Model Answer – Computer Vision 2004, Paper 9 Question 11 (JGD)

(a) Face recognition is intrinsically difficult because the within-class variability is large (faces are surfaces of 3D objects whose appearances change with pose and viewing angle, illumination geometry, expression, age, and accoutrements), and because the between-class variability is small (different faces generally have the same basic set of features, and in a standard canonical geometry). Whenever the within-class variability is larger than the between-class variability, pattern recognition is intrinsically difficult and error rates are high.

A major shortcoming of most algorithms to date is that they have approached the face recognition problem as a 2D problem instead of as a 3D one. They have been appearance-based (e.g. the "eigenfaces" linear combination of 2D images approach), rather than seeking to build 3D models that would achieve pose-invariance and perspective invariance because 2D appearances could be projected down from such 3D models with any specified illumination, pose, and perspective. An active 3D model might also incorporate expression dynamics and aging. Another factor is that algorithms to date have focused on facial landmark features that are universal for all faces, rather than on features of high randomness and variability amongst different faces. A still further factor is that data fusion over multiple frames has not yet been well-developed; this might allow us to incorporate evidence from dynamic perspectives, including different pose angles but also dynamic behaviour of a face (e.g. the evolution of its expressions during speech or facial gesture). Yet another factor is the fact that even the prior problem of detecting a face as such (versus some other kind of object) is not yet performed very reliably. The over-arching issue is that we do not yet understand well what are the degrees-of-freedom that are generic (universal for the category "face") versus the particular ones that we would like to extract as identifying a particular face. [8 marks]

(b) Using Fourier Transforms allows convolutions to be replaced by mere multiplications. But the cost of performing a 2D FFT on an image having n^2 pixels is on the order of $n^2 \log_2(n)$ multiplications. A further n^2 multiplications are needed in the Fourier domain as the substitute for convolution, once the image FT has been obtained. Thus the total complexity of the Fourier approach is $n^2(2\log_2(n) + 1)$ This becomes more efficient than explicit convolution once the size of the kernel array is larger than about (5 x 5) taps. [2 marks]

(c) The aligned stereo pair of cameras with parameters as specified would compute a target 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.

[5 marks]

(d) The Spectral Co-Planarity Theorem asserts that rigid translational visual motion has the 3D spatio-temporal Fourier domain consequence of collapsing all the spectral energy (normally a 3D distribution) onto just 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. [2 marks]

(e) 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. But the inference of shape and orientation parameters depends on the assumption that the texture is uniform on the surface itself, so that its gradients reveal local surface normals. An example is the use of "wire mesh" to reveal complex shapes in 3D, based either on the assumption of a uniform mesh on the object surface or that the mesh is a uniform cartesian grid system. [3 marks]