR, invisible) laser grid into the scene, whose resulting pitch in the image sensed by an IR camera is a cue to depth and shape, as we saw in discussing shape from texture. Here we consider only depth computation from stereoscopic disparity.

- Solving the correspondence problem can require very large searches for matching features under a large number of possible permutations
- We seek a relative registration which generates maximum correlation between the two scenes acquired with the spatial offset, so that their disparities can then be detected and measured
- The multi-scale image pyramid is helpful here
- It steers the search by a coarse-to-fine strategy to maximise its efficiency, as only few features are needed for a coarse-scale match
- The permutation-matching space of possible corresponding points is greatly attenuated, before refining the matches iteratively, ultimately terminating with single-pixel precision matches

(Structure from Stereo Vision, con't)

- If the optical axes of the 2 cameras converge at a point, then objects in front or behind that point in space will project onto different parts of the two images. This is sometimes called parallax
- The disparity becomes greater in proportion to the distance of the object in front, or behind, the point of fixation
- Clearly it depends also on the convergence angle of the optical axes
- Even if the optical axes parallel each other ("converged at infinity"), there will be disparity in the image projections of nearby objects
- Disparity also becomes greater with increased spacing between the two cameras, as that is the base of triangulation

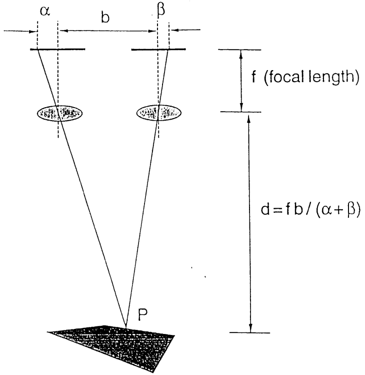


(Structure from Stereo Vision, con't)

In the simplifying case that the optical axes are parallel, once the correspondence problem has been solved, plane geometry enables calculation of how the depth d of any given point depends on:

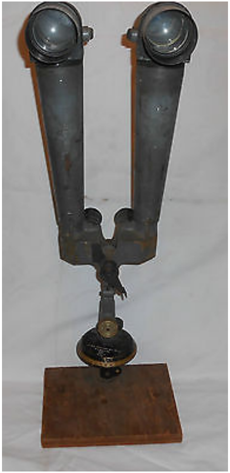
- camera focal length f
- base distance b between the optical centres of their lenses
- disparities (α, β) in the image projections of some object point (P) in opposite directions relative to the optical axes, outwards

Namely: $\boxed{d = fb/(\alpha + \beta)}$



Note: P is "at infinity" if $(\alpha, \beta) = 0$

(Structure from Stereo Vision, con't)



In World War I, stereo trench periscopes were used not only to peer "safely" over the parapets, but by increasing the base of triangulation (increasing the angle of the V), to try to "break camouflage".

Functional streaming: colour and motion pathways