ENGINEERING TRIPOS PART IB JUNE 2012

CHAIRMAN PROF. P DAVIDSON

FRIDAY 8TH JUNE 2012   9 TO 11.30

PAPER 8 SOLUTIONS – SELECTED TOPICS

AUTHORS :

PROF. J ROBERTSON

Introductory Business Economics 2012

CRIB

1 (a) Explain the concept of profit maximisation. [4]

Profits are maximised when the next unit produced and sold adds as much to total revenue as it does to total cost. Hence, profit maximisation occurs when marginal revenue = marginal cost (MR=MC). IF MR exceeds MC, then profit can be increased by increasing production. If MC exceeds MR, then profits can be increased by cutting back on production.

(b) Explain why firms may not seek to maximise profits. [6]

Firms may not be able to identify marginal revenue or marginal costs as this requires knowledge of cost and revenue conditions in the market so that MR and MC can be found. Furthermore, firms may not be seeking to maximise profits, instead they may be seeking to maximise revenues - this is particularly the case for manager-controlled businesses as salaries and other perks are often more closely linked with sales. Also firms may be engaged in non-maximising behaviour. A business is an organization with various groups such as employees, managers; shareholders and customers – and each group may have different objectives/goals. In this situation, maximising behaviour may be replaced by 'satisficing' (satisfy and suffice) – setting minimum acceptable outcomes for all groups.

(c) Using the Keynesian Consumption Function Model, explain the potential impact of a temporary increase in income tax on the level of aggregate demand. [5]

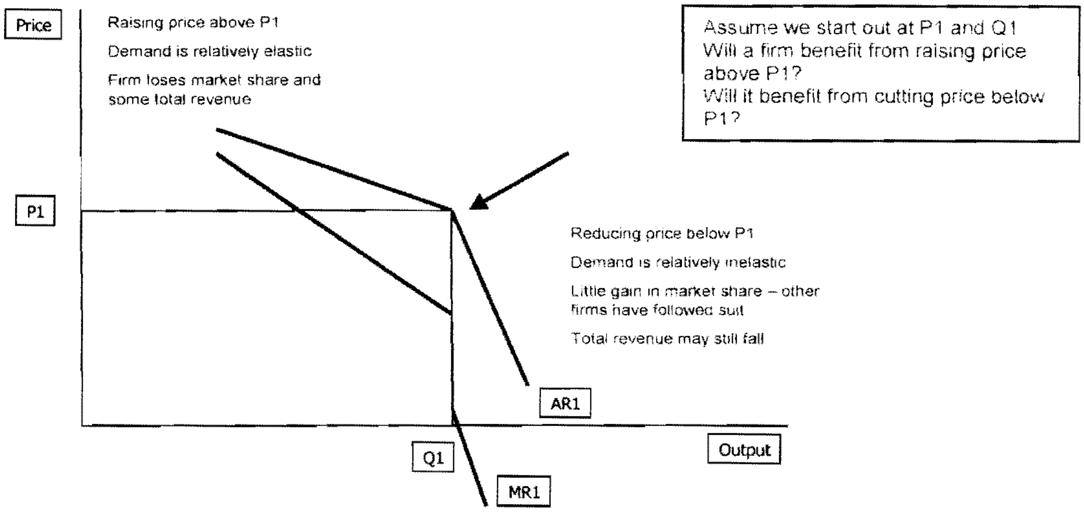
In the Keynesian case, current consumption (C) is determined by real current personal disposable income (Y). A temporary increase in income tax should lead to a fall in current disposable income. This will lead to a fall in consumption and thus a reduction in aggregate demand. This may lead to a fall in output or inflation (or possibly both) depending on the level of economic activity in the economy.

(d) Using either the Life Cycle Model or the Permanent Income Model, explain the potential impact of a temporary increase in income tax on the level of aggregate demand. [5]

The Life Cycle and Permanent Income Models are based on the notion that consumption is determined by long-run or normal income. Thus, temporary income tax changes will have less impact on current consumption than permanent income tax changes. Thus the impact of the temporary tax increase will have little impact on the level of aggregate demand – and hence output or inflation. Candidates may comment on the limitations of the Models – for instance, the assumptions of perfect foresight and no liquidity constraints, and the differences between consumption and consumers expenditure.

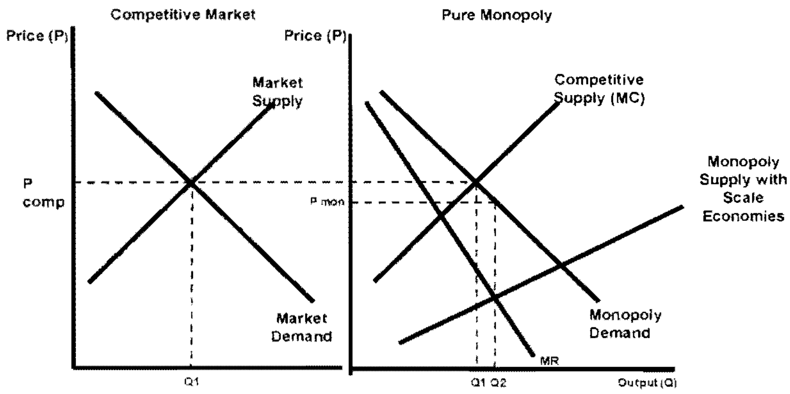
2 (a) Outline the kinked demand curve theory of oligopoly. [6]

With the kinked demand theory, and as shown in the diagram, firms believe that their competitors would follow them if they were to reduce their price below the prevailing market price, and that it is therefore not possible to gain much in the way of extra sales by reducing price (i.e. demand is relatively inelastic below the market price). Also, as shown below, firms believe that their competitors would not follow them if they were to raise their price above the prevailing market price (i.e. demand is relatively elastic above the market price).



(b) Under what circumstances might monopolies improve consumer welfare? [6]

Although normally competitive markets are preferable in terms of consumer welfare a monopoly may be preferable if it can exploit economies of scale and deliver lower prices to consumers (as shown in the diagram below). Natural monopolies can be used as examples (such as the utilities). Some candidates may also comment that monopolies may be more likely to invest in R&D and are stronger in internationally competitive markets.



(c) Outline the accelerator model of investment. [8]

The accelerator model suggests that total capital investment in an economy varies directly with the rate of change of output i.e. investment is largely income-induced. The basic accelerator model assumes: given technological conditions; given relative prices of capital and labour; and a fixed size of capital stock needed to produce a given level of output. If the level of output changes, then the desired size of capital stock will also change. Net capital investment is the amount by which the required capital stock changes so it follows that the amount of investment depends on the size of the change in output. When the rate of growth of demand is strong, the size of the capital stock needs to be increased - boosting demand for capital goods.

Candidates may note the limitations of the accelerator model. First, the capital-output ratio is not fixed – technological change will alter the amount and cost of capital required to produce a given output. Second, relative prices of capital and labour change. Third, changes in demand in the short term can be met by: using up existing spare capacity; and using up stocks of finished goods (inventory changes). Third, firms do not always react immediately to a change in demand.

3. a) Tunnels could be constructed open face, in which the strength of the soil is sufficient to support the tunnel face, or closed-face, in which the soil is supported by a shield using either slurry pressure or compressed air.

At 20m depth to bottom of tunnel, stability number is less than 5, so open face tunnelling is possible

At 12m depth to tunnel base, top of tunnel is at 4m depth in sand, so closed face tunnelling is needed to prevent collapse and water ingress.

b) Segmental lining – tunnel segments are made offsite from either steel or reinforced concrete. These are assembled as the TBM progresses and normally bolted together although they may be held together by earth pressure

NATM. Steel mesh is placed on the exposed soil and concrete is sprayed onto the surface to create the tunnel lining

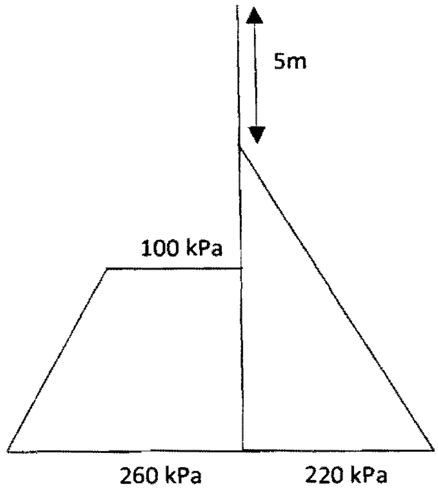
c) Cut and cover tunnelling can use low-tech equipment and can progress over a long length of tunnel simultaneously. It is efficient at small depths, but becomes problematic at depth. It must follow a clear path on the ground surface and so causes great disruption in cites. Removal of spoil can cause congestion. Ground movements may be greater than for bored tunnelling due to shallow depths.

d) Masonry structures are brittle as they cannot sustain tension. Above a tunnel, sagging can in general be sustained by masonry buildings as tension is carried in the foundations, hogging to the sides of the tunnel causes tension at the top of structures which causes more damage.

Damage can be prevented by: Compensation grouting, jacking of structures, tie-rods to strengthen building.

e) Instrumentation such as inclinometers in the soil or strain gauges on buildings can be used to measure movements. These measurements can be used to determine when and where compensation grouting should be used. This can also be used to inform future design or to modify the design using the observational method.

4.



Wall is safe

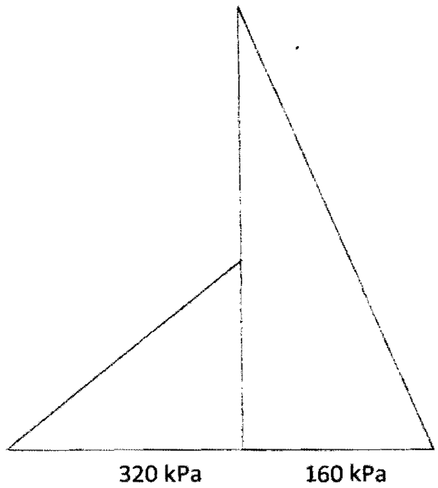
a)

$M_{Passive}$= 100 x 8 x 12 + 80 x 8 x (8+16/3) = 18133 kNm/m

M_{Active}= 220 x 11/2 x (5+22/3) = 14923 kNm/m

FoS = 18133/14923 = 1.2

b)



Ka= 0.5

Kp=2

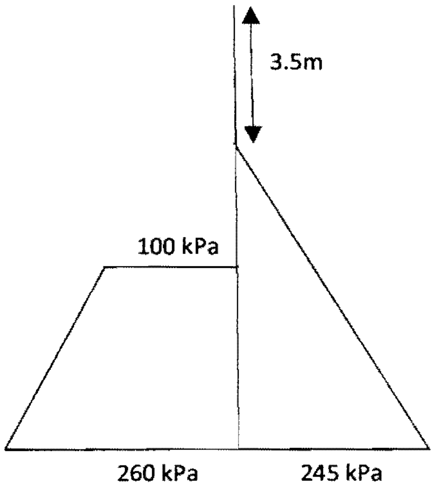
$M_{Passive}$= 320 x 8/2 x (8+16/3) = 17067 kNm/m

M_{Active}= 160 x 8 x (32/3) = 13653 kNm/m

FoS = 17067/13653 = 1.25

Wall is safe

c)



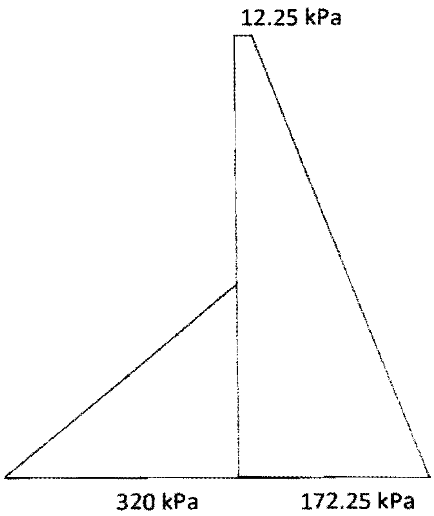
$M_{Passive}$= 100 x 8 x 12 + 80 x 8 x (8+16/3) = 18133 kNm/m (as before)

M_{Active}= 245 x 12.5/2 x (3.5+25/3) = 18119 kNm/m

FoS = 18133/18119 = 1.0

Wall is very marginally safe

d)



$M_{Passive}$= 320 x 8/2 x (8+16/3) = 17067 kNm/m (as before)

M_{Active}= 13653 + 12.25 x 16 x 8 = 15221

FoS = 17067/15221 = 1.12

Wall is marginally safe

5. a) Taking moments about B:

300 x 5/2 x (10/3-1) = x x 3/2 x (4 + 6/3)

x = 194.4 kPa

b) by horizontal equilibrium:

T = 2 x (-194.4 x 3/2 + 300 x 5/2) = 916.75 kN

c) S_A= 0 M_A=0

S_B= 30 kN/m and -428kN/m M_B= 30 x 1/3 =10 kNm/m

S_C= -194.4 x 3/2= -291.7 kN/m M_C= 291.6 x 2 = -583.3 kNm/m

S_A= 0 M_A=0

Maximum moment between B & C at point where S=0:

Let distance above C be y:

291.7 = (300 – 60 y) x y + 60 y x y/2

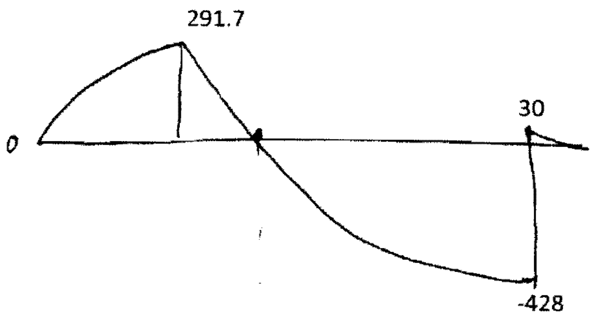
30 y^2 -300y +291.6 =0

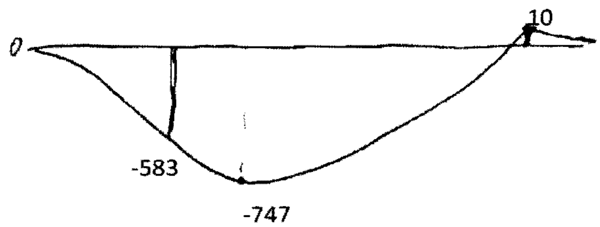
y = 1.09m

M_{max} = (300-1.09x60) x 1.09/2 +(1.09 x 60) x (1.09 / 2) x (2 x 1.09/3) - 291.7 x (1.09+2)

=127.86 + 25.9 – 901.04 = -747.28 kNm/m

M_{min} = 10 kNm/m at B





d) M < 0.15 $f_{cu}bd^2$

$d > (747.28/(0.15*40000))^{0.5} = 353$ mm

e) d= 530 mm

Guess x = 0.4

M = 0.87 fy Ad (1- x/2)

A = M/(0.87 fy d (1- x/2)) = 0.0044 m^2/m = 4404 mm^2/m

x = 2.175 x (fy/fcu)(A/bd) = 0.207

A = 3932 mm^2/m

x = 0.186

A = 3884 mm^2/m (OK) 25 mm bars at 125 mm centres

Also need some hogging reinforcement at top of wall, extra reinforcement at Tie-rods to transmit force into main reinforcement and shear links.

Engineering Tripos Part IB, 2012
Crib for Paper 8, Part C - Mechanics, Materials and Design

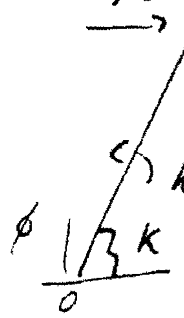
Michael Sutcliffe, Digby Symons, Tim Flack June 2012

6 (a) (i)

Stiffness

$$\delta = \frac{WL^3}{3EI}$$

D'Alembert



Static

$$\delta = \frac{WL^2}{k}$$

Dynamic

$$\text{mo}\circlearrowleft \; -ML^2\ddot{\phi} - k\phi = 0$$

$$\Rightarrow \text{SHM}, \quad f = \frac{1}{2\pi L}\sqrt{\frac{k}{M}}$$

Static:

Match stiffnesses:

$$\frac{\delta}{W} = \frac{L^3}{3EI} = \frac{L^2}{k}$$

$$\Rightarrow \boxed{k = \frac{3EI}{L}}$$

Match frequencies:

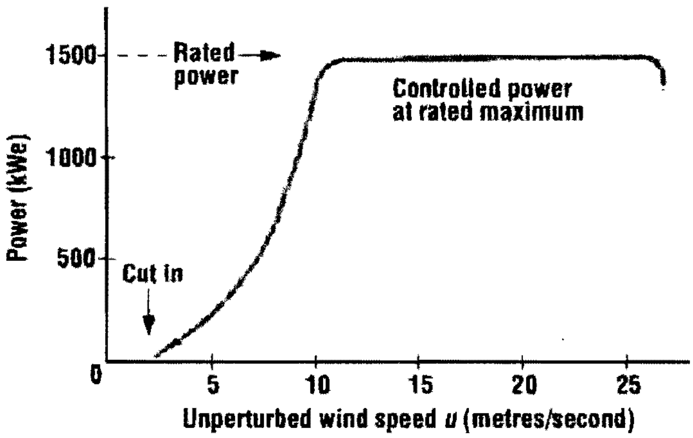
$$\frac{3.52}{2\pi}\sqrt{\frac{EI}{mL^4}} = \frac{1}{2\pi L}\sqrt{\frac{k}{M}}$$

$$\Rightarrow M = \frac{1}{3.52^2}\frac{1}{L^2}\frac{mL^4}{EI}k = 0.24\,Lm$$

substituting for k

(ii) If we want to include the substantial effect of the tower head mass, then we need to get the tower-head deflections right in our model. Hence the model matching the stiffness due to a load at the top of the tower is better.

6b The 'cut-in' speed of a wind turbine is the wind speed above which it becomes worthwhile producing power. 'Rated' wind speed is the wind speed at which the turbine-generator produces its rated output power. 'Stall' wind speed is the wind speed above which it becomes unsafe to continue to operate the wind turbine, and so it is stalled. A typical power vs wind speed characteristic is sketched below.



In the range of wind speeds between rated and stall the turbine must be controlled so that it develops no more than its rated power. This can be achieved by furling the blades (twisting them to reduce their power coefficient), by yawing the turbine to reduce the component of wind normal to the blades or by designing the blades to twist in winds which are greater than rated (passive control).

(d) (i) Turbine needs to produce 5MW when $v = 14\ ms^{-1}$ and with power coefficient 0.45. Using power equation and taking $\rho = 1.23\ kgm^{-3}$:

$5\times10^6 = 0.5\times0.45\times1.23\times A\times14^3$

giving swept area $A = 6584m^2$ and diameter from $A = \pi d^2/4$ as $d = 91.6$ m.

(ii) Tip-speed ratio is 10 and from equation for tip-speed ratio, with $R = d/2$:

$10 = \omega\times45.8/14$ giving $\omega = 3.1\ rads^{-1} = 29$ rpm

(iii) Discount wind speeds below cut-in and above stall since no power is produced at these. Power produced at 16 ms^{-1} wind speed is rated power of 5MW. At 7 ms^{-1} and 12 ms^{-1} wind speeds use fact that power scales with wind speed cubed, and the system produces 5 MW at a 14ms^{-1} wind speed. Thus, power at 7 ms^{-1} wind speed is $(7/14)^3\times5MW = 0.625$ MW and at 12 ms^{-1} wind speed is $(12/14)^3\times5MW = 3.15$ MW. Now complete table:

Wind speed(ms^{-1})	Power(MW)	Days	Hours	Energy (MWhr)
7	0.625	185	4440	2775
12	3.15	100	2400	7560
16	5	50	1200	6000

giving a total of 16335 MWhr = 16.3 GWhr

The capacity factor is $16335/(365\times24\times5) = 0.373$

Q7 (a) Sources of noise

- aerodynamic noise (e.g. turbulence)
- gear noise
- shaft noise associated with out-of-balance
- bearings and generator noise

Characteristics – a spectrum, measured for example with a microphone, represents the contribution of different frequencies to the noise. Some types of noise (e.g. gears) are at specific frequencies but others span a broader spectrum. The dB scale is used for noise.

Managing impact:

- reduce gear noise using helical gears
- vibration isolation
- keep generators away from houses (e.g. 100m separation)
- mapping noise propagation and background noise can help assess impact

Examiners' comments

Q6 Electrical/noise

Part (a) of this question considered simple vibration models of tower sway. The main downfall was failure to draw separate diagrams and perform separate analyses for static and dynamic loading. The second half of this question, sections (b) to (d) addressed electrical/power aspects of wind turbines. Many excellent answers were received, with lots of candidates achieving full marks. The numerical parts of the question caused the most problems, but in most cases the errors were of the careless rather than lack of understanding type.

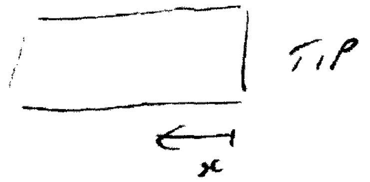
Q7 Noise/materials

(a) Lots of good solutions discussing noise. Clearly candidates had revised this topic well. Marks were lost by not addressing the various aspects listed in the question.
Part (b) was a straightforward materials selection question, reasonably well answered. A sizeable minority didn't write down the appropriate questions, and very few dealt rigorously with the variation in resultant torque along the blade (although still getting the right material grouping). Part (c) asked about composites. The evaluation of stiffness from the carpet plot was well done, but the assessment of what laminates to put on which sections was not well answered. It seemed that the message that the shell can take the torque loading, while the spar takes the bending loads (but with shear loading on the webs) did not get through. The final part describing vacuum infusion was mostly OK.

Q8 Loading and fatigue analysis

Generally well answered with a number of candidates gaining full marks. Most errors seemed to be small mistakes made under the time pressure of an exam, rather than fundamental misunderstandings. However, some candidates failed to identify that N was the wind direction and therefore T the tangential direction producing useful torque. Not all candidates realised that part (b) was a biaxial bending problem and only considered bending about one axis. The fatigue calculation was apparently straightforward for most although a fairly common mistake was to use the self-weight stress amplitude rather than the stress range (twice the amplitude).

7 (b)(i) $T = \dfrac{4 G A_e^2 \phi}{\int \frac{ds}{t}}$ ← $= \frac{s}{t}$



Here $T = Qx$, $\theta = \int \phi \, dx$

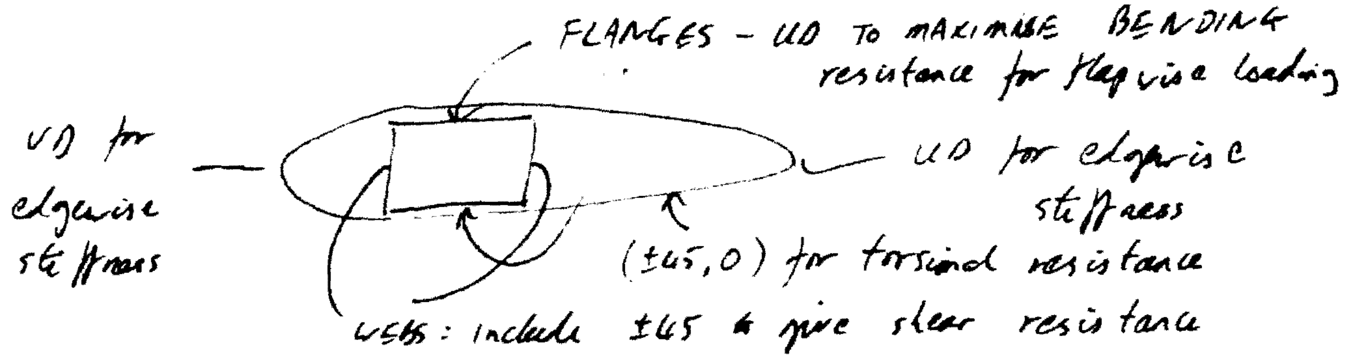
Combining $\theta = \int_0^L x \, dx \, \dfrac{Qs}{t} \dfrac{1}{4 G A_e^2} = \dfrac{L^2 Q s}{8 t G A_e^2}$

Mass $m = \rho t s L$. Eliminate free variable $t \Rightarrow$

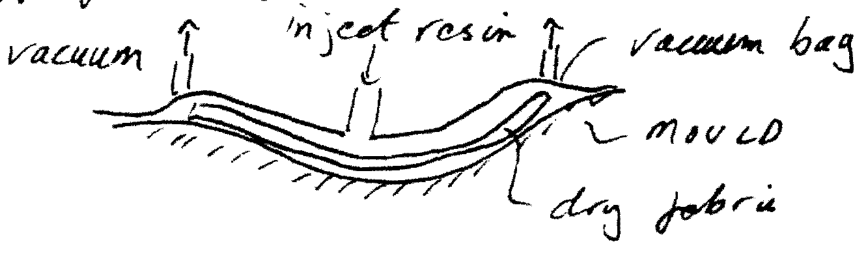
$m = \rho s L . \dfrac{L^2 Q s}{8 \theta G A_e^2}$. To minimise mass, maximise material group G/ρ

(ii) Changes in angle of attack associated with torsion will affect aerodynamic efficiency and may lead to vibration (flutter).

(c) (i) UD - 45 GPa; 50% ±45, 50% 0° - 30 GPa



(d) Widely used as relatively cheap and appropriate for large parts.



Vacuum helps draw resin through dry fabric.

8 (a) Aerodynamic loads: $F_N = 1000\frac{r}{20} = 50r$ N/m

$F_T = 100$ N/m

Moment due to F_T @ hub (r=0):

$$M_T = \int_0^{r=20m} F_T\, r\, dr = 100\left[\frac{r^2}{2}\right]_0^{20} = 20\times10^3 \text{ Nm}$$

Angular velocity $\omega = 30$ RPM $= 30\times 2\pi/60 = \pi$ rads/s

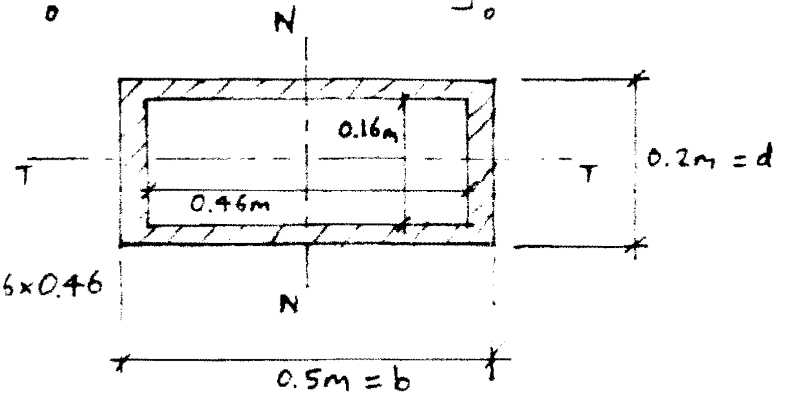
Mechanical power $P = T\omega = 3M_T\omega$ (3 blades)

$= 3\times 20\times10^3 \times \pi = 188$ kW

(b) Moment due to F_N @ hub:

$$M_N = \int_0^{20} F_N\, r\, dr = \int_0^{20} 50r^2 dr = 50\left[\frac{r^3}{3}\right]_0^{20} = 133\times10^3 \text{ Nm}$$

Structural spar:
$t = 0.02$m



Area $A = 0.2\times0.5 - 0.16\times0.46$
$= 0.0264$ m²

2nd moments of area: $I_{NN} = (0.2\times0.5^3 - 0.16\times0.46^3)/12 = 0.786\times10^{-3}$ m⁴

$I_{TT} = (0.5\times0.2^3 - 0.46\times0.16^3)/12 = 0.176\times10^{-3}$ m⁴

Maximum stress due to aerodynamic loads:

$$\sigma_{max,\,aero} = \frac{M_N}{I_{TT}}\frac{d}{2} + \frac{M_T}{I_{NN}}\frac{b}{2} = \left(\frac{133}{0.176}\times0.1 + \frac{20}{0.786}\times0.25\right)\times\frac{10^3}{10^{-3}}$$

$$= (75.6 + 6.36)\times10^6 \text{ N/m}^2 = 82.0 \text{ MPa}$$

(c) Mass / unit length of blade: $m = 300\left(1 - \frac{r}{20}\right)$ kg/m

Centrifugal force in blade:

(i) $$F_c = \int_0^{20} m r \omega^2 \, dr = \int_0^{20} 300\left(1-\frac{r}{20}\right) r \omega^2 \, dr = 300\omega^2 \int_0^{20} r - \frac{r^2}{20} \, dr$$

$$= 300\pi^2 \left[\frac{r^2}{2} - \frac{r^3}{60}\right]_0^{20} = 300\pi^2 \left\{\frac{200}{3}\right\} = 197 \times 10^3 \text{ N}$$

∴ Stress due to centrifugal load: $\sigma_c = \frac{F_c}{A} = \frac{197 \times 10^3}{0.0264} = 7.46$ MPa

(ii) Moment due to self-weight loading when blade is horizontal:

$$M_{sw} = \int_0^{20} m g r \, dr = \int_0^{20} 300\left(1-\frac{r}{20}\right) g r \, dr = 300 g \int_0^{20} r - \frac{r^2}{20} \, dr$$

$$= 300 \times 9.81 \left\{\frac{200}{3}\right\} = 196 \times 10^3 \text{ Nm}$$

∴ Maximum stress due to self-weight is:

$$\sigma_{max, sw} = \frac{M_{sw}}{I_{NN}} \frac{b}{2} = \frac{196}{0.786} \times 0.25 \times 10^6 = 62.3 \text{ MPa}$$

(d) Aero & centrifugal loads are constant, self-weight alternates once per revolution of turbine. Hence:

mean stress $\sigma_m = \sigma_{max, aero} + \sigma_c = 82.0 + 7.46 = 89.5$ MPa

stress range $\Delta\sigma = 2\sigma_{max, sw} = 2 \times 62.3 = 125$ MPa

Equivalent stress range $$\Delta\sigma_o = \frac{\Delta\sigma}{1 - \sigma_m/\sigma_{TS}} = \frac{125}{1 - \frac{89.5}{400}} = 161 \text{ MPa}$$

Fatigue life $$N_f = \left(\frac{C_1}{\Delta\sigma_o}\right)^{1/\alpha} = \left(\frac{800}{161}\right)^{1/0.08} = 505 \times 10^6 \text{ cycles}$$

Hence $505 \times 10^6 / (30 \times 60 \times 24 \times 365) = 32.0$ years

9. a) Lift coefficient is defined as $C_L = L/0.5\rho AV^2$. The maximum lift coefficient at take-off, with a safety margin to avoid stalling is about 1.6. Take-off speed is limited by runway length to around 90 m/s. Air density is set by atmospheric conditions (and is lower at hot and high airports). Thus, the wing area is given by:

$$\Rightarrow A = \frac{W_{TO}}{0.5\rho_a V_{TO}^2 C_{L,\max}}$$

[3]

b) There is an optimum lift coefficient for minimum fuel burn (or maximum range) at cruise. This is typically around 0.5. From rearranging the lift coefficient expression, the cruise speed is given by:

$$V_{cruise} = \sqrt{\frac{W_{cruise}}{0.5\rho_a C_L}}$$

Thus, to make the cruise speed high (and give high efficiency transport), the density should be low, which is achieved by flying at high altitude.
In other words, reducing the density allows both M and C_L to be set at their optimum values (~ 0.5 and 0.85 respectively), such that $ML/D = f(M, C_L)$ is a maximum.
During a long range flight, the weight of the aircraft W reduces as fuel is burnt. Thus, to maintain the same, optimum lift coefficient, $C_L = W/0.5\rho AV^2$ the altitude should be increased to reduce the air density in proportion with the reduction in weight.

[4]

c) The sfc, assumed constant throughout cruise, is given by:

$$\text{sfc} = \frac{\textit{mass flow rate of fuel}}{\textit{engine net thrust}} = \frac{\dot{m}_f}{F_N}$$

$$\frac{dw}{dt} = -g\,\dot{m}_f = -g\,sfc\,F_N = -g\,sfc \times drag = -g\,sfc \times \frac{lift}{L/D} = -\frac{g\,sfc}{L/D} \times w$$

Assuming cruise is at constant velocity, level flight,

$$\frac{dw}{w} = -\frac{g\,sfc}{L/D} \times dt = -\frac{g\,sfc}{L/D} \times \frac{ds}{V}$$

Integrating, assuming the L/D is also constant throughout cruise,

$$[\ln w]_{start}^{end} = -\frac{g\,sfc}{V\,L/D} \times [s]$$

Rearranging, $\underline{s = \frac{V\,L/D}{g\,sfc} \ln\left(\frac{W_{start}}{W_{end}}\right)}$

[5]

d) Thermal efficiency: $\eta_{\text{th}} = \frac{\Delta KE}{\textit{fuel heat release}} = \frac{\Delta KE}{\dot{m}_f\,LCV}$

Propulsive efficiency: $\eta_{\text{p}} = \frac{\textit{power to aircraft}}{\textit{power to jet}} = \frac{VF_N}{\Delta KE}$

Specific fuel consumption: $sfc = \frac{\dot{m}_f}{F_N} = \frac{\dot{m}_f LCV}{VF_N} \times \frac{V}{LCV} = \frac{1}{\eta_{th}\eta_p} \times \frac{V}{LCV}$

Thus, maximising the product $\eta_{\text{th}}\eta_{\text{p}}$ minimises sfc.

[4]

e) High propulsive efficiency is achieved by having a turbofan engine with a high bypass ratio (or a large diameter fan / low specific thrust). This gives an engine which produces thrust using a large total mass flow rate but low jet velocity, leading to high Froude efficiency. An upper limit of propulsive efficiency exists since increasing bypass ratio increases fan diameter and engine weight. This also increases the engine and installation drag contribution. A larger engine may also require a larger undercarriage, pylon, etc.

High thermal efficiency is achieved through improving engine core thermodynamics: high turbine entry temperature (TET) with cooled turbines, optimum, high overall pressure ratio (~40), high component isentropic efficiencies. The upper limit to thermal efficiency exists since turbine blades and cooling technology limits TET, isentropic efficiencies are limited by turbomachinery technology and inherent irreversibilities.

[4]

10. a)

$$\eta_{is,c} = 0.88 = \frac{\left(\frac{p_{03}}{p_{023}}\right)^{\frac{\gamma-1}{\gamma}} - 1}{\frac{T_{03}}{T_{023}} - 1} \Rightarrow \frac{T_{03}}{T_{02}} = \frac{(25)^{\frac{\gamma-1}{\gamma}} - 1}{0.88} + 1 = 2.714 \Rightarrow T_{03} = \underline{795.3 \text{ K}}$$

$$n_{stage} > \frac{c_p(T_{03} - T_{023})}{\left[\Delta h_0/U^2\right]_{\min} \times U^2} = \frac{1010 \times (795.3 - 293)}{0.45 \times 355^2} = 9.2$$

Therefore the minimum number of stages required is <u>10</u>

[5]

b) Work balance for HP shaft, $\dot{m}c_p(T_{03} - T_{023}) = \dot{m}c_p(T_{04} - T_{045})$

$$\Rightarrow T_{045} = T_{04} - (T_{03} - T_{023}) = 1650 - (795.3 - 293) = \underline{1147.7 \text{ K}}$$

$$\eta_{is,t} = 0.85 = \frac{1 - \frac{T_{045}}{T_{04}}}{1 - \left(\frac{p_{045}}{p_{04}}\right)^{\frac{\gamma-1}{\gamma}}} \Rightarrow \frac{p_{045}}{p_{04}} = \left(1 - \frac{1 - \frac{T_{045}}{T_{04}}}{0.85}\right)^{\frac{\gamma}{\gamma-1}} = \left(1 - \frac{1 - \frac{1147.7}{1650}}{0.85}\right)^{\frac{1.4}{.4}} = 0.212$$

$$\Rightarrow \quad p_{045} = 70 \times 25 \times 0.212 = \underline{370.7 \text{ kPa}}$$

[4]

c) For the LP Turbine $\eta_{is,t} = 0.9 = \dfrac{1 - \frac{T_{05}}{T_{045}}}{1 - \left(\frac{p_{05}}{p_{045}}\right)^{\frac{\gamma-1}{\gamma}}}$

$$\Rightarrow \frac{T_{05}}{T_{045}} = 1 - 0.9 \times \left(1 - \left(\frac{p_{05}}{p_{045}}\right)^{\frac{\gamma-1}{\gamma}}\right) = 1 - 0.9 \times \left(1 - \left(\frac{1}{8.5}\right)^{\frac{.4}{1.4}}\right) = 0.588 \Rightarrow T_{05} = 675.2 \text{ K}$$

For the exhaust jet, given the nozzle is isentropic:

$$0.5V_9^2 = c_p(T_{05} - T_9) = c_p T_{05}\left(1 - (p_9/p_{05})^{(\gamma-1)/\gamma}\right)$$

$$V_9 = \sqrt{2c_p T_{05}\left(1 - (p_9/p_{05})^{(\gamma-1)/\gamma}\right)} = \sqrt{2 \times 1010 \times 675.2 \times \left(1 - (28.7/(370.7/8.5))^{.4/1.4}\right)}$$

$= \underline{392 \text{ m/s}}$

[6]

d) Work balance for the LP shaft:

$$\dot{m}c_p(T_{023} - T_{02}) + BPR.\dot{m}c_p(T_{013} - T_{02}) = \dot{m}c_p(T_{045} - T_{05})$$

$$BPR(T_{013} - T_{02}) = (T_{045} - T_{05}) - (T_{023} - T_{02})$$

If the outer fan is isentropic, $T_{013} - T_{02} = \left(V_j^2 - V^2\right)/2c_p$

$$\Rightarrow BPR = \frac{(T_{045} - T_{05}) - (T_{023} - T_{02})}{\left(V_j^2 - V^2\right)/2c_p} = \frac{1147.7 - 675.2 - 41}{\left(392^2 - 240^2\right)/2020} = \underline{9.1}$$

(A reasonable value for a modern turbofan. It would be lower if the fan was not isentropic)

[5]

11 a)

$V = M\sqrt{\gamma R T_a} = 0.8\times\sqrt{1.4\times 287\times 223} = \underline{239.5\text{ m/s}}$

$F_G = \dot{m}V_j = 400\times 350 = \underline{140\text{ kN}}$

$F_N = \dot{m}(V_j - V) = 400\times(350 - 239.5) = \underline{44.2\text{ kN}}$

$$T_{02} = T_a\left(1+\frac{\gamma-1}{2}M^2\right) = 223\left(1+0.2\times 0.8^2\right) = \underline{251.5\text{ K}}$$

$$p_{02} = p_a\left(1+\frac{\gamma-1}{2}M^2\right)^{\frac{\gamma}{\gamma-1}} = 26\left(1+0.2\times 0.8^2\right)^{3.5} = \underline{39.6\text{ kPa}}$$

[5]

b)

Applying conservation of momentum between the exit of the propelling nozzle (19) and far downstream, in the jet, we see that: $\dot{m}_a(V_j - V_{19}) = A_N(p_{19} - p_a)$

Hence, $\dot{m}_a V_{19} + A_N p_{19} = \dot{m}_a V_j + A_N p_a = F_G + p_a A_N$

Therefore, the expression $(F_G + p_a A_N)$ only depends on conditions <u>inside</u> the engine $(\dot{m}_a, V_{19}, p_{19})$ provided the nozzle is choked (thus preventing any influence from downstream of the engine, within the atmosphere)

$$\tilde{F} = \frac{140 + 26\times 2.8}{39.6\times 2.8} = \underline{1.919} \qquad \tilde{\dot{m}} = \frac{400\times\sqrt{1010\times 251.5}}{39.6\times 10^3\times 2.8} = \underline{1.818}$$

[5]

c) $F_N = F_G - \dot{m}V$

$$\frac{F_N}{p_a A_N} = \frac{F_G}{p_a A_N} - \frac{\dot{m}V}{p_a A_N} = \left[\tilde{F}\left(\frac{p_{02}}{p_a}\right) - 1\right] - \frac{\dot{m}\sqrt{c_p T_{02}}}{p_{02} A_N}\frac{p_{02}}{p_a}M\sqrt{\frac{\gamma R}{c_p}\frac{T_a}{T_{02}}}$$

Using $\gamma R / c_p = \gamma - 1$,

$$\underline{\frac{F_N}{p_a A_N} = \left[\tilde{F}\left(\frac{p_{02}}{p_a}\right) - 1\right] - \tilde{\dot{m}}M\sqrt{\gamma - 1}\frac{p_{02}}{p_a}\left(\frac{T_{02}}{T_a}\right)^{-\frac{1}{2}}}$$

[4]

d) $T_{02,new} = T_a\left(1+\frac{\gamma-1}{2}M^2\right) = 216\left(1+0.2\times 0.85^2\right) = 247.2\text{ K}$

$$p_{02,new} = p_a\left(1+\frac{\gamma-1}{2}M^2\right)^{\frac{\gamma}{\gamma-1}} = 18\left(1+0.2\times 0.85^2\right)^{3.5} = 28.9\text{ kPa}$$

$$\dot{m}_{new} = \tilde{\dot{m}}\frac{p_{02,new}A_N}{\sqrt{c_p T_{02,new}}} = 1.818\frac{28.9\times 10^3\times 2.8}{\sqrt{1010\times 247.2}} = \underline{294.4\text{ kg/s}}$$

$$\frac{F_N}{p_a A_N} = 1.919(1.1445)^{3.5} - 1 - 1.818\times 0.85\sqrt{0.4}(1.1445)^{3.5-0.5} = 0.6126$$

$$\Rightarrow F_N = 0.6126\times 18\times 2.8 = \underline{30.9\text{ kN}}$$

(Note that the available thrust and engine mass flow are reduced, as expected for a higher altitude and Mach number flight condition)

[6]

Dr C. A. Hall

Section 8E

12 (a) Al good conductivity, good contact to Si, low melting point but suffers from (1) corrosion effects – when water enters a chip package an electrochemical corrosion of Al can occur in the presence of either chlorine or phosphorus. Both of these are often used in semiconductor processing. (2) Electromigration – at high current densities metal atoms can actually be swept along by the current carriers leading to breakages...........

Al/Cu: Aluminium doped with a small amount of copper can be used to reduce electromigration because its smaller grain size gives better resistance to electromigration. It does however have a higher resistivity than Al.

Al/Si: Aluminium doped with Si is used to control the alloying of aluminium which is in direct control with Si. This reduces the chance of aluminium spiking which can damage shallow junctions – see fig. 1

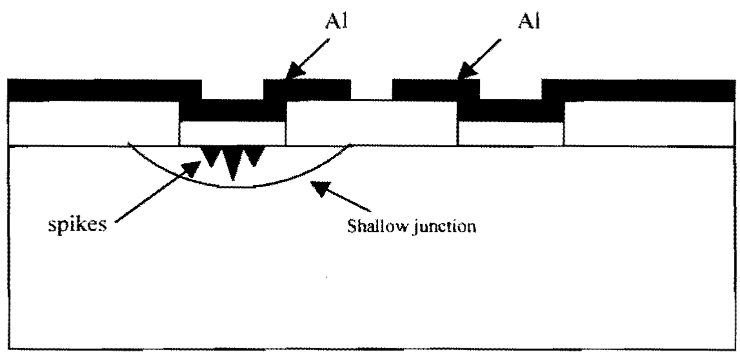


Fig. 1

(b) ignore parasitic effects

$$RC = \left(\frac{\rho \ell}{A}\right)\left(\frac{\varepsilon_o \varepsilon_r A}{\text{spacing}}\right)$$

$$= \left(\frac{2.7\times10^{-6}\times2\times10^{-1}}{0.35\times0.35\times10^{-8}}\right)\left(\frac{8.85\times10^{-14}\times3.9\times2\times10^{-1}\times0.35\times10^{-4}}{200\times10^{-7}}\right)$$

$$= 5.32\times10^{-11} s$$

(c) Self Aligned Structures
It is essential for the correct functioning of the NMOST that the channel region extends all the way between the source and drain. Because of alignment tolerances between masking levels it is possible that one of the edges of the gate region, B or B' may not overlap the source or drain edges A and A' unless a sufficient margin of error is allowed. Thus in the mask design the gate overlap is made larger than is ideally necessary.

Large overlap means an increase in gate-source and more importantly gate-drain parasitic capacitances.

Reduces potential speed of device.

This led to the design of the polysilicon self-aligned gate structure.

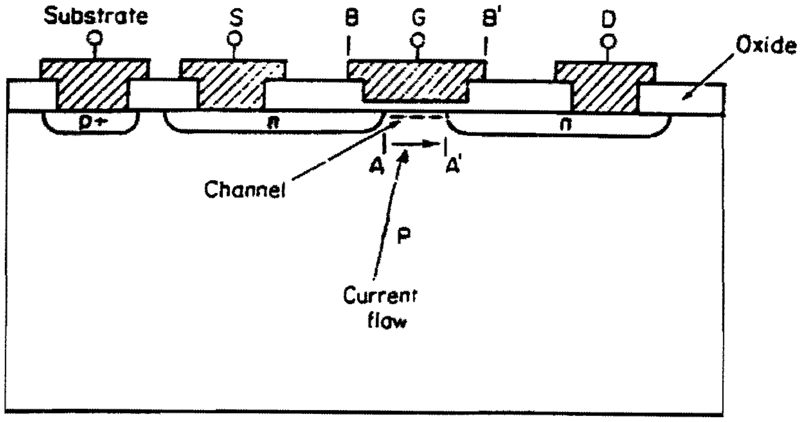
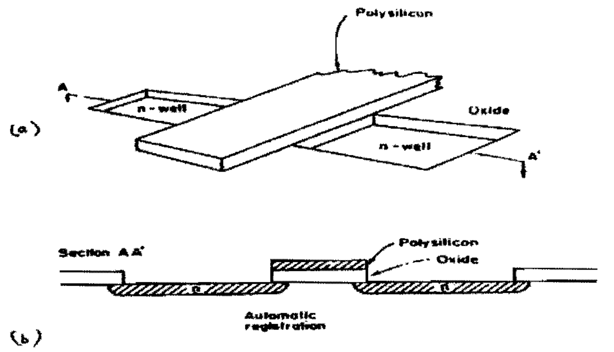


Fig. 2



(a) source and drain window opened; (b) source and drain diffusion

Fig. 3

Polysilicon is used because it can withstand much higher temperatures than conventional contact metals and this is important as it is present during the source and drain diffusion or post implantation anneal.

The sideways movement of dopants at gate edges will always ensure that the gate-source and gate-drain overlap sufficiently **but** not too much.

Need some overlap as any unaccumulated gate length will lead to parasitic resistance which will also affect speed.

The other clever aspect of the process is that the gate polysilicon is also simultaneously doped with the source and drain and its resistance reduced.

13 a) semiconductors and insulators have band gaps. Insulators have bigger band gaps, so they can't be doped. Metals have band overlap so that they conduct even at 0^0 K.

b) intrinsic semiconductors conduct by thermal excitation of their undoped electrons and holes into the relevant bands,
In doped semiconductors, the dopants atoms provide the free carriers, one carrier (electron or hole) per dopant atom, so that carrier density = number of dopant atoms. Conductivity is controllable unlike in a metal, over many orders of magnitude, limited by the purity at the lowest level and by the dopant solubility at the top.

c) Electrons are accelerated by applied field. They are then randomly scattered by phonons or defects after an average time τ. This slows them down to zero velocity, after which they go through the cycle again.

$F = eE = m^* v/\tau$, which gives
$v = eE\tau/m^*$

m^* is the effective mass, a quantity that plays role of electron mass in this equation, v is the average electron velocity.

Using the analogy to incompressible water passing along a pipe,
the total charge flow = number of electrons time the electronic charge.
Number of electrons = density of electrons times volume that passes,
or A x L, L = v.t, so

Conductivity $\sigma = n.e.\mu$

Taking mobility from above, $\mu = v/E = e\tau/m^*$ (eqn in data sheet)

$\sigma = n.e.\mu = \mathbf{n.e^2\tau/m^*}$

The carrier velocity has a maximum value of around the scattering limited velocity, v_s.

d) $v = \mu E$, critical field $E = v/\mu = 1.25\text{x}10^5/0.5 = 2.5\text{x}10^5$ V/m

channel length $L = V/E = 0.5/ 2.5\text{x}10^5 = \mathbf{2\ x10^{-6}\ m}$.
transit time $t = L/v = 2\ \text{x}10^{-6} / 1.25\ \text{x}\ 10^5 = \mathbf{1.6\ x\ 10^{-11}\ s}$

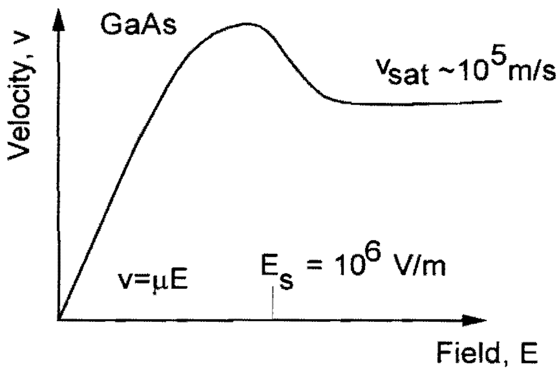
e) $\mu = e\tau/m^*$ so $\tau = \mu m^*/e = 0.5 \text{ x } 0.2 \text{ x } 9.1\text{x}10^{-31} / 1.6\text{x}10^{-19} = \mathbf{5.6x10^{-13}\ s}$

this is an order of magnitude less than transit time, so the approximations are OK.

14 a,b,c) scaling – book work.
a) Simple scaling = constant field scaling.
Device density = $1/k^2$, transit time $1/k$, voltage k, power per chip 1, wafer size undefined.

b) lithography, originally
now power dissipation due to lack of voltage scaling.
also limited by introduction of new materials to increase material parameters.

c)



d) field $E = V/L = 0.5/5.10^{-8} = 10^7$ V/m. this is above critical field
critical field = $2.10^5/4 = 0.5\times10^5$ V/m.

hence carrier velocity $v = 2\times10^5$ m/s
$t = 5\times10^{-8}/2\times10^5$ = **2.5×10^{-13} s**

e) $V_g = eNd^2/(2\varepsilon)$, so $N = 2\varepsilon V/ed^2$ = **4.68×10^{23} m^{-3}**

Q15.

Professor Roberto Cipolla
Department of Engineering
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ
Tel: 01223 332849
Email: cipolla@eng.cam.ac.uk

(a) 1-D linear interpolation:

The pixel at p should change linearly from $x(a)$ to $x(b)$ as p moves from a to b. Hence

$$x(p) = x(a) + \frac{p-a}{b-a} \cdot \left(x(b) - x(a) \right)$$

because $\frac{p-a}{b-a}$ changes linearly from 0 to 1 as p moves from a to b. Simplifying the above gives

$$x(p) = \frac{1}{b-a}\left[(b-p)\, x(a) + (p-a)\, x(b) \right]$$

5

(b) If we have 4 pixels located as follows:

$\cdot\, x(a,c)$ $\qquad\qquad$ $\cdot\, x(a,d)$

$\qquad\qquad \cdot\, x(p,q)$

$\cdot\, x(b,c)$ $\qquad\qquad$ $\cdot\, x(b,d)$

& we wish to find the bi-linearly interpolated pixel $x(p,q)$ as shown, then first we must find pixels located at (p,c) & (p,d) using vertical linear interp, and then we must find $x(p,q)$ by horizontal linear interp between $x(p,c)$ & $x(p,d)$

5

(b) (cont).

Hence
$$x(p,c) = \frac{1}{b-a}\Big[(b-p)\,x(a,c) + (p-a)\,x(b,c)\Big]$$

&
$$x(p,d) = \frac{1}{b-a}\Big[(b-p)\,x(a,d) + (p-a)\,x(b,d)\Big]$$

And so

$$x(p,q) = \frac{1}{d-c}\Big[(d-q)\,x(p,c) + (q-c)\,x(p,d)\Big]$$

$$= \frac{(d-q)\big[(b-p)\,x(a,c) + (p-a)\,x(b,c)\big] + (q-c)\big[(b-p)\,x(a,d) + (p-a)\,x(b,d)\big]}{(b-a)(d-c)}$$

(c) The image processing function which adjusts locations of facial features is known as '<u>morphing</u>'. It is necessary to generate a smooth field of displacement vectors, centred on the region that is to be displaced/distorted, such that pixels near the centre of the region are displaced by the desired amount and in the desired direction, while other nearby pixels are displaced by a progressively smaller extent as one moves away from the region centre.

A <u>Gaussian function in 2-D</u> is circularly symmetric and has the desired property of smooth

(c) (cont)

decay as one moves away from the region centre.

A Gaussian function of standard deviation σ, centred on (p, q) in an image is given by
control points

$$g(s,t) = \exp\left(-\frac{(s-p)^2 + (t-q)^2}{2\sigma^2}\right)$$

This has a maximum value of 1 when $s = p$ and $t = q$ at the desired centre.

If the pixel at (p, q) needs to come from (u, v) in the original image, then the displacement vector at (p, q) should be

$$\hat{\underline{d}}(p, q) = (u - p, v - q)$$

Hence the field at an arbitrary location (s, t) should be given by $g(s,t) . \hat{\underline{d}}(p,q)$, so that

$$\underline{d}(s,t) = \exp\left(-\frac{(s-p)^2 + (t-q)^2}{2\sigma^2}\right) . (u-p, v-q)$$

5

(d) If the displacement field at (s,t) is $\underline{d}(s,t)$, then this means that the pixel at (s,t) in the output image should come from location $(s,t) + \underline{d}(s,t)$ in the original image. To avoid destroying the sense of the image, the ordering of the pixels in the output image must not change from that in the input image. Hence the gradient of the above expression w.r.t. s & t must remain positive.

Hence
$$\frac{\partial}{\partial s}\left[(s,t) + \underline{d}(s,t)\right] > 0$$
$$+ \quad \frac{\partial}{\partial t}\left[(s,t) + \underline{d}(s,t)\right] > 0$$

Considering just the $\frac{\partial}{\partial s}$ term first

$$\frac{\partial}{\partial s}\left[(s,t) + \underline{d}(s,t)\right] = 1 +$$

(d) (cont.)

Since the gradient of the first term (s,t) is unity, it is sufficient to require that the gradient of $|\underline{d}(s,t)|$ is greater than -1 everywhere. Considering a simplified 1-D Gaussian

$$g(r) = \exp\left(\frac{-r^2}{2\sigma^2}\right)$$

$$\frac{dg}{dr} = -\frac{2r}{2\sigma^2} \cdot \exp\left(\frac{-r^2}{2\sigma^2}\right) = -\frac{r}{\sigma^2}\exp\left(\frac{-r^2}{2\sigma^2}\right)$$

To find the steepest (max) gradient, we must differentiate again to get:

$$\frac{d^2g}{dr^2} = -\frac{1}{\sigma^2}\exp\left(\frac{-r^2}{2\sigma^2}\right) + \frac{r}{\sigma^2} \cdot \frac{2r}{2\sigma^2}\exp\left(\frac{-r^2}{2\sigma^2}\right)$$

$$= -\frac{1}{\sigma^2}\left(1 - \frac{r^2}{\sigma^2}\right)\exp\left(\frac{-r^2}{2\sigma^2}\right)$$

$\therefore \frac{d^2g}{dr^2} = 0$ when $r = \pm\sigma$, so the max + min values of $\frac{dg}{dr}$ are

$$\mp \frac{1}{\sigma}\exp\left(\frac{-1}{2}\right) = \mp \frac{0.6065}{\sigma}$$

Hence, to ensure that $|\underline{d}(s,t)|$ is greater than -1 everywhere, we require

$$\left(\frac{0.6065}{\sigma}\right) \cdot |(u-p, v-q)| < 1$$

(5)

$$\therefore \sigma > 0.6065 \cdot |(u-p, v-q)|$$

In practice we allow a margin of safety so $\sigma > 0.8\,|(u-p, v-q)|$

Q16 Feature detection

(a) (i) - Smoothing — reduce high frequency noise before differentiation (latter amplifies noise)

- Use a low-pass filter; Gaussian Kernel

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}}$$

$$= g_\sigma(x)\, g_\sigma(y)$$

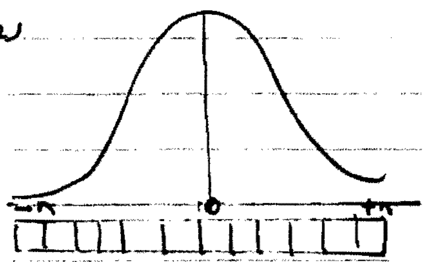
— Increasing σ increases amount of blurr, i.e. suppresses high frequencies + cut-off frequency is reduced and only ~~very~~ larger spatial structures survive.

(a)(ii).

$$S(x,y) = \sum_{-n}^{n}\sum_{-n}^{n} I(x-u, y-v)\, G_\sigma(u,v)$$

$$= \sum_{-n}^{n} g_\sigma(u) \sum_{-n}^{n} g_\sigma(v)\, I(x-u, x-v)$$

i.e. implement as 2 1D discrete convolutions where $g_\sigma(x)$ is 1D gaussian samples



2n + 1 samples of Kernel

16(a)(iii)

For edge detection we require $\nabla S = \nabla (G_\sigma * I)$

We compute derivatives by finite differences:

By Taylor Series expansion (In 1D)

$$S(x+\Delta x) = S(x) + \frac{dS}{dx}\Delta x + \frac{d^2S}{dx^2}\frac{\Delta x^2}{2} \ldots$$

$$S(x-\Delta x) = S(x) - \frac{dS}{dx}\Delta x + \frac{d^2S}{dx^2}\frac{\Delta x^2}{2} \ldots$$

$$\therefore \frac{dS}{dx} \simeq \frac{S(x+\Delta x) - S(x-\Delta x)}{2} + O(\Delta x^2) \quad (1)$$

For $S(x,y)$

$$\frac{\partial S}{\partial x}(x,y) \simeq \frac{S(x+1,y) - S(x-1,y)}{2} \quad (2)$$

and

$$\frac{\partial S}{\partial y}(x,y) \simeq \frac{S(x,y+1) - S(x,y-1)}{2} \quad (3)$$

Equivalent to convolution with masks (remember flipping operation in convolution)

$+\frac{1}{2}$	0	$-\frac{1}{2}$

and

$-\frac{1}{2}$
0
$+\frac{1}{2}$

(4)

$$\frac{\partial S(x,y)}{\partial x} = -\frac{1}{2}S(x-1,y) + \frac{S(x+1,y)}{2}$$

$$\frac{\partial S}{\partial y}(x,y) = -\frac{1}{2}S(x,y-1) + \frac{S(x,y+1)}{2}$$

P 3

Q16(b) Image pyramids and blob-detection (scale-space)

(i). Need to generate a discrete set of images with difference amount of blur. We sample $S(x, y, \sigma^2)$, logarithmically spaced;

$$\sigma_i = 2^{\frac{i}{s}} \sigma_0 \quad , \quad \sigma_{i+1} = 2^{\frac{1}{s}} \sigma_i \qquad (1)$$

with s images per octave (i.e. after s images, σ has doubled).

— apply incremental blur (gaussian σ_K) between images in octave to get images with increasing amount of blur.

$$G(\sigma_{i+1}) = G(\sigma_i) * G(\sigma_K) \qquad \sigma_{i+1}^2 = \sigma_i^2 + \sigma_K^2 \text{ and } \sigma_{i+1} = 2^{\frac{1}{s}}\sigma_i$$

$$\sigma_K = \sigma_i \sqrt{2^{\frac{2}{s}} - 1} \qquad (2)$$

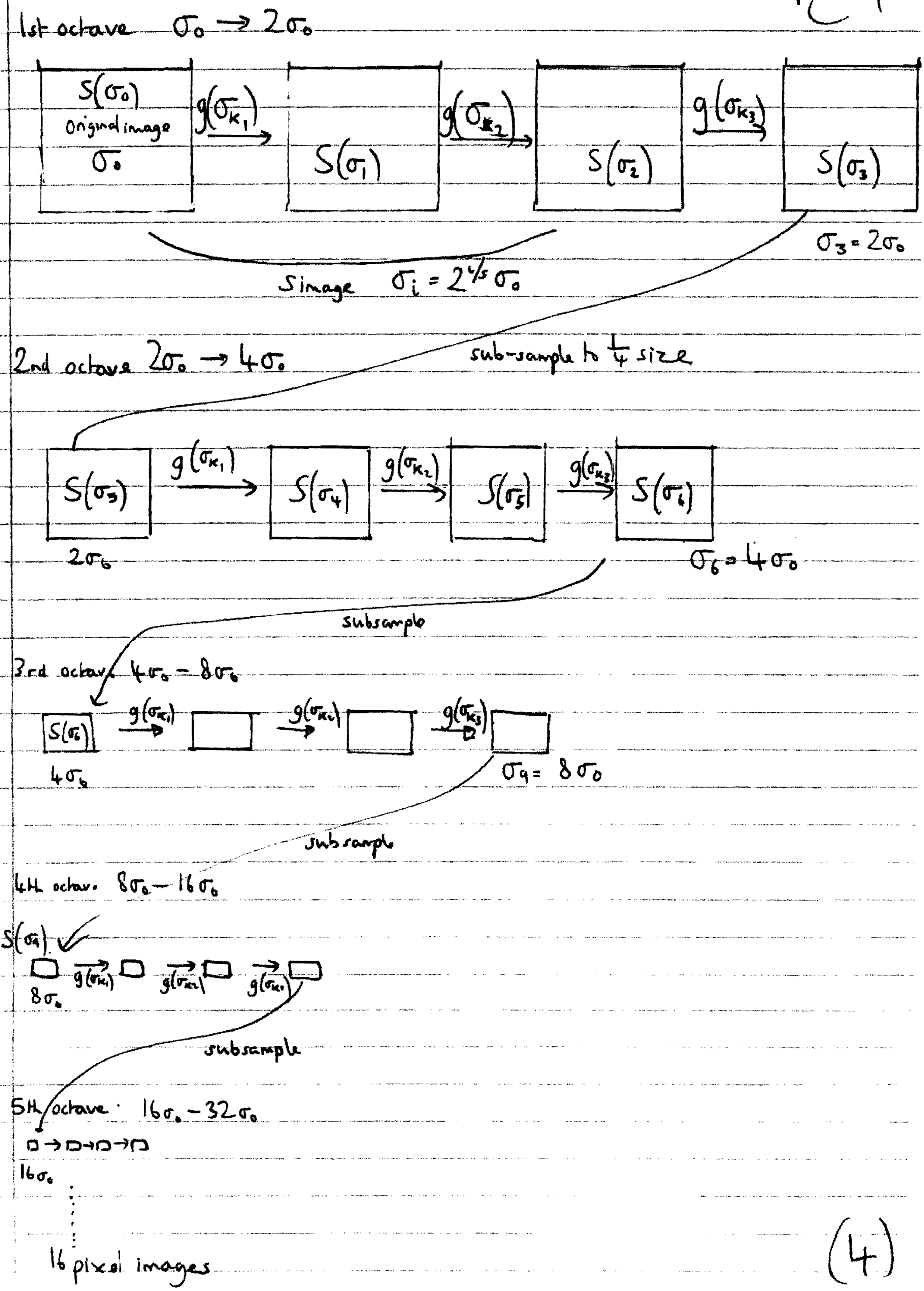
— Each blur is performed as 2 1D convolutions (see (a) ii).

— After scale has doubled (i.e. s blurs) resize image by subsampling by 2 (i.e. ¼ size image). We can represent blurred images with fewer pixels without loss of information (Nyquist). [biggest saving] (3)

— same (small) incremental blur kernels used in each octave. σ_{K_1}, σ_{K_2}, σ_{K_3} etc. on sub-sampled images, but really represent filtering with larger kernels, $2\sigma_{K_1}$ .... $4\sigma_{K_1}$ .. $8\sigma_{K_1}$.

R 4

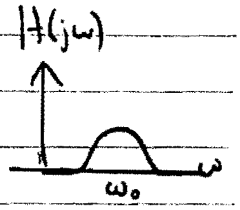
16b(i)


1st octave $\sigma_0 \rightarrow 2\sigma_0$
$S(\sigma_0)$
original image
σ_0
$g(\sigma_{k_1})$
$S(\sigma_1)$
$g(\sigma_{k_2})$
$S(\sigma_2)$
$g(\sigma_{k_3})$
$S(\sigma_3)$
$\sigma_3 = 2\sigma_0$
S image $\sigma_i = 2^{i/s}\sigma_0$
2nd octave $2\sigma_0 \rightarrow 4\sigma_0$
sub-sample to $\frac{1}{4}$ size
$S(\sigma_3)$
$g(\sigma_{k_1})$
$S(\sigma_4)$
$g(\sigma_{k_2})$
$S(\sigma_5)$
$g(\sigma_{k_3})$
$S(\sigma_6)$
$2\sigma_0$
$\sigma_6 = 4\sigma_0$
subsample
3rd octave $4\sigma_0 - 8\sigma_0$
$S(\sigma_6)$
$g(\sigma_{k_1})$
$g(\sigma_{k_2})$
$g(\sigma_{k_3})$
$4\sigma_0$
$\sigma_9 = 8\sigma_0$
subsample
4th octave $8\sigma_0 - 16\sigma_0$
$S(\sigma_9)$
$g(\sigma_{k_1})$
$g(\sigma_{k_2})$
$g(\sigma_{k_3})$
$8\sigma_0$
subsample
5th octave $16\sigma_0 - 32\sigma_0$
$16\sigma_0$
16 pixel images

(4)

16 b(ii) Band-pass filtering

— tuned to a small band of spatial frequencies (difference of high pass and low pass filtered o/p)



$$-\nabla^2 G_{\sigma_i} * I = \nabla^2 S(\sigma_i^2) \simeq S(\sigma_{i+1}^2) - S(\sigma_i^2)$$

i.e. subtract neighbouring images in same octave.

— generate a pyramid of DOG images by subtraction in same octave (2)

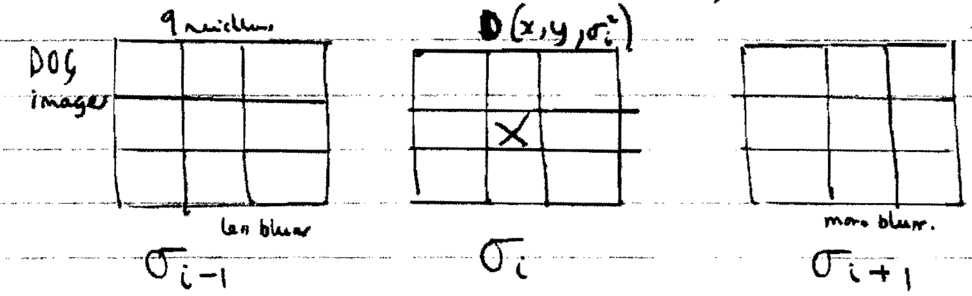
b(iii) Blobs are localised at max/min of $\nabla^2(G_\sigma * I)$ response.
Need to search over $\nabla^2 S(x, y, \sigma_i^2)$ for (max/min) local.

— More efficient to search over difference of gaussian (DOG) pyramid

$$\nabla^2 S \approx D(x,y,\sigma_i^2) = S(x,y,\sigma_{i+1}^2) - S(x,y,\sigma_i^2)$$

for max/min in x, y and σ.

— Evaluate 26 neighbours of $D(x,y,\sigma_i^2)$ to see if local max/min



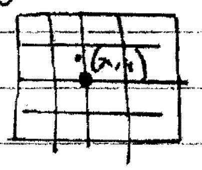
— Local max/min is blob location (x_0, y_0); scale (size of feature) is σ_i

— SIFT descriptor looks at 16×16 pixels sampled from $S(x,y,\sigma_i^2)$ image in correct octave.

RC 1

(Q17) SIFT, matching and classification

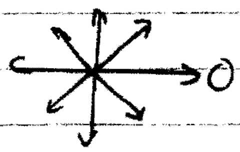
a (i) – For each interest point at location (x, y) and scale σ estimate the dominant orientation θ, by looking at histogram of edge gradients from $\nabla S(x, y, \sigma^2)$.
Bins 10° apart + smoothed by gaussian.



– Sample 16x16 gradients from (x,y) and aligned with θ at image of scale, σ

– Smooth these gradients with $g(1.5\sigma)$ to emphasise gradients at interest point.

– produce orientation histogram (HOG) for each 4x4 block (cell)
Each bin records grad. magnitude interpolated over 8 dir's (quadrants) only.



– concatenate 16 histograms of gradients (HOG's) to vector of 128D

– normalize to unit length, truncate to 0.2 to avoid illumination effects and normalize to 1.

– SIFT encodes 2D shape – invariant to lighting by using edges + normalization step + to exact 2D position by histogram / bin effects [pooling]
Edges taken around a blob in centre.

(4

17 a(ii) Each descriptor $\underline{d}_q$ is a 128D vector. Search for <u>nearest-neighbour</u> (first nearest $\underline{d}_1$ and second nearest $\underline{d}_2$) in all descriptors seen.
Done efficiently by arranging descriptors in a <u>kd tree</u> (binary search tree data structure).

Accept match if $\frac{\cos^{-1}(\underline{d}_1 . \underline{d}_q)}{\cos^{-1}(\underline{d}_2 . \underline{d}_q)} < r$ (threshold 0.6 → 0.7)

This is used to distinguish reliable matches from ambiguous matches (LOWE '99)
Alternatively measure <u>euclidean distance</u> between descriptors $\underline{d}_q$ and $\underline{d}_1$ and accept if close.

(iii). Each <u>feature</u> (SIFT descriptor) points to a target image.
Each match will vote for a target. (See <u>inverted file index</u>)

Can use total votes to hypothesise targets. Verify by estimating pose of target and checking agreement [geometric verification]

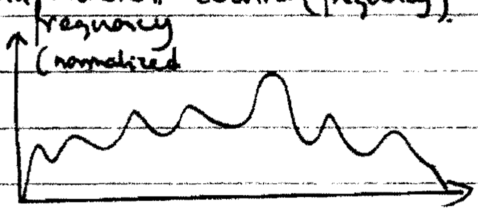
17(b) Visual words

(i). - Key idea of "visual words" is to quantize the descriptor (SIFT 128D) space by K-means clustering. Each word represents image patches which are similar in low-level appearance.

- Cluster centres make up "visual vocabulary" or "visual words". Once centres are extracted the corpus of descriptors can be discarded and hence very efficient in storage + matching.

- K is chosen between 200 to 1 million depending on application: categories vs specific target

(ii). Extract interest points, scales and orientations. Compute SIFT descriptor. For each descriptor find visual word (efficient Nearest Neighbour). and, Increment counter (frequency).



Visual word #
(K = 200 → 1000,000)

Each image rep. by histogram of visual word occurrences (ie. a vector of fixed length K in length)

Also: — modify by term frequency — inverse document frequency weighting

— remove "words" that are common to all images, ie. a stop list.

(iii). Use training data (supervised — ie. images with known categories)

Assign based on similarity of vectors:

— Nearest Neighbour

— SVM

— Boosting or pLSA.

SECTION G *Engineering for the Life Sciences*

Answer not more than two questions from this section

18 The cornea of the eye is largely comprised of the natural material collagen.

(a) Describe the structure and self-assembly of collagen. [4]

Answer: Collagen is a protein, a polymer made up of repeating amino acid subunits in the typical pattern Gly-X-Y where Gly is glycine and the other amino acids are commonly proline and hydroxyproline. (Hydroxyproline is unique to collagen and is thus used in collagen quantification assays in vitro.) Collagen molecules self-assemble into a triple helix, where three individual protein chains coil together to form a repeating subunit 1.5 nm in diameter and 300 nm in length. The three chains can be identical (type II collagen) or differ (type I collagen, with two identical chains and a third different chain). These triple helices then self assemble to form fibrils with a quarter-staggered array of helices giving rise to a measurable gap of 67 nm between consecutive helices, which results in the typical striped appearance of collagen in transmission electron microscopy and atomic force microscopy. Typical collagen fibrils are around 50 nm in diameter. Because of the strong self-assembly motive, collagen "ends" are difficult to locate experimentally and are nearly never observed, so there is no way to define the length of a collagen fibril, only the diameter. There is cross linking both within the triple helix and between adjacent triple helices to stabilize the protein assembly into a mechanically functional rope-like structure with excellent mechanical properties, akin to the "carbon fiber" of nature.

(b) How is collagen in the cornea organized? How does this differ from the collagen organization in the sclera? How does this relate to the function of the cornea? [4]

Answer: Collagen in the cornea is crystalline: it is organized into very regular perpendicular lamellae at 90 degrees cross-ply orientation, with uniform spacing between individual collagen fibrils within each lamella and regularly sized collagen fibrils (constant diameter). The regular structure is important both mechanically and optically. Mechanically, the lamellar organization provides resistance to intraocular pressure and allows the cornea to serve its critical function of providing two thirds of the optical power (fixed focus) of the eye overall. Optically, the regular crystalline structure of the collagen allows for corneal transparency, again required for unobstructed vision. The lamellar collagen organization is obviously not isotropic in the cornea, nor is it identical in the left and right eyes, but instead the collagen patterns in the two eyes are mirror images and the fibrils are directed towards the extra-ocular muscles. The sclera and the cornea are co-continuous structures, but have very different collagen fibril patterns according to their function. The collagen in the sclera is much less well-organized and this gives rise to the "white" appearance of the "whites" of the eyes as light is scattered rather than transmitted as in the regular structure of the transparent cornea. In disorders and diseases of the cornea, the tissue can become cloudy (i.e. more like the sclera) due to changes in the collagen organization resulting in a loss of transmissibility of light and the need for corneal transplant.

(c) Describe LASIK surgery in relation to the cornea. [3]

Answer: LASIK stands for Laser-ASsisted In-situ Keratomileusis, a minimally invasive surgical procedure used to correct vision defects, and falling into the general category of refractive surgeries. The basic procedure modifies corneal curvature, essentially changing the effective "lens" shape by which the cornea controls 2/3 of optical (fixed) focusing power. The procedure has three steps. First, a flap is created in the cornea. The corneal bed (stroma, or collagen layer) beneath the flap is then reshaped by excising tissue with a precisely controlled eximer laser. The corneal flap is then repositioned, it naturally adheres to the modified stroma and the tissue is allowed to heal. Because the cornea has been modified, a person who has had LASIK surgery is no longer eligible to be a cornea donor, leading to an increasing donor shortage worldwide.

(d) Both the collagen density and the collagen fibril diameter can influence different mechanical properties of a soft collagenous tissue such as the cornea.

(i) Consider a hypothetical disease in which the collagen fibril diameter is unchanged, but in which the collagen density increases. What effect does this change have on the elastic modulus of the tissue? What effect does this change have on the intrinsic permeability? How do the elastic modulus and intrinsic permeability relate to the function of the cornea?

Answer: Elastic modulus: collagen in soft tissues can be modeled as a series of springs in parallel, where not all of the springs are 'recruited' simultaneously, but instead sequentially. This gives rise to a stiffening response with increasing deformation, as effectively more springs are recruited as the deformations get larger. For springs in parallel, the overall stiffness is the sum of the spring stiffness of all recruited springs at that point of deformation,, and the maximum total elastic modulus is simply $k_{total} = n * k_{spring}$ where n is the number of springs and k_{spring} is the individual spring modulus. Thus, for an increase in the collagen density with no change in fibril size, n would be larger with no change in k_{spring} and the overall stiffness k_{total} would be greater. The stiffness would affect the equilibrium shape of the cornea, and thus the optical focus. Permeability: the intrinsic permeability goes as $a^2 f(\phi)$ where $f(\phi)$ is a decreasing function of the solids fraction, or approximate collagen density. Thus, for increasing collagen density, ϕ decreases, and the overall permeability p also decreases. Decreased permeability means that fluid is less easily transported through the cornea at the same pressure level and this may affect the circulation of tear fluid within the eye detrimentally, as well as influencing the optical transparency of the tissue if it is not hydrated to equilibrium hydration, required for maintaining collagen fibril spacing and transparency.

(ii) Now consider a different hypothetical disease, in which the collagen density overall does not change, but for which the diameter of individual collagen fibrils is decreased. What effect does this change have on the elastic modulus and intrinsic permeability, and how does this differ from the change in part (d)(i) above?

Answer: The intended interpretation of "density" here was for fixed number of fibrils per area. In this case, the following holds. Elastic modulus: The spring stiffness for a single collagen fibril k_{spring} goes as EA/L where A is the cross-sectional area, proportional to the square of the collagen fibril size a. Thus, $k_{spring} \propto a^2$ for constant intrinsic material modulus E. Thus, as $k_{total} = n * k_{spring}$, k_{total} would be decreased with a decrease in collagen fibril diameter. This effect would be more dramatic than the effect of increasing the collagen density, in part (i), as the effect is quadratic instead of linear for k_{total}. In this case the intrinsic permeability would be increased and there would be more room for fluid to flow through the tissue. This is precisely the opposite as in part (i).

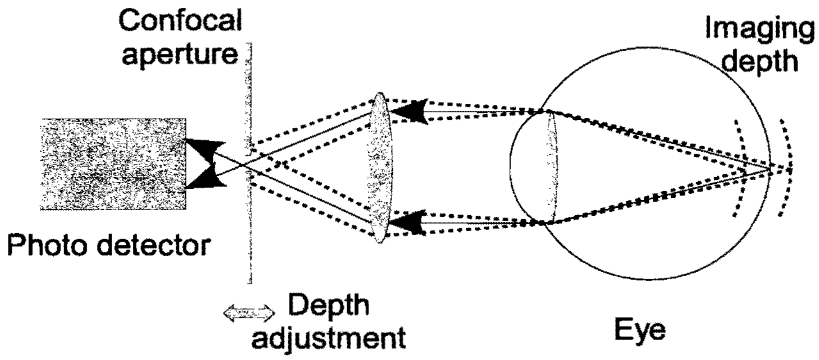
An alternative interpretation of this question would be that the areal density was unchanged, such that there were many more fibres of smaller size.

In this instance the overall modulus would not have been altered since the total cross sectional area of fibrils was not changed. As intrinsic permeability is proportional to a^2, the permeability will decrease with a decrease in fibril diameter (at fixed solids fraction) in a quadratic manner. Whether this is a greater or smaller effect than the collagen density effect will depend on the precise nature of $f(\phi)$, which depends on precise details of the tissue collagen organization. Thus this differs from case (i) in that in both cases the permeability changes, affecting fluid flow and nutrition in the eye, but at fixed modulus, which means there is no change in reaction to the intraocular pressure in the eye.

[9]

19 (a) Describe, with an appropriate diagram, how the confocal optics in a Scanning Laser Ophthalmoscope (SLO) allows depth sectioning of the fundus. [4]

Answer: Depth sectioning describes the process of being able to record light scattered from tissue at a specific depth, rather than the total light scattered from all depths. This enables the collection of 3D data, since it allows independent measurements to be made at different depths in the tissue.



In the Scanning Laser Ophthalmoscope this is achieved by confocal optics. The diagram above shows the *reflected* light path in this case. A spot at the back of the retina is illuminated with laser light. The reflected light is focused onto the photo detector through a small hole in a sheet known as the confocal aperture. Any light which is scattered from deeper within the retina will have a focal point *before* the aperture, and hence only a small fraction of this light passes through the aperture. Any light which is scattered from shallower depths in the retina has a focal point *after* the aperture, and once again only a small fraction of this light passes through the aperture. Only light at the focal depth passes through the aperture un-altered.

Hence the sectioning depth (i.e. the depth at which the instrument is responsive to scattered light) can be altered by moving the confocal aperture towards or away from the preceding lens.

(b) A light beam is focused at a thin confocal aperture of diameter d. The beam cross-section is a disc of radius r given by:

$$r^2(z) = r_0{}^2\left(1+\left(\frac{\lambda z}{\pi r_o{}^2}\right)^2\right)$$

where z is the axial distance from the confocal aperture, r_0 is the radius at the focal point, and λ is the wavelength of light. Assume that the lens is placed in air.

(i) What is the approximate radial resolution of the lens, $\Delta x = 2r_0$, in terms of the numerical aperture NA and wavelength λ? [5]

Answer: The numerical aperture NA is defined in terms of the angle that light from the focal point to the diameter of the lens makes with the axis:

$$\text{NA} = sin\theta \approx \frac{D}{2f}$$

where D is the lens diameter and f the focal length, and the lens is placed in air. Noting that $r_0 = \frac{\Delta x}{2}$, and that the beam radius must be $\frac{D}{2}$ at the lens location $z = f$:

$$\left(\frac{D}{2}\right)^2 = \left(\frac{\Delta x}{2}\right)^2 \left(1 + \left(\frac{4\lambda f}{\pi \Delta x^2}\right)^2\right)$$

However, $D \gg \Delta x$, hence the numerical value of the $(1 + \ldots)$ term must be very large, and we can therefore ignore the 1 in this term, giving:

$$\left(\frac{D}{2}\right)^2 \approx \left(\frac{\Delta x}{2}\right)^2 \left(\frac{4\lambda f}{\pi \Delta x^2}\right)^2$$

Taking the square root of both sides gives:

$$\frac{D}{2} = \frac{2\lambda f}{\pi \Delta x}$$

And hence:

$$\Delta x = \frac{4\lambda}{\pi}\frac{f}{D} \approx \frac{2\lambda}{\pi \text{NA}}$$

(ii) The axial resolution Δz is defined as the range in z over which at least half of the beam cross-sectional area would pass through the aperture. Show that:

$$\Delta z = \frac{2\lambda}{\pi(\text{NA})^2}\sqrt{\frac{1}{2}\left(\frac{d\pi \text{NA}}{\lambda}\right)^2 - 1}$$

and evaluate Δz for NA $= 0.4$, λ = 800 nm and $d = 20\mu$m. [4]

Answer: At $z = \frac{\Delta z}{2}$, the radius must be $r = \frac{d}{\sqrt{(2)}}$ so that half the beam area passes through the aperture. Hence:

$$\left(\frac{d}{\sqrt{2}}\right)^2 = \left(\frac{\Delta x}{2}\right)^2 \left(1 + \left(\frac{2\lambda \Delta z}{\pi \Delta x^2}\right)^2\right)$$

Re-arranging in terms of Δz:

$$\Delta z = \frac{\pi \Delta x^2}{2\lambda}\sqrt{\left(\frac{\sqrt{2}d}{\Delta x}\right)^2 - 1}$$

Substituting for Δx in terms of NA from (i) gives:

$$\Delta z = \frac{2\lambda}{\pi(\mathrm{NA})^2}\sqrt{\frac{1}{2}\left(\frac{d\pi\mathrm{NA}}{\lambda}\right)^2 - 1}$$

Evaluating this with NA $= 0.4$, $\lambda = 800\,\mathrm{nm}$ and $d = 20\mu\mathrm{m}$ gives:

$$\Delta z = 70.64\mu\mathrm{m}$$

(iii) First by increasing NA alone, then by decreasing d alone, find the minimum achievable Δz (in each case keeping other values as in (ii)). Also list in each case the value of NA and d which you used. [2]

Answer: If we set NA $= 1$, this gives:

$$\Delta z = 28.28\mu\mathrm{m}$$

By reducing d we can achieve an apparent $\Delta z = 0$, in which case:

$$\left(\frac{d\pi\mathrm{NA}}{\lambda}\right)^2 = 2, \quad \Rightarrow \quad d = \frac{\sqrt{2}\lambda}{\pi\mathrm{NA}} = 0.9\mu\mathrm{m}$$

(iv) What are the problems with the improvements outlined in (iii)? [2]

Answer: When imaging in air, we can not increase NA to above 1.0, in fact it is difficult to produce lenses with NA greater than 0.95. This can be improved by placing the lens in oil, in which case an NA of 1.55 is possible. However at such a high NA we are also at risk of reducing the depth range of the SLO. Using a very small confocal aperture will limit the amount of light and hence dramatically reduce the sensitivity of the system. In practice we need the aperture to be at least twice the light wavelength, if not more. The $\Delta z = 0$ result is not really valid: this just implies that even at the focus, only half the cross-sectional area of the beam is passing through the aperture.

(c) What factors, other than lens resolution and aperture size, limit the radial resolution and the depth range of an SLO? [3]

Answer: The radial resolution of an SLO is also limited by the spot size of the laser. This is a more major factor then Δx, since aberrations in the eye limit the size to about

15 μm or so. Hence we really need the confocal aperture to be large enough to let the light from this entire spot through.

The depth range is also limited by how far the light will penetrate into the tissue. This is affected by the light intensity, wavelength and material attenuation. Light intensity is limited since we do not want to damage the fundus. Longer wavelengths (lower frequencies) will tend to penetrate further into tissue, but at the cost of reduced radial resolution.

20 (a) With regard to the mechanisms of visual processing:

(i) Describe the differences between ganglion cell, simple cell and complex cell receptive fields.

Answer: Ganglion cells have roughly circular receptive fields from a few minutes of arc at fovea, to few degrees at periphery. On-centre ganglion cells have an excitatory receptive field centre and inhibitory surround thereby respond optimally to differential illumination. Off-centre ganglion cells have an inhibitory receptive field centre and excitatory surround. There are also colour opponent cells which are either red- green or blue-yellow. In the visual cortex simple cells respond to either a line or edge detector in a particular location and complex cell respond to a line with a particular orientation moving in a particular direction.

(ii) What is colour opponency and why might it be a useful way for the visual system to code colour?

Answer: Colour is processed in an antagonistic manner with red vs green, blue vs yellow and black vs white. This is an efficient coding scheme based on Principal Components Analysis (PCA) of the reflectance properties of natural objects and corresponds to the first three principal components.

(iii) Describe the evidence that the visual system processes different spatial frequencies within an image independently.

Answer: The theory proposes that the two-dimensional luminance of an image is represented as a combination of elements such as sinusoidal gratings of different spatial frequencies, amplitudes, phases and orientations. Therefore the neurons are effectively doing a two-dimensional Fourier analysis. Importantly the theory suggests that the different components are processed by separate channels. Evidence for this comes from fatiguing one particular frequency leads to a dominant perception of the other spatial frequencies such as in the spatial frequency aftereffect. In addition the processing of each frequency is independent of others and so if two different frequencies sinusoids are combined the detectability of the stimulus is independent of the phase of the combination. In addition, neurophysiological evidence from the cells in visual cortex show receptive field properties which are wavelet like (sinusoid windowed by a Gaussian) of varying spatial frequencies.

(cont.

(iv) Describe the cues to object depth that are missing from a photograph.

Answer: The cues that are missing from a photograph are

- Motion parallax which is the result of changing positions of an object in space due to either the motion of the object or of the viewer's head
- Accommodation cues due to the adjustment of the focal length of the lens
- Convergence cues due to the angle made by the two viewing axes of a pair of eyes
- Binocular disparity: the disparity between images of the same object projected onto the retinas

[10]

(b) With regard to optimality in visual processing:

(i) Describe the source of the resolution limit of compound eyes (with the corresponding mathematical formula).

Answer: The resolution of compound eyes is limited by diffraction, because for small lenses (such as those in the ommatidia of compound eyes) the point spread function in inversely proportional to the diameter of the lens (or aperture), D: $\frac{2\lambda}{D}$, where λ is the wavelength of the light.

(ii) Derive the formula for the response of a neuron, r, as a function of the stimulus (contrast level of light), s. Assume that the neuron encodes this stimulus with an information theoretically optimal tuning curve, with negligible noise in its response, under the constraint that its response lies between r_{min} and r_{max} and is a monotonically increasing function of the stimulus, and the stimulus is uniformly distributed between s_{min} and s_{max}. In your derivation:

- explain what makes a tuning curve optimal in this case (i.e. what is the quantity that the tuning curve needs to optimise),
- state what is the response distribution of the neuron in this case,
- explain the steps taken to derive the final formula for the tuning curve.

Answer:

- An information theoretically optimal tuning curve maximises the mutual information between stimulus and response. Since response noise is negligible, the tuning curve that maximises mutual information is the same as the one that maximises response entropy.

•The maximum entropy distribution under constraints of a minimal and maximal response is the uniform distribution.

•To achieve a uniform response distribution from a uniform stimulus distribution a linear tuning curve is needed. With the constraints on the minimal and maximal stimulus and response values, this linear tuning curve must have the following formula: $r = r_{\min} + (r_{\max} - r_{\min}) \cdot \frac{s - s_{\min}}{s_{\max} - s_{\min}}$

(iii) Describe the conditions under which two line segments may appear to belong to the same physical object despite not being connected, and how this reflects statistically optimal computations in the brain.

Answer: Two unconnected line segments appear to belong to the same contour when they are close to each other and approximately co-linear. This is statistically optimal because in the natural environment the contours of objects frequently give rise to such line segments due to occlusion and partial view.

[10]

[redacted]

2012 IB Paper 8, Section H – Crib

21 (a) List and describe four sources of invention that could be classified as technology push, and four sources of invention that could be classified as market pull. Provide one example for each of the sources.

Technology push:

- The post-it note arose by **accident**
- The dvd player arose by **analogy** and the Dyson was a **transfer** of an existing industrial technology to a new domestic applications
- The domestic breadmaker arose from a **structured search** for new kitchen appliances
- The 'inerter' arose due to a gap in an existing **map** of possibilities
- **New materials** allowed the hair dryer to move from an expensive metal body to a cheaper plastic body.

Market pull:

- The 'aural' thermometer for babies arose from the difficult **experience of using** conventional mercury thermometers measuring babies temperatures with
- The chopper bicycle arose from modifications to existing bikes by enthusiastic **users**
- Fridges and washing machines are now sold as **fashion** items as the kitchen has become the main entertaining room
- The ink-jet printing industry around Cambridge has grown due to **legislation** on sell-by dates for food
- The model-T Ford was successful because Ford found ways that by making cars cheaper he could turn a **luxury product into a common one**

(b) Describe the four generic ways in which a market may be segmented using a perceptual map.

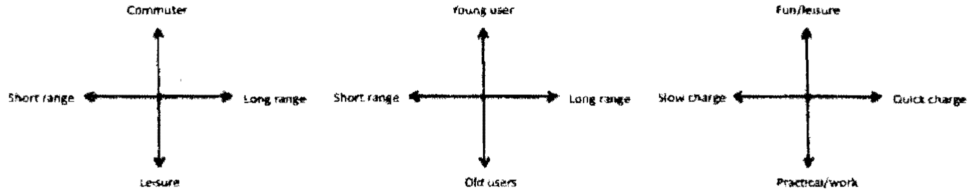
There are four possible ways in which this can be considered.

- By the **benefits** that are delivered: what benefits do customers and users derive from the product? For a hedge trimmer, it might be a neat hedge, or perhaps versatility, or perhaps a great service deal.
- By particular **product attributes**: perhaps the easiest way to segment a market is to compare product attributes. This however tends to say little about the customers and is often the weakest approach. For lawn mowers, this could include product performance, width of cut etc.
- By **characteristics of the consumer**: this can be split to include demographics, and psychographics. Demographics relates to aspects such as social class, age, house size, sex etc. Psychographics relates to the user's attitudes and beliefs, what they feel, their lifestyle etc.
- And by **product use**: describing ways in which a product is used. There might be strong customer loyalty, products might be used rarely or frequently, it could describe purchase behaviour (e.g. Distress purchase, seasonal patterns or regular upgrades etc).

(c) List the two characteristics of a well-defined market segment.

Homogeneous within a segment (all customers display similar characteristics or preferences towards the product). Heterogeneous between segments: each segment has clearly defined differences in preferences and characteristics as compared with other segments.

(d) Sketch three different perceptual maps that could be used to segment the markets for a low cost electrical bicycle. Explain your choice of axes used in each of your perceptual maps.



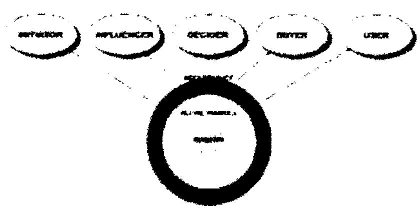
The choice of axes provides the ability to segment the market into groups that would be real and substantial. The choice of axes in the first map allows grouping of potential customers based upon characteristics of product use, and product attributes. The second map compares characteristics of consumers (age demographics) against product attributes. The final map compares use and attributes, with the key issue of battery charge being used rather than range.

The maps shown here are examples rather than being the only answers. Other equally valid axes include speed; radical or traditional configuration.
Examiner's Report: *Parts (a) and (b): generally well answered. Answers to part (c) were mixed, with some missing the point altogether. There were some very good answers in part (d), with some thoughtful proposal for maps. However, marks were lost by many others who provided no explanation of the axes and what the maps were trying to achieve.*

22 A team of university researchers have developed a novel device for detecting very small quantities of explosives in the air with very high levels of reliability. The researchers believe that this device could be used to improve the effectiveness of airport security systems.

(a) Describe who the potential stakeholders would be in the market for such an airport security device.

Generic types of stakeholders can be described in this chart:



For this particular market, stakeholders could include: security personnel at the airport (to get the views of those who will actually use the technology – the user); their employers (in case specific training/skill-set is required); airport management company procurement department (the purchaser of the technology); the decider (the airport management company); the airlines, media and passengers (influencing the decision to purchase); and government (ensuring that air travel is safe – and may actually be the initiator if they wish to see new standards in place for airport security).

(b) Describe how interviews and observations could be used to understand the needs of customers and users of such a device.

Interviews: these are useful for capturing basic information on behaviours (frequency, where, when, how much questions) and attitudes (but need to be very careful with wording of questions). Interviews could be used with the security staff, airport management company staff, and airlines to capture information on particular concerns and needs. Advantages of interviews are that they are relatively quick and cheap.

Observation: these are useful for uncovering issues / needs that consumers might not be able to identify if just asked a question. For this situation, user observation of airline security staff interacting with equipment and passengers could be revealing, but also observation of installing/ maintenance staff and the way in which they install and manage the technology could also be very helpful. The downside of observation is that it can be time consuming and costly to analyse.

(c) The device combines standard and bespoke components, and the reliability of the device is highly dependent on the quality of the components and the assembly process. It is anticipated that the market for this device will be around 100 units per year. Discuss the issues that would need to be considered in designing a production system for this device.

Reliability of this technology will be critical – failure to detect an explosive could have serious consequences. Customers are likely to be quite price insensitive if the have the confidence that this system will work. This means that there should be no unnecessary cost-cutting in materials, assembly, installation and maintenance.

The production system will need to be designed to ensure that the quality of devices installed is very high. Therefore, the team could consider minimising use of sub-contractors to help in ensuring that tight control is kept over ever step of the production process. However, this may not be possible if highly specialised skills are required for specific stages of the process. If that is the case, strong links would need to be maintained with key suppliers, and having suppliers located close by could be an advantage.

As the volumes are quite low, automation may not be appropriate. However, if precision assembly is required for which manual operation is less reliable, automation may be appropriate at certain steps in the assembly process.

Given the safety critical nature of the technology, and the problems that would be incurred if the device breaks down, the team will need to maintain accessible stock of components, and to have replacement systems available at all times.

Attention would also need to be given to ensure that installation is a tightly integrated step in the production process, and that this should either be done by the team themselves or by specially trained (and monitored) third party installation engineers.

In addition, the process should not stop at the installation stage: consideration would need to be given to how the technology would be monitored, maintained and upgraded over its lifetime of use.

Examiner's Report: Part (a) was generally well answered. Answers to part (b) often demonstrated rather weak appreciation of the use of interviews and observation for gathering information on user needs. Part (c) was not well answered. Many candidates focused on just one aspect, normally whether or not to automate, and ignored all other considerations.

23 (a) List the four tests an invention must satisfy in order for it to be

1. It must be **novel**;
2. involve an **inventive step**: i.e. not be obvious to someone in the light of what has been done before (the 'prior art');
3. have a **practical application**: be capable of being made or used in some kind of industry; and
4. **not be excluded** (e.g. scientific theory or mathematical method, method of doing business, perpetual motion machine....)

patentable.

(b) Explain why it is desirable to have more than one claim in a patent.

- The claims define in precise words the exact scope of the monopoly
- The first claim usually includes all the *essential* elements of the invention – it should not include non-essential elements as it should aim to be as general as possible without including prior art (because if it does, it will not be valid)
- If the first claim includes non-essential features, then a competitor can simply omit those features in his/her product and will not have infringed the patent
- Subsequent ('dependent') claims usually add extra features to the earlier claims, narrowing the scope – the aim being to provide a fall-back in case the earlier claims are invalidated – either during the patenting process or later.

(c) Provide one strength and one weakness for each of the following types of business model:

(i) product; (ii) service; (iii) service enabled by a product; (iv) product with consumables.

Sell a product: Strength = Increase revenue with higher volumes; Weakness = Need to develop, make, distribute, support, etc.

Sell a service: Strength = No manufacturing costs; Weakness = Can be hard to scale up the business (in many cases can only grow through recruiting and training lots of people).

Sell a product plus services: Strength = Long-term revenues on the back of the sale of each product, Weakness = Need infrastructure to provide services.

Sell a product plus consumables: Strength = Long-term revenues on the back of the sale of each product; Weakness = Need infrastructure to deliver consumables.

(d) The website of the UK semiconductor firm ARM plc states that its business model "*[..] involves the designing and licensing of IP [intellectual property] rather than the manufacturing and selling of actual semiconductor chips*".

(i) Describe what is meant by 'licensing of IP' in this context.

ARM plc maintains ownership of the IP, but allows others to use it in return for a fee. This fee typically combines an up-front payment, and an on-going stream payments for each time the IP is embodied in a product.

(ii) Discuss the relative merits of a business model based around licensing of IP in comparison to a business model that is focused upon the manufacturing and selling of semiconductor chips.

Licence the IP

Capital investment required is relatively low. There are the costs of generating the IP and finding/managing customers, but these are far lower than the cost of building and running a semiconductor plant. In addition, this business model is quite flexible; if new IP is generated relevant to new markets, new customers can be targeted without the substantial costs of reconfiguring or replacing manufacturing equipment. This business model can also be developed quite rapidly, and customers may be willing to try licensing a technology from a new firm, and if they don't end up making any products, they don't have to make the on-going payments.

The downside is that the firm may only get relatively small % of actual value of the technology. In addition, the challenges of finding and manage licensees can be substantial. Licencees sometimes do not like to be over-reliant on technology from outside and may seek to find ways to do things themselves.

Manufacturing and selling

By being an integrated chip company (like Intel and AMD), these firms are able to overcome the weaknesses of the licensing model and get 'all' the value generated. They are in control of the whole process and do not need to be reliant upon external parties to get their product to market. By being involved in the manufacturing process, they may also be better positioned to spot opportunities for development that would not be visible to external technology providers.

However, the downside of this business model is that you need to raise all the investment/capital to build and run a factory (or factories) and will be responsible for the whole process of designing, sourcing, assembling, distributing, after sales, etc. By having tangible capital assets (factories and machines) that need to be making money, this may make the firm more rigid and less able to respond to changes in demand, and threats from new technologies.

Examiner's Report: *Part (a) was generally well answered. Answers to part (b) were patchy: most candidates had some idea of the purpose of claims, but understanding was imperfect. Part (c) was very well answered. Many candidates were so enthusiastic to impart their knowledge that they wrote disproportionately long answers. Full comparison between the business models in part (d) was not well done. Most answers covered only aspects relating to IP advantage, and there was little discussion of the manufacturing model beyond the most basic mention of manufacturing costs.*