High-Level Vision (Artificial)

1. Below is shown a template T and an image I. Calculate the result of performing template matching on the image, and hence, suggest the location of the object depicted in the template assuming that there is exactly one such object in the image. Use the following similarity measures (a) normalised cross-correlation (b) sum of absolute differences.

$$T = \begin{bmatrix} 100 & 150 & 200 \\ 150 & 10 & 200 \\ 200 & 200 & 250 \end{bmatrix}, \qquad I = \begin{bmatrix} 60 & 50 & 40 & 40 \\ 150 & 100 & 100 & 80 \\ 50 & 20 & 200 & 80 \\ 200 & 150 & 150 & 50 \end{bmatrix}$$

(a) Normalised cross-correlation.

Similarity =

$$\frac{\sum_{i,j} T(i,j)I(i,j)}{\sqrt{\sum_{i,j} T(i,j)^2}\sqrt{\sum_{i,j} I(i,j)^2}}$$

$\sqrt{\sum_{i,j} T(i,j)^2} = \sqrt{100^2+150^2+200^2+150^2+10^2+200^2+200^2+200^2+250^2} = 526.9$

At pixel (2,2) Similarity =

$$\frac{100\times60+150\times50+200\times40+150\times150+10\times100+200\times100+200\times50+200\times20+250\times200}{526.9\sqrt{60^2+50^2+40^2+150^2+100^2+100^2+50^2+20^2+200^2}}$$

$= \frac{129000}{526.9\times305.1} = 0.80$

At pixel (3,2) Similarity =

$$\frac{100\times50+150\times100+200\times20+150\times40+10\times100+200\times200+200\times40+200\times80+250\times80}{526.9\sqrt{50^2+100^2+20^2+40^2+100^2+200^2+40^2+80^2+80^2}}$$

$= \frac{115000}{526.9\times280.9} = 0.78$

At pixel (2,3) Similarity =

$$\frac{100\times150+150\times50+200\times200+150\times100+10\times20+200\times150+200\times100+200\times200+250\times150}{526.9\sqrt{150^2+50^2+200^2+100^2+20^2+150^2+100^2+200^2+150^2}}$$

$= \frac{205200}{526.9\times412.8} = 0.94$

At pixel (3,3) Similarity =

$$\frac{100\times100+150\times20+200\times150+150\times100+10\times200+200\times150+200\times80+200\times80+250\times50}{526.9\sqrt{100^2+20^2+150^2+100^2+200^2+150^2+80^2+80^2+50^2}}$$

$= \frac{134500}{526.9\times347.4} = 0.73$

Hence, object at location (2,3).

(b) Sum of absolute differences.

Distance =

$$\sum_{i,j} abs(T(i,j) - I(i,j))$$

At pixel (2,2) Distance =

$\|100-60\|+\|150-50\|+\|200-40\|+\|150-150\|+\|10-100\|+\|200-100\|+\|200-50\|+\|200-20\|+\|250-200\|$

$= 870$

At pixel (3,2) Distance =

$\|100-50\|+\|150-40\|+\|200-40\|+\|150-100\|+\|10-100\|+\|200-80\|+\|200-20\|+\|200-200\|+\|250-80\|$

$= 930$

At pixel (2,3) Distance =

$\|100-150\|+\|150-100\|+\|200-100\|+\|150-50\|+\|10-20\|+\|200-200\|+\|200-200\|+\|200-150\|+\|250-150\|$

$= 460$

At pixel (3,3) Distance =

$\|100-100\|+\|150-100\|+\|200-80\|+\|150-20\|+\|10-200\|+\|200-80\|+\|200-150\|+\|200-150\|+\|250-50\|$

$= 910$

Hence, object at location (2,3).

2. A computer vision system uses template matching to perform object recognition. The system needs to detect 20 different objects each of which can be seen from 12 different viewpoints, each of which requires a different template. If an image is 300 by 200 pixels, and templates are 11 by 11 pixels, how many floating-point operations are required to process one image if cross-correlation is used as the similarity measure?

The system uses 20×12 *templates. Each template is matched to* $(300-10) \times (200-10)$ *image locations (assuming we don't allow the template to fall off the edge of the image). Hence, similarity is calculated* $20 \times 12 \times (300-10) \times (200-10) = 13,224,000$ *times when processing one image.*

For cross-correlation, each matching operation requires 11×11 *multiplications plus* $11 \times 11 - 1$ *additions, i.e.,* $(11 \times 11)^2 - 1 = 14640$ *floating-point operations.*

Therefore, processing one image requires $13,224,000 \times 14640 \approx 1.9 \times 10^{11}$ *floating point operations.*

3. Below are shown three binary templates T_1, T_2 and T_3 together with a patch I of a binary image. Determine which template best matches the image patch using the following similarity measures (a) cross-correlation, (b) normalised cross-correlation, (c) correlation coefficient, (d) sum of absolute differences.

$$T_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}, \quad T_2 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}, \quad T_3 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad I = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

(a) Cross-correlation.

Similarity =

$$\sum_{i,j} T(i,j) I(i,j)$$

For T_1 Similarity = 7
For T_2 Similarity = 5
For T_3 Similarity = 7
Both T_1 and T_3 match equally well.

(b) Normalised cross-correlation.

Similarity =

$$\frac{\sum_{i,j} T(i,j) I(i,j)}{\sqrt{\sum_{i,j} T(i,j)^2}\sqrt{\sum_{i,j} I(i,j)^2}}$$

For T_1 Similarity = $\frac{7}{\sqrt{7}\times\sqrt{7}} = 1$
For T_2 Similarity = $\frac{5}{\sqrt{5}\times\sqrt{7}} = 0.85$
For T_3 Similarity = $\frac{7}{\sqrt{8}\times\sqrt{7}} = 0.94$
T_1 is the best match.

(c) Correlation coefficient.

Similarity =

$$\frac{\sum_{i,j}(T(i,j)-\bar{T})(I(i,j)-\bar{I})}{\sqrt{\sum_{i,j}(T(i,j)-\bar{T})^2}\sqrt{\sum_{i,j}(I(i,j)-\bar{I})^2}}$$

$$T_1 - \bar{T}_1 = \begin{bmatrix} 0.22 & 0.22 & 0.22 \\ 0.22 & -0.78 & -0.78 \\ 0.22 & 0.22 & 0.22 \end{bmatrix}, \quad T_2 - \bar{T}_2 = \begin{bmatrix} 0.44 & -0.56 & -0.56 \\ 0.44 & -0.56 & -0.56 \\ 0.44 & 0.44 & 0.44 \end{bmatrix},$$

$$T_3 - \bar{T}_3 = \begin{bmatrix} 0.11 & 0.11 & 0.11 \\ 0.11 & -0.89 & 0.11 \\ 0.11 & 0.11 & 0.11 \end{bmatrix}, \quad I - \bar{I} = \begin{bmatrix} 0.22 & 0.22 & 0.22 \\ 0.22 & -0.78 & -0.78 \\ 0.22 & 0.22 & 0.22 \end{bmatrix}$$

For T_1 *Similarity* $= \frac{7(0.22^2)+2(-0.78^2)}{\sqrt{(7(0.22^2)+2(-0.78^2))}\sqrt{(7(0.22^2)+2(-0.78^2))}} = 1$

For T_2 *Similarity* $= \frac{5(0.44\times 0.22)+2(-0.56\times 0.22)+2(-0.56\times -0.78)}{\sqrt{(5(0.44^2)+4(-0.56^2))}\sqrt{(7(0.22^2)+2(-0.78^2))}} = \frac{1.111}{1.491\times 1.247} = 0.60$

For T_3 *Similarity* $= \frac{7(0.11\times 0.22)+(0.11\times -0.78)+(-0.89\times -0.78)}{\sqrt{(8(0.11^2)+(-0.89^2))}\sqrt{(7(0.22^2)+2(-0.78^2))}} = \frac{0.777}{0.943\times 1.247} = 0.66$

T_1 *is the best match.*

(d) Sum of absolute differences
Distance =

$$\sum_{i,j} \| T(i,j) - I(i,j) \|$$

For T_1 *Distance* $= 7(1-1)+2(0-0) = 0$
For T_2 *Distance* $= 5(1-1)+2(1-0)+2(0-0) = 2$
For T_3 *Distance* $= 7(1-1)+1(1-0)+1(0-0) = 1$
T_1 *is the best match.*

4. Below is shown an edge template T and a binary image I which has been pre-processed to extract edges. Calculate the result of performing edge matching on the image, and hence, suggest the location of the object depicted in the edge template assuming that there is exactly one such object in the image. Calculate the distance between the template and the image as the average of the minimum distances between points on the edge template (T) and points on the edge image (I).

$$T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \qquad I = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

At pixel (2,2) Distance =

$$\frac{1}{8}[1+1+1+0+0+1+1+0] = 0.625$$

At pixel (3,2) Distance =

$$\frac{1}{8}\left[1+1+\sqrt{2}+0+1+1+0+1\right] = 0.802$$

At pixel (2,3) Distance =

$$\frac{1}{8}[0+0+0+1+0+0+0+0] = 0.125$$

At pixel (3,3) Distance =

$$\frac{1}{8}[0+0+1+1+1+0+0+1] = 0.5$$

Hence, object at location (2,3).

5. One method of object recognition is comparison of intensity histograms. Briefly describe two advantages and two disadvantages of this method.

Advantages:

- *Fast*
- *Unaffected by viewpoint changes*

Disadvantages:

- *Sensitive to illumination changes*
- *Insensitive to spatial configuration changes*

6. In a very simple feature-matching object recognition system each keypoint has x,y-coordinates and a 3 element feature vector. Two training images, one of object A and the other of object B, have been processed to create a database of

known objects, as shown below:

Object	Keypoint Number	Coordinates (pixels)	Feature Vector
A	**A1**	**(20,5)**	**(1,6,10)**
	A2	**(10,40)**	**(7,8,15)**
	A3	**(40,25)**	**(2,9,3)**
B	**B1**	**(20,10)**	**(6,1,12)**
	B2	**(30,5)**	**(13,4,8)**
	B3	**(30,45)**	**(3,8,4)**

The keypoints and feature vectors extracted from a new image are as follows:

Keypoint Number	Coordinates (pixels)	Feature Vector
N1	**(16,50)**	**(5,8,15)**
N2	**(25,14)**	**(2,6,11)**
N3	**(30,31)**	**(12,3,8)**
N4	**(40,45)**	**(5,2,11)**
N5	**(44,34)**	**(2,8,3)**

Perform feature matching using the sum of absolute differences as the distance measure and applying the following criterion for accepting a match: that the ratio of distance to first nearest descriptor to that of second is less than 0.4.

It is known that objects in different images are related by a pure translation in the image plane. Hence, use the RANSAC algorithm to assess the consistency of the matched points and so determine which of the two training objects is present in the new image. Apply RANSAC exhaustively to all matches, rather than to a subset of matches chosen at random and assume the threshold for comparing the model's prediction with the data is 3 pixels.

SAD distance between each keypoint in the training image database and each keypoint in the new image:

Feature Vectors	*N1(5,8,15)*	*N2(2,6,11)*	*N3(12,3,8)*	*N4(5,2,11)*	*N5(2,8,3)*
A1(1,6,10)	*11*	*2*	*16*	*9*	*10*
A2(7,8,15)	*2*	*11*	*17*	*12*	*17*
A3(2,9,3)	*16*	*11*	*21*	*18*	*1*
B1(6,1,12)	*11*	*10*	*12*	*3*	*20*
B2(13,4,8)	*19*	*16*	*2*	*13*	*20*
B3(3,8,4)	*13*	*10*	*18*	*15*	*2*
Ratio 1st to 2nd best:	*2/11 = 0.18*	*2/10 =0.2*	*2/12=0.17*	*3/9=0.33*	*1/2=0.5*

Hence, the following matches are found:

Keypoints	*Coordinates*
N1 - A2	*(16,50) - (10,40)*
N2 - A1	*(25,14) - (20,5)*
N3 - B2	*(30,31) - (30,5)*
N4 - B1	*(40,45) - (20,10)*

Applying RANSAC.

Choose N1. Model is a translation of (16,50)-(10,40) = (6,10)
Locations matching points predicted by the model are:
For N2: (25,14)-(6,10)=(19,4)
actual match is at (20,5), hence, this is an inlier for this model.
Hence, consensus set =1.
Choosing N2, will also predict a translation consistent with N1.

Choose N3. Model is a translation of (30,31)-(30,5) = (0,26)
Locations matching points predicted by the model are:
For N4: (40,45)- (0,26) = (40,19)
actual match is at (20,10), hence, this is an outlier for this model.
Hence, consensus set =0.
Choosing N4, will predict a translation inconsistent with N3.
The new image contains object A, as this is the only object for which the locations of the matching keypoints are consistent.

7. In a simple bag-of-words object recognition system images are represented by histograms showing the number of occurrences of 10 "codewords". The number of occurrences of the codewords in three training images are given below:
ObjectA = (2,0,0,5,1,0,0,0,3,1)

ObjectB = (0,0,1,2,0,3,1,0,1,0)
ObjectC = (1,1,2,0,0,1,0,3,1,1)
A new image is encoded as follows:
New = (2,1,1,0,1,1,0,2,0,1)
Determine the training image that best matches the new image by finding the cosine of the angle between the codeword vectors.

Similarity $= cos(\theta) = \frac{\sum_i A(i)N(i)}{\sqrt{\sum_i A(i)^2}\sqrt{\sum_i N(i)^2}}$ *(i.e., the normalised cross-correlation).*
Similarity between New image and ObjectA is:

$$\frac{(2\times 2)+(1\times 0)+(1\times 0)+(0\times 5)+(1\times 1)+(1\times 0)+(0\times 0)+(2\times 0)+(0\times 3)+(1\times 1)}{\sqrt{2^2+1^2+1^2+0^2+1^2+1^2+0^2+2^2+0^2+1^2}\sqrt{2^2+0^2+0^2+5^2+1^2+0^2+0^2+0^2+3^2+1^2}}$$
$$=\frac{6}{3.6\times 6.3}=0.26$$

Similarity between New image and ObjectB is:

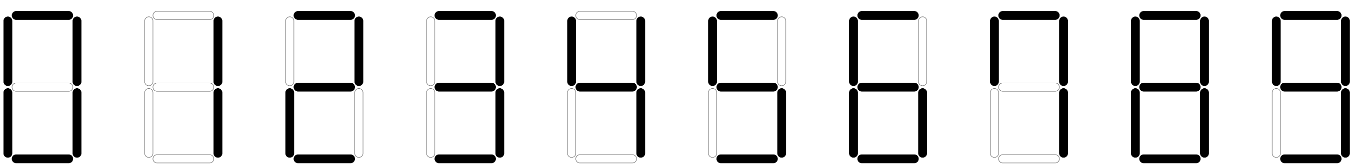
$$\frac{(2\times 0)+(1\times 0)+(1\times 1)+(0\times 2)+(1\times 0)+(1\times 3)+(0\times 1)+(2\times 0)+(0\times 1)+(1\times 0)}{\sqrt{2^2+1^2+1^2+0^2+1^2+1^2+0^2+2^2+0^2+1^2}\sqrt{0^2+0^2+1^2+2^2+0^2+3^2+1^2+0^2+1^2+0^2}}$$
$$=\frac{4}{3.6\times 4}=0.28$$

Similarity between New image and ObjectC is:

$$\frac{(2\times 1)+(1\times 1)+(1\times 2)+(0\times 0)+(1\times 0)+(1\times 1)+(0\times 0)+(2\times 3)+(0\times 1)+(1\times 1)}{\sqrt{2^2+1^2+1^2+0^2+1^2+1^2+0^2+2^2+0^2+1^2}\sqrt{1^2+1^2+2^2+0^2+0^2+1^2+0^2+3^2+1^2+1^2}}$$
$$=\frac{13}{3.6\times 4.2}=0.86$$

Hence, the new image is most similar to ObjectC.

8. A computer vision system is to be developed the can read digits from an 7-segment LCD display (like that on a standard calculator). On such a display, the numbers 0 to 9 are generated by turning on specific combinations of segments, as shown below.



A simple bag-of-words object recognition system is to be used. The codeword dictionary consists of two features: (1) a vertical line, and (2) a horizontal line.

(a) How would the digits 0 to 9 be encoded?

(b) Does this system succeed in recognising all 10 digits?

(c) Suggest an alternative object recognition method that might work better?

(d) If the camera capturing images of the LCD display gets rotated 180 degrees around the optical axis, what effect does this have on the bag-of-words solution and your alternative method?

(e) If the LCD display shows multiple digits simultaneously, what effect does this have on the bag-of-words solution and your alternative method?

(a) encoding = (vertical,horizontal)
0= (4,2)
1= (2,0)
2= (2,3)

3= (2,3)
4= (3,1)
5= (2,3)
6= (3,3)
7= (3,1)
8= (4,3)
9= (3,3)

(b) No. Digits 2, 3, and 5 all have the same encoding, as do digits 4 and 7 and digits 6 and 9. These digits cannot be told apart.

(c) Template matching.

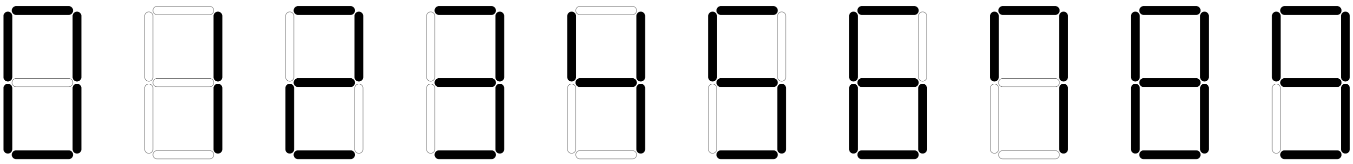
(d) If the image is rotated 180 degrees, this:

- *has no effect on the bag-of-words solution (inverted input images will be encoded in the same way as upright training images, but the system still fails for the reason given in part (b)).*
- *has a large effect on template matching as most templates will fail to match the upside down input, and some templates (such as those for 2 and 5) will match the wrong inputs.*

(e) If the image contains multiple digits, this:

- *makes the bag-of-words solution even more infeasible, as the encoding of the input image will contain features from separate objects.*
- *has no effect on template matching, except that multiple matches may now be found.*

9. A computer vision system is to be developed the can read digits from an 7-segment LCD display (like that on a standard calculator). On such a display, the numbers 0 to 9 are generated by turning on specific combinations of segments, as shown below.



A simple bag-of-words object recognition system is to be used. The SIFT feature detector has been used to locate features in the 10 training digits in order to create a codeword dictionary. Due to the rotation invariance of the SIFT descriptor only three distinct features are identified: (1) an "L" shaped corner (at any orientation), (2) a "T" shaped corner (at any orientation), (3) a line termination, or end point, (at any orientation).

(a) How would the digits 0 to 9 be encoded?

(b) Does this system succeed in recognising all 10 digits?

(a) encoding = (L,T,end)
0= (4,0,0)
1= (0,0,2)
2= (4,0,2)
3= (2,1,3)
4= (1,1,3)
5= (4,0,2)
6= (4,1,1)
7= (2,0,2)
8= (4,2,0)
9= (4,1,1)

(b) No. Digits 2 and 5 have the same encoding, as do digits 6 and 9. These digits cannot be told apart.

10. Projective geometry does not preserve distances or angles. However, the cross-ratio (which is a ratio of ratios of distances) is preserved. Given four collinear points p_1, p_2, p_3, and p_4, the cross-ratio is defined as:

$$Cr(p_1, p_2, p_3, p_4) = \frac{\Delta_{13}\Delta_{24}}{\Delta_{14}\Delta_{23}}$$

Where Δ_{ij} is the distance between two points p_i and p_j.

Four co-linear points are at the following 3D coordinates relative to the camera reference frame: $p_1 = [40, -40, 400]$, $p_2 = [23.3, -6.7, 483.3]$, $p_3 = [15, 10, 525]$, $p_4 = [-10, 60, 650]$.
(a) Calculate the cross-ratio for these points in 3D space.
(b) Calculate the cross-ratio for these points in the image seen by the camera. The image principal point is at coordinates [244,180] pixels, and the magnification factors in the x and y directions are 925 and 740. Assume that the camera does not suffer from skew or any other defect.

(a)

$$\Delta_{13} = \sqrt{(40-15)^2 + (-40-10)^2 + (400-525)^2} = 136.9$$
$$\Delta_{24} = \sqrt{(23.3-(-10))^2 + (-6.7-60)^2 + (483.3-650)^2} = 182.6$$
$$\Delta_{14} = \sqrt{(40-(-10))^2 + (-40-60)^2 + (400-650)^2} = 273.9$$
$$\Delta_{23} = \sqrt{(23.3-15)^2 + (-6.7-10)^2 + (483.3-525)^2} = 45.7$$

$$Cr(p_1, p_2, p_3, p_4) = \frac{\Delta_{13}\Delta_{24}}{\Delta_{14}\Delta_{23}} = \frac{136.9 \times 182.6}{273.9 \times 45.7} = 2$$

(b)

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \frac{1}{z}\begin{pmatrix} \alpha & 0 & o_x \\ 0 & \beta & o_y \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \frac{1}{z}\begin{pmatrix} 925 & 0 & 244 \\ 0 & 740 & 180 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \frac{1}{z}\begin{pmatrix} 925 & 0 & 244 & 0 \\ 0 & 740 & 180 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} u_1 \\ v_1 \\ 1 \end{pmatrix} = \frac{1}{400}\begin{pmatrix} 925 & 0 & 244 & 0 \\ 0 & 740 & 180 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 40 \\ -40 \\ 400 \\ 1 \end{pmatrix} = \begin{pmatrix} 429 \\ 32 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} u_2 \\ v_2 \\ 1 \end{pmatrix} = \frac{1}{483.3}\begin{pmatrix} 925 & 0 & 244 & 0 \\ 0 & 740 & 180 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 23.3 \\ -6.7 \\ 483.3 \\ 1 \end{pmatrix} = \begin{pmatrix} 333.2 \\ 159.5 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} u_3 \\ v_3 \\ 1 \end{pmatrix} = \frac{1}{525}\begin{pmatrix} 925 & 0 & 244 & 0 \\ 0 & 740 & 180 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 15 \\ 10 \\ 525 \\ 1 \end{pmatrix} = \begin{pmatrix} 296.9 \\ 208.2 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} u_4 \\ v_4 \\ 1 \end{pmatrix} = \frac{1}{650}\begin{pmatrix} 925 & 0 & 244 & 0 \\ 0 & 740 & 180 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} -10 \\ 60 \\ 650 \\ 1 \end{pmatrix} = \begin{pmatrix} 215.5 \\ 316.6 \\ 1 \end{pmatrix}$$

$$\Delta_{13} = \sqrt{(429-296.9)^2 + (32-208.2)^2} = 220.2$$
$$\Delta_{24} = \sqrt{(333.2-215.5)^2 + (159.5-316.6)^2} = 196.3$$
$$\Delta_{14} = \sqrt{(429-215.5)^2 + (32-316.6)^2} = 355.8$$
$$\Delta_{23} = \sqrt{(333.2-296.9)^2 + (159.5-208.2)^2} = 60.8$$

$$Cr(p_1, p_2, p_3, p_4) = \frac{\Delta_{13}\Delta_{24}}{\Delta_{14}\Delta_{23}} = \frac{220.2 \times 196.3}{355.8 \times 60.8} = 2$$