Module 3G4: Medical Imaging & 3D Computer Graphics

Solutions to 2017 Tripos Paper

1. **Locating the source of the imaging response.**

   The following notes cover a full range of possible answers to the question: an individual answer would not be expected to contain all these points.

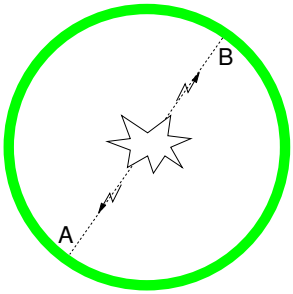
   (a) (i) The collimator on a gamma camera absorbs all radiation apart from that emitted in from a narrow cone-shaped region of space. It is therefore known that the radiation reaching the detector at the back of the collimator has come from a specific, narrow, conical region of space. The purpose of the collimator is to block some of the incident radiation in this way, so that the location of the possible source is seriously constrained. This enables the resulting data to be used to form an image or as part of a reconstruction algorithm.

   The collimator takes the form of thick sheet of lead with narrow holes passing though it, aligned with each detector. The radiation sources are on the opposite side of the sheet from the detectors. Radiation that is incident normally on sheet and aligned with a hole will penetrate to the detector at the back. Obliquely incident radiation will simply be absorbed by the lead. Hence, the collimator enables each detector to provide information about the radiation in a specific region of space, along the axis of the cylindrical hole.

   This sort of collimator is a simple and reliable device, with predictable performance. It has the disadvantage of absorbing a large proportion of the incident radiation, which increases the patient dose necessary to form a satisfactory image. Because of the conical sensitivity volume, the resolution degrades with increasing distance from the surface of the collimator. [15%]

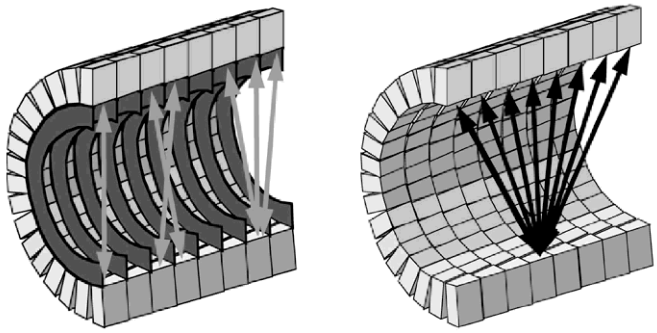
   (ii) The collimator in a PET scanner is designed to determine the line in three-dimensional space on which a positron has been emitted. This can be achieved without a mechanical collimator device by using a method based on coincidence detection that is much more sensitive.

   When a positron is emitted, it rapidly mutually annihilates with an electron creating two 511 keV photons travelling in opposite directions. The collimator system works by detecting *both* photons from the annihilation event at independent modules of the photomultiplier tube array. If two photons are detected within 10 ns of each other, the event is recorded and used in the PET reconstruction. The line of projection is the line linking the two detectors A and B which "lit up" at the same time.



This approach is more complicated than a mechanical collimator, but much more sensitive because radiation arriving at the sensor from almost any direction can be used. Resolution does not degrade in the same way with increasing distance from the surface of the sensor. The positron may be assumed to have been emitted on the line joining the two coincident 511 keV photons of annihilation received at the detectors. The only exception to this is when the positron happens to be travelling at speed when it hits an electron, and this does not happen often in a medical scanning context. [15%]

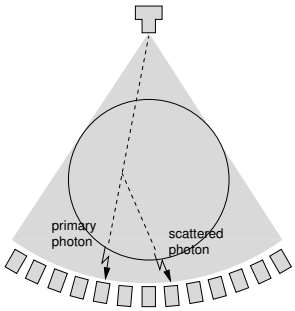
(iii) Between the detector rings in a PET scanner are retractable septa. With the septa deployed, only coplanar coincidences are possible and the machine operates in 2D mode. Note, however, that the vast majority of emitted photon pairs are absorbed by the septa.



(Figure reproduced from *P. Suetens, Fundamentals of Medical Imaging, Cambridge University Press, 2002*, by kind permission of the author and publishers.)

The sensitivity can be dramatically increased by retracting the septa. Then all possible projection lines are accepted, but full 3D reconstruction algorithms are required. Also, measuring the subsequent high count rates becomes difficult. There are more single module rejections (multiple impacts within the 300 ns scintillation time), and more multiple pair rejections (more than two photons hit the array within the 10 ns coincidence window). [15%]

(iv) X-ray CT scanners do not need a mechanical collimator to determine the source of the radiation because this is known to be the X-ray source. The direction of travel is also known, provided the radiation has not undergone scattering in the subject.

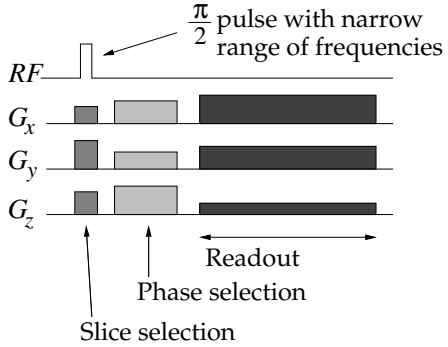


However, CT scanners *do* have a detector collimator. this is to stop scattered photons from reaching the detector and degrading the reconstructed image. Scattered photons will often arrive from a direction that is not aligned with the X-ray source, and can therefore be easily eliminated. Used in this way, a mechanical collimator is an efficient device that does not absorb a significant proportion of the primary imaging radiation. It is, however, fulfilling a fundamentally different role to that of a collimator in a gamma camera. [15%]

(b) In MRI imaging, spatial location is achieved using the link between the magnetic field a nucleus is in and the frequency of RF radiation that it is able to absorb or emit. The constant of proportionality in this relationship is the *gyromagnetic ratio*.

An MRI machine has a fixed strong magnetic field, plus a set of "gradient coils" that create variations in this field. All the gradient coils create fields that are in the same direction as the primary strong field. The gradient coils vary the intensity of the field as a function of position in any direction. This is achieved by having three gradient coils which each control the variation of the intensity of the field in one of the coordinate directions.

A single configuration of gradient field intensities provides planes which experience the same magnetic field strength. Using two gradient field combinations, in different directions, it is possible to isolate a response to some point along a line. Hence, using three gradient field combinations, the response of a single point can be identified.



This is done as follows. The first stage is called *slice selection* and takes place when the magnetic spins are disturbed at the start of the MRI imaging sequence. It makes used of the fact that only the slice in the field that implies a resonant (lamor) frequency the same as the input RF pulse gets rotated by the required angle ($\pi/2$ in the simplest case).

The second stage is called *phase selection.* A new combination of gradient settings are

applied for a short time to set the phases of the spins to different values in different planes.

The third phase is called *readout.* Further different gradient settings are used to generate frequency encoding during reception of the emitted NMR signal. [20%]

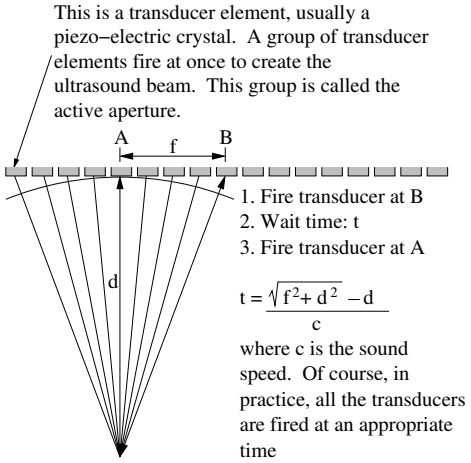
(c) The location of a scatterer in 3D space is done through the way that the ultrasound beam is transmitted and received. In a specific transmit cycle, the beam is transmitted in such a way that only certain known points in space are excited by the wave at each point in time. This is called transmit focussing.

Similarly, during the receive process, the transducer responses are combined so as to produce data that is only sensitive to limited spatial regions at each point in time. Hence the localisation of a response can be further improved. This is called receive focussing.

It is best to consider each of the three directions individually. They are: *axial:* in the direction of wave propagation. *Lateral:* the other direction within the B-scan plane. *Elevational:* perpendicular to the B-scan plane.

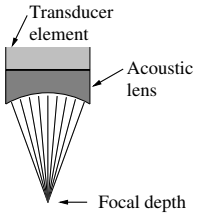
Location in the axial direction is determined by the time of arrival of the backscattered pulse. Signals that arrive first come from more superficial scatterers; signals that arrive later come from scatterers that are deeper in the body.

Location in the lateral direction is determined by performing both transmit and receive focussing with the transducer elements across the face of the ultrasound probe. Transmit focussing determines the times at which the elements fire. Receive focussing determines the delays imposed before the received responses are added together.



During the receive process, a dynamic technique is used to adapt the focussing delays and enable focussing over a wider range of depths.

Location in the elevational direction is achieved by focussing (in both transmit and receive) using an acoustic lens.



This provides an elevational focus at a single axial depth for both transmit and receive. Some more recent machines have a row of 3, 5, or 7 transducer elements in the *elevational* direction and so can achieve elevational focus using delays, as for the lateral direction. This is called $2\frac{1}{2}$D imaging. [20%]

Assessor's remarks: Part (a) was generally well answered. A smaller number of candidates were successful in answering part (b) on spatial location in MRI; even fewer got full marks in part (c) on the focussing and axial resolution in ultrasound.

2. **Marching cubes**

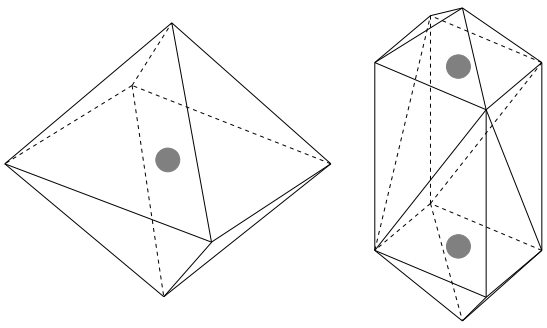
(a) Surface normals are important because these allow us to calculate how the light reflects off the surface and hence how the surface should be shaded when visualising it. The direction in which a surface normal points is also important: this can be used to indicate whether the surface is facing towards or away from the camera. For closed surfaces, triangles facing away from the camera do not need to be rendered.

Surface normals can be calculated from each triangle by taking the cross-product of any two sides. However, in order to get the direction right, the sides have to be referenced in a consistent order. Point normals can be calculated by taking the average triangle normals of all triangles which contain that point. [15%]

(b) Marching cubes is a process for creating a surface from 3D sampled data, consisting of a closed mesh of connected triangles. The surface is an iso-surface within the data at a particular data threshold. A single cube of data is considered, with samples at the cube vertices. If these samples are either all above or all below the threshold, there is no more to do for this cube. Otherwise, for any cube edge with one sample above and one sample below, the intersection of the surface with that edge is found by linear interpolation along that edge. Then a set of triangles are fitted within the cube, connecting these intersection points, which separate the space within the cube into regions which are either above or below the threshold. The arrangement of triangles is from a look-up table, which stores the 15 possible cases according to the values of the samples at the corners of the cube. The process is repeated at every cube location, and the resulting triangles accumulated to make the whole surface.

The triangles from Marching Cubes can have very different sizes and aspect ratios. This is important, since triangles with very poor aspect ratios (very thin) result in very poor estimation of surface normals, and also poor interpolation of these normals over the surface. They hence make surface renderings look less effective. [25%]

(c) (i) The function is spherical, with value equal to $0.6^2 - r^2$, and a centre at $\{x_0, y_0, z_0\}$.

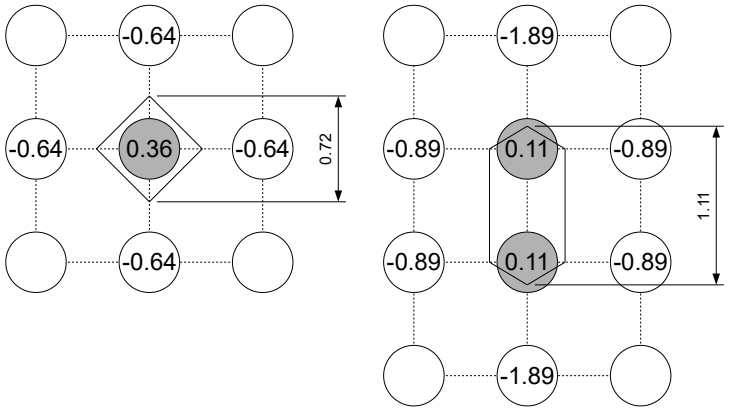


Surfaces are shown above. In the first case (left hand image), there is a sample at exactly the centre of the function with value 0.6^2 and the nearest samples, which are 1 unit away in either x, y or z, have value $0.6^2 - 1$ and are hence negative (outside the surface).

In the second case (right hand image), the centre of the function is exactly between two samples, so each of these samples has value $0.6^2 - 0.5^2$. This is still above zero, so these are considered to be inside the surface, though the sample values are much smaller. All other samples are outside the surface.

In the third case, the centre of the function is equidistant from samples in each of the x, y and z dimensions, so the closest samples have values $0.6^2 - 3 \times 0.5^2$. Hence all the samples are below zero and no surface is extracted at all. [20%]

(c) (ii) In order to find the intersection points, we need the function values either side of the surface. These are given below for a 2D plane which includes the maximum width of the surface in each case:



The first case is shown in the left hand image, and the maximum width is hence $2 \times \frac{0.36}{0.36+0.64} = 0.72$.

The second case is shown in the right hand image - the maximum width is in the vertical direction, and is $2 \times \frac{0.11}{0.11+1.89} + 1 = 1.11$.

The third case has no surface, so the maximum width is 0. [25%]

(c) (iii) At a much higher resolution, we measure the actual zero iso-surface of the function, which is spherical, and the diameter is hence $2 \times 0.6 = 1.2$. [5%]

(c) (iv) Clearly Marching Cubes must be used with care when looking at very small features. The real surface in this case should be spherical, but the actual surface is either octahedral, or long and thin, or does not exist at all. In addition, Marching Cubes relies on linear interpolation, and this will not always reveal the actual surface location, which is why the width is incorrect even for the case where the spherical location is actually centred on a data sample. [10%]

Assessor's remarks: This was a less popular question. The bookwork in the first two parts of the question was generally answered well. There were also some excellent sketches for part (c)(i), with most candidates noting that the surface in C did not exist. The linear interpolations in (c)(ii) were more variable, with some perfect answers but many which

simply used the spacing between pixel centres. Most noted the difficulty of sampling small features in (c)(iv) though few also noted the errors as a result of using linear interpolation on a spherical function.

3. **Catmull-Rom Splines**

(a) A Catmull-Rom spline is a type of parametric cubic curve which consists of a segment, defined in a parameter t, and with starting point $t = 0$ and ending point $t = 1$. Each of the x and y (and possibly further) coordinates is defined as a separate cubic function of t. The coefficients of this function are found by a matrix multiplication, so that the curve $C(t)$ is:

$$C(t) = \frac{1}{2}\begin{bmatrix} t^3 & t^2 & t & 1 \end{bmatrix}\begin{bmatrix} -1 & 3 & -3 & 1 \\ 2 & -5 & 4 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 0 \end{bmatrix}\begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \\ \mathbf{p}_4 \end{bmatrix}$$

and $\mathbf{p}_1$ to $\mathbf{p}_4$ are a set of control points with (in this case) x and y coordinates. This will generate a segment of curve which links the middle two points, i.e. $\mathbf{p}_2$ and $\mathbf{p}_3$.

The C-R spline creates a curve which interpolates the two mid-points, and whose gradient at each end is in the same direction as a vector joining the neighbouring points. Where two C-R segments join, the C-R has C1 continuity - that is parametric continuity in the first derivative. In contrast, the B-spline exhibits C2 continuity, but does not interpolate the control points, it only passes close to them. The B-spline is also contained in the convex hull of a polygon connecting all the control points, whereas this is not necessarily the case for a C-R spline. [25%]

(b) A closed loop can be formed from four segments of a C-R spline, where the control points are rotated each time, starting with $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4$ then $\mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4, \mathbf{p}_1$, etc. [5%]

(c) (i) In this case, the control points are exactly as shown in (a) above, with the geometry matrix:

$$\mathbf{G} = \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \\ \mathbf{p}_4 \end{bmatrix} = \begin{bmatrix} -b & 0 \\ 0 & a \\ b & 0 \\ 0 & -a \end{bmatrix}$$

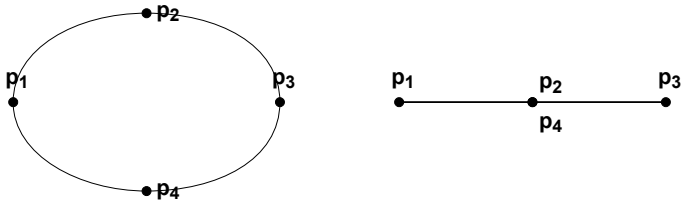
hence:

$$C(t) = \left\{-bt^3 + bt^2 + bt, at^3 - 2at^2 + a\right\}$$

so the end-points are:

$$C(0) = \{0, a\}, \quad C(1) = \{b, 0\}$$

which, as expected, are the points $\mathbf{p}_2$ and $\mathbf{p}_3$. [20%]



(c) (ii) A sketch of the curve is in the left hand diagram above. The gradients at the control points are either horizontal or vertical. [10%]

(c) (iii) A sketch of the curve for $a = 0$ is in the right hand diagram above. This is a completely flat loop from $\mathbf{p}_1$ to $\mathbf{p}_3$. [10%]

(c) (iv) In the middle of each of the four curve segments, the parametric and geometric continuity is guaranteed to be up to the third derivative, since these are cubic functions. Hence the only points of note are where the segments join, and there are only two different cases (due to symmetry), at $\mathbf{p}_2$ (similarly $\mathbf{p}_4$) and at $\mathbf{p}_3$ (similarly $\mathbf{p}_1$).

First looking at $\mathbf{p}_2$, by taking the $t = 0$ end of the curve from $\mathbf{p}_2$ to $\mathbf{p}_3$:

$$\begin{aligned} C_{2,3}(t) &= \left\{-bt^3 + bt^2 + bt, 0\right\} \\ C_{2,3}(0) &= \{0, 0\} \\ C'_{2,3}(0) &= \{b, 0\} \\ C''_{2,3}(0) &= \{2b, 0\} \end{aligned}$$

Now the $t = 1$ end of the curve from $\mathbf{p}_1$ to $\mathbf{p}_2$:

$$\begin{aligned} C_{1,2}(t) &= \left\{-bt^3 + 2bt^2 - b, 0\right\} \\ C_{1,2}(1) &= \{0, 0\} \\ C'_{1,2}(1) &= \{b, 0\} \\ C''_{1,2}(1) &= \{-2b, 0\} \end{aligned}$$

Hence there is parametric continuity here only up to C1, but clearly from the sketch the geometric continuity is up to at least G2.

Now looking at $\mathbf{p}_3$, by taking the $t = 1$ end of the curve from $\mathbf{p}_2$ to $\mathbf{p}_3$:

$$\begin{aligned} C_{2,3}(t) &= \left\{-bt^3 + bt^2 + bt, 0\right\} \\ C_{2,3}(1) &= \{b, 0\} \\ C'_{2,3}(1) &= \{0, 0\} \\ C''_{2,3}(1) &= \{-4b, 0\} \end{aligned}$$

Now the $t = 0$ end of the curve from $\mathbf{p}_3$ to $\mathbf{p}_4$:

$$\begin{aligned} C_{3,4}(t) &= \left\{bt^3 - 2bt^2 + b, 0\right\} \\ C_{3,4}(0) &= \{b, 0\} \\ C'_{3,4}(0) &= \{0, 0\} \\ C''_{3,4}(0) &= \{-4b, 0\} \end{aligned}$$

Hence there is parametric continuity here up to C2, but clearly from the sketch the geometric continuity is only up to G0. [30%]

Assessor's remarks: The comparison between Catmull-Rom and B-splines in the first part of the question was answered well, though several candidates omitted any general description of how splines are used. Part (c) was very well answered, though a few candidates laboured this by not using the given coordinates for the points. Sketches in (ii) and (iii) demonstrated a good understanding of the properties of the spline, with many candidates correctly noting that (iii) was just a straight line. (iv) was less well answered: some candidates made the maths far harder by not concentrating on the specific case ($a = 0$, $b > 0$), or failing to note that C1 continuity is guaranteed from the outset. Answers to this part also often lacked any commentary to the various equations, or realisation that continuity had to be proven at two of the segment joins.

4. **Illumination, reflection and shading**

(a) I is the intensity of the reflected light. I depends on several terms. First, there is the ambient reflection term, $I_a k_a$, which models indirect illumination of the surface. I_a is the intensity of the general background illumination, and k_a is the surface's ambient reflection coefficient. The next two terms in the model are calculated for a point light with intensity I_p. First there is the diffuse reflection term, $k_d \mathbf{L}.\mathbf{N}$, which models even reflection of the light source in all directions. $\mathbf{L}$ is the unit vector from the surface point towards the light source, $\mathbf{N}$ is the unit surface normal and k_d is the surface's diffuse reflection coefficient (small for dark surfaces, high for bright surfaces). Finally, there is the specular reflection term, $k_s \left(\mathbf{R}.\mathbf{V}\right)^n$, which models directional reflection of the light source along the unit mirror vector $\mathbf{R}$. $\mathbf{V}$ is the unit vector from the surface point towards the viewer. The viewer only perceives the specular highlight when looking along the mirror direction, or at least close to it. k_s is the surface's specular reflection coefficient (small for matte surfaces, high for shiny surfaces), and n is the specular exponent that determines the tightness of the glint. n is high for a tight highlight (e.g. a perfect mirror) and small for a more blurred highlight (e.g. aluminium). [25%]

(b) Both Gouraud and Phong shading work with *vertex normals*, which are found by averaging the normals of all polygons incident at a vertex. Gouraud shading proceeds by calculating an intensity at each vertex using the vertex normal and the Phong model. Intensities for interior pixels are found by bilinear interpolation. For efficiency, the interpolation can be formulated using fast, incremental calculations.

Phong shading interpolates the normals instead of the intensities. This tends to restore the original curvature of a surface, so that highlights can be reproduced accurately. The disadvantage of Phong shading is its expense. Even though the normals can be interpolated using incremental calculations, the interpolation considers the three components independently, so the vector must be renormalized at each pixel. Then, a *separate* intensity for each pixel is calculated using the Phong model.

Gouraud shading is comparatively fast, though it produces less photo-realistic renderings. It is particularly poor with the specular component. If a highlight should impinge on a polygon but not extend to its vertices, Gouraud shading will miss the highlight. [20%]

(c) The bilinear interpolation used in Gouraud shading produces a constant intensity gradient within each polygon. At polygon edges, therefore, there should be a step change in the intensity gradient, but not a step change in the intensity itself. That viewers do, nevertheless, perceive a step change in intensity, must be due to an idiosyncrasy of the human visual system triggered by gradient discontinuities (this is known as the *Mach band effect*). Intensity gradients are particularly strong around specular highlights, where the intensity can fall away rapidly on either side of an edge. [15%]

(d) (i) The light source must be at infinity for $\mathbf{L}$ to be the same at all three vertices. [5%]

(ii) Phong shading uses bilinear interpolation of the vertex normals at A, B and C to calculate the intensity at P. Since we are doing diffuse shading ($\mathbf{L}.\mathbf{N}$) only, and $\mathbf{L} = [\ 0\ \ 0\ \ 1\ ]^T$,

we need only consider the third element of the normal at P. The scan line containing P is exactly half way along the edges BA and CA, and P is half way along this scan line. The third element of the interpolated normal will be zero half way between C and A, and negative half way between B and A, and therefore negative at P. So **L**.**N** will be negative at P, indicating that this point on the surface is facing away from the light source. The rendered intensity will therefore be zero. We have assumed that the given vertex normals are those directed towards the viewer, as required by the Phong model. [25%]

(iii) Gouraud shading would give zero intensities at A and B, but a nonzero intensity at C, and therefore a nonzero intensity at P. So the Gouraud-shaded intensity would be higher than the Phong-shaded intensity at P. [10%]

Assessor's remarks: The more familiar parts of the question were very well answered by the vast majority of candidates, indicating a good grasp the basic principles. In (c), fewer than half of the candidates made a coherent case for the illusion originating in the human visual system, since there is no conceivable way that Gouraud shading could render the pixels as shown. In (d), a majority of candidates correctly calculated the negative scalar product for the intensity at P, but then failed to deduce that the surface must therefore be turned away from the light source and hence not illuminated.

Andrew Gee, Richard Prager & Graham Treece
January 2017