Relative Positioning from Model Indexing

Stefan Carlsson
Computational Vision and Active Perception Laboratory (CVAP)*
Royal Institute of Technology (KTH), Stockholm, Sweden

Abstract

We show how to determine the position of a camera relative to a model of the environment using features extracted from one image. The method is based on view variation of cross-ratios and is independent of camera calibration.

1 Model Indexing From Images

Recognition and pose estimation of objects in images is in general based on extracting features such as points and lines from the image and finding corresponding features in model objects. The combinatorial complexity of matching image descriptors to models can be considerable. A way to reduce this is to use index tables where image descriptors are used as indexes to a table containing corresponding model descriptors. These tables can be constructed off-line which means that space is traded off for time [1] [2] [3] [4].

Another factor affecting the complexity of recognition is the fact that the image of an object depends on the viewpoint of the camera. This has initiated work in finding feature descriptors invariant to viewpoint [1] [2]. For configurations of points and lines invariants can in general be found only when they are contained in a planar surface. For general configurations of points and lines, any descriptor computed from image data will vary with viewpoint. [3] [5] . Since camera viewpoint in 3D is parameterized by 6 parameters, variation of image descriptors will in general be 6-dimensional. It is however possible to choose descriptors that have far less dimensionality of variation than 6. In the case when imaging is approximated by an affine mapping, it has been shown [3] [4] that it is possible to compute 4 image descriptors from 4 points that exhibit only 2-dimensional variation of viewpoint. For a specific configuration of 4 points, the 4-D image descriptor will be contained in a 2-D surface.

In this work we will consider the more general case of imaging by perspective mapping. Since one of the main applications we have in mind is moving platform navigation, perspective effects can be substantial. In projective and perspective transformations the cross ratio is a fundamental invariant of point and line sets. We will show that for 6 points in an image, 4 cross ratios can be computed. These 4 cross ratios will exhibit 3-dimensional variation with viewpoint. This is a very simple result due to the fact that the cross ratios are invariant to rotations of the camera around the point of projection. The variation is therefore due entirely to camera translation. In the limit of large relative viewing distances the image mapping will become parallel and the dependence of image data on camera position

*Address: NADA, KTH, S-100 44 Stockholm, Sweden
Email: stefanc@bion.kth.se

BMVC 1993 doi:10.5244/C.7.12

will reduce to that of the direction to the camera described by 2 parameters, The 4 cross ratios will therefore asymptotically be contained in a 2-D surface.

An interesting aspect of using cross ratios for indexing is the fact that they are independent of camera calibration. This is assuming that the imaging is linear. Cross ratios will be related to camera position in a non linear way. When image-model feature correspondence has been established, this relation can be inverted and the camera position relative to the model can be found in a calibration independent way. Due to the complexity of non linear inversion this is implemented as a table lookup where cross ratios are used to index directly into a table of camera positions.

In an application we will consider the problem of finding the position of a camera relative to a model of the environment. By considering camera motion restricted to a planar ground, the variation of cross ratios on viewpoint will be 2-dimensional even in the perspective mapping case, and 1-dimensional in the affine limit. By using only the horizontal coordinates of vertical edge segments, cross ratios can be computed from 4 points. This fact has previously been exploited for navigation. [6]

In our case, we can use 3 cross ratios computed from 6 points as entries to the index table. This will reduce the complexity of the index tables considerably.

2 View Variation of Cross Ratios

Since a general viewpoint in R^3 has 6 degrees of freedom, the $2n$ image coordinates of a set of n points will be confined to 6-dimensional surface in R^{2n}. Having a 6-parameter variation is definitely too large to be manageable for recognition and pose estimation using indexing. We therefore consider the cross ratio of a group of 5 points arbitrarily placed in R^3.

A cross-ratio of a set of 5 points in R^3 with image coordinates $(x_1, y_1) \ldots (x_5, y_5)$ can be defined as:

$$\sigma = \frac{\begin{vmatrix} x_1 & x_2 & x_5 \\ y_1 & y_2 & y_5 \\ 1 & 1 & 1 \end{vmatrix} \begin{vmatrix} x_3 & x_4 & x_5 \\ y_3 & y_4 & y_5 \\ 1 & 1 & 1 \end{vmatrix}}{\begin{vmatrix} x_1 & x_3 & x_5 \\ y_1 & y_3 & y_5 \\ 1 & 1 & 1 \end{vmatrix} \begin{vmatrix} x_2 & x_4 & x_5 \\ y_2 & y_4 & y_5 \\ 1 & 1 & 1 \end{vmatrix}} \tag{1}$$

If all points are coplanar in R^3 a change of viewpoint will correspond to a projective transform of the image coordinates. The projective transform is linear in the homogeneous coordinates $(x_i, y_i, 1)$ and the cross ratio will therefore be invariant. If the points are not all coplanar, a change of viewpoint will no longer correspond to a projective transformation. The cross ratio will therefore vary with the viewpoint.

The image coordinates are related to the position of a point in a camera centered coordinate system as:

$$x_i = \frac{X_i^C}{Z_i^C} \qquad y_i = \frac{Y_i^C}{Z_i^C} \tag{2}$$

The coordinates of the points in a world centered reference frame are denoted as $(X_1, Y_1, Z_1) \ldots (X_5, Y_5, Z_5)$ and the coordinates of the projection point of the camera is (X_c, Y_c, Z_c) This point is taken as the origin of the camera centered

system. The coordinates in the camera centered system are related to those of the world centered system by a linear transformation:

$$\begin{pmatrix} X_i^C \\ Y_i^C \\ Z_i^C \end{pmatrix} = \begin{pmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{pmatrix} \begin{pmatrix} X_i - X_c \\ Y_i - Y_c \\ Z_i - Z_c \end{pmatrix} = PR \begin{pmatrix} X_i - X_c \\ Y_i - Y_c \\ Z_i - Z_c \end{pmatrix} \tag{3}$$

This transformation can be factored into camera rotation R and projection matrix P The projection matrix contains scale parameters for image coordinates and other parameters that can in principle be obtained by calibration. By using cross ratios as image descriptors however, both the rotation and projection matrix will cancel out and there will be no need for camera calibration. Using eq. 2 and 3 in the expression 1 for the cross ratio we get:

$$\begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} = \frac{1}{Z_i^C} \begin{pmatrix} X_i^C \\ Y_i^C \\ Z_i^C \end{pmatrix} = \frac{1}{Z_i^C} PR \begin{pmatrix} X_i - X_c \\ Y_i - Y_c \\ Z_i - Z_c \end{pmatrix} \tag{4}$$

$$\sigma = \frac{\begin{vmatrix} X_1 - X_c & X_2 - X_c & X_5 - X_c \\ Y_1 - Y_c & Y_2 - Y_c & Y_5 - Y_c \\ Z_1 - Z_c & Z_2 - Z_c & Z_5 - Z_c \end{vmatrix} \begin{vmatrix} X_3 - X_c & X_4 - X_c & X_5 - X_c \\ Y_3 - Y_c & Y_4 - Y_c & Y_5 - Y_c \\ Z_3 - Z_c & Z_4 - Z_c & Z_5 - Z_c \end{vmatrix}}{\begin{vmatrix} X_1 - X_c & X_3 - X_c & X_5 - X_c \\ Y_1 - Y_c & Y_3 - Y_c & Y_5 - Y_c \\ Z_1 - Z_c & Z_3 - Z_c & Z_5 - Z_c \end{vmatrix} \begin{vmatrix} X_2 - X_c & X_4 - X_c & X_5 - X_c \\ Y_2 - Y_c & Y_4 - Y_c & Y_5 - Y_c \\ Z_2 - Z_c & Z_4 - Z_c & Z_5 - Z_c \end{vmatrix}} \tag{5}$$

where the factors $1/Z_i^C$ and determinant of the linear transformation matrix PR are cancelled out.

We see that the cross ratio computed from the image coordinates of 5 points in general position in R^3 is the ratio of two polynomials in the camera coordinates (X_c, Y_c, Z_c) and totally independent of camera rotation. The set of cross ratios $\sigma_1 \ldots \sigma_k$ computed from a set of points are therefore in general contained in a 3-dimensional surface in k-space. This 3-dimensional surface is a characteristic of the point set, independent of viewpoint. Using cross ratios we have therefore reduced the parametric variation on viewpoint from 6 to 3.

In the limit of parallel projection this variation is reduced even further, from 3 to 2. The determinants in eq. 5 can be expanded as:

$$\begin{vmatrix} X_1 - X_c & X_2 - X_c & X_5 - X_c \\ Y_1 - Y_c & Y_2 - Y_c & Y_5 - Y_c \\ Z_1 - Z_c & Z_2 - Z_c & Z_5 - Z_c \end{vmatrix} =$$

$$= \begin{vmatrix} X_1 & X_2 & X_5 \\ Y_1 & Y_2 & Y_5 \\ Z_1 & Z_2 & Z_5 \end{vmatrix} - \begin{vmatrix} X_c & X_c & X_c \\ Y_1 & Y_2 & Y_5 \\ Z_1 & Z_2 & Z_5 \end{vmatrix} - \begin{vmatrix} X_1 & X_2 & X_5 \\ Y_c & Y_c & Y_c \\ Z_1 & Z_2 & Z_5 \end{vmatrix} - \begin{vmatrix} X_1 & X_2 & X_5 \\ Y_1 & Y_2 & Y_5 \\ Z_c & Z_c & Z_c \end{vmatrix}$$

etc.

All determinants in eq. 5 can therefore be expressed in X_c, Y_c, Z_c as:

$$|\ldots| = A_n X_c + B_n Y_c + C_n Z_c + D_n \qquad n = 1 \ldots 4 \tag{6}$$

and the cross ratio:

$$\sigma = \frac{(A_1 X_c + B_1 Y_c + C_1 Z_c + D_1)(A_2 X_c + B_2 Y_c + C_2 Z_c + D_2)}{(A_3 X_c + B_3 Y_c + C_3 Z_c + D_3)(A_4 X_c + B_4 Y_c + C_4 Z_c + D_4)} \tag{7}$$

If we fix the coordinates of the points in the world centered system and introduce spherical coordinates for the camera position (X_c, Y_c, Z_c), we get:

$$X_c = r\, f_x(\theta,\phi) \quad Y_c = r\, f_y(\theta,\phi) \quad Z_c = r\, f_z(\theta,\phi) \tag{8}$$

Inserting this into eq. 7 and taking the limit of large observation distance r we get:

$$\lim_{r\to\infty} \sigma = \frac{(A_1 f_x(\theta,\phi) + B_1 f_y(\theta,\phi) + C_1 f_z(\theta,\phi))(A_2 f_x(\theta,\phi) + B_2 f_y(\theta,\phi) + C_2 f_z(\theta,\phi))}{(A_3 f_x(\theta,\phi) + B_3 f_y(\theta,\phi) + C_3 f_z(\theta,\phi))(A_4 f_x(\theta,\phi) + B_4 f_y(\theta,\phi) + C_4 f_z(\theta,\phi))}$$

The variation of the cross ratio with viewpoint is therefore reduced to variation with respect to the two parameters θ and ϕ, i.e. the angular position of the camera relative to the world reference frame. This result should be compared with that of [3] where it is shown that assuming para perspective projection, it is possible to choose 4 image descriptors from 4 points that will have 2-dimensional variation with viewpoint. The price for generalizing to perspective projection is that we need 6 points instead of 4.

3 Feature Indexing and Pose Estimation for a Camera on Planar Ground

3.1 Construction of Index Tables and Feature Matching

The cross ratios computed from image data can be used as indexes to a table containing model features. In order to be efficient these tables must not be too large. The tables are quantized versions of the index space with each axis representing an index feature. A critical factor is of course the dimensionality of the table. For points in general position in R^3 and perspective projection we would in general need a 4-D index space with each axis representing a cross ratio. Various 6 point configurations will be represented as 3-D surfaces in this 4-D index space. In the limit when parallel projection is valid we can consider 2-D surfaces in the 4-D index space.

The dimensionality of the index space can also be reduced by restricting the camera positions. This will be the case e.g. when a camera is constrained to move on a planar surface. Our problem is to find the position relative to a model of the environment, of a camera moving on a planar ground floor. In this case perspective effects cannot in general be neglected, since the extent of the object, in this case the spatial environment, will be of the same order of magnitude as the distance to the object features. The fact that we are moving on a planar ground floor means however that only two position parameters X_c, Y_c need to be considered. Features groups in the form of cross ratios can therefore be represented as a 2-D surface and the dimensionality of the index space can be taken to be 3. As features we will consider horizontal coordinates v_i of vertical lines in the image. Cross ratios can then be computed from groups of 4 coordinates $x_1 \ldots x_4$ and from eq. 5 we see that they will vary with camera position as:

$$\sigma = \frac{\begin{vmatrix} x_1 & x_2 \\ 1 & 1 \end{vmatrix} \begin{vmatrix} x_3 & x_4 \\ 1 & 1 \end{vmatrix}}{\begin{vmatrix} x_1 & x_3 \\ 1 & 1 \end{vmatrix} \begin{vmatrix} x_2 & x_4 \\ 1 & 1 \end{vmatrix}} = \frac{\begin{vmatrix} X_1 - X_c & X_2 - X_c \\ Y_1 - Y_c & Y_2 - Y_c \end{vmatrix} \begin{vmatrix} X_3 - X_c & X_4 - X_c \\ Y_3 - Y_c & Y_4 - Y_c \end{vmatrix}}{\begin{vmatrix} X_1 - X_c & X_3 - X_c \\ Y_1 - Y_c & Y_3 - Y_c \end{vmatrix} \begin{vmatrix} X_2 - X_c & X_4 - X_c \\ Y_2 - Y_c & Y_4 - Y_c \end{vmatrix}}$$

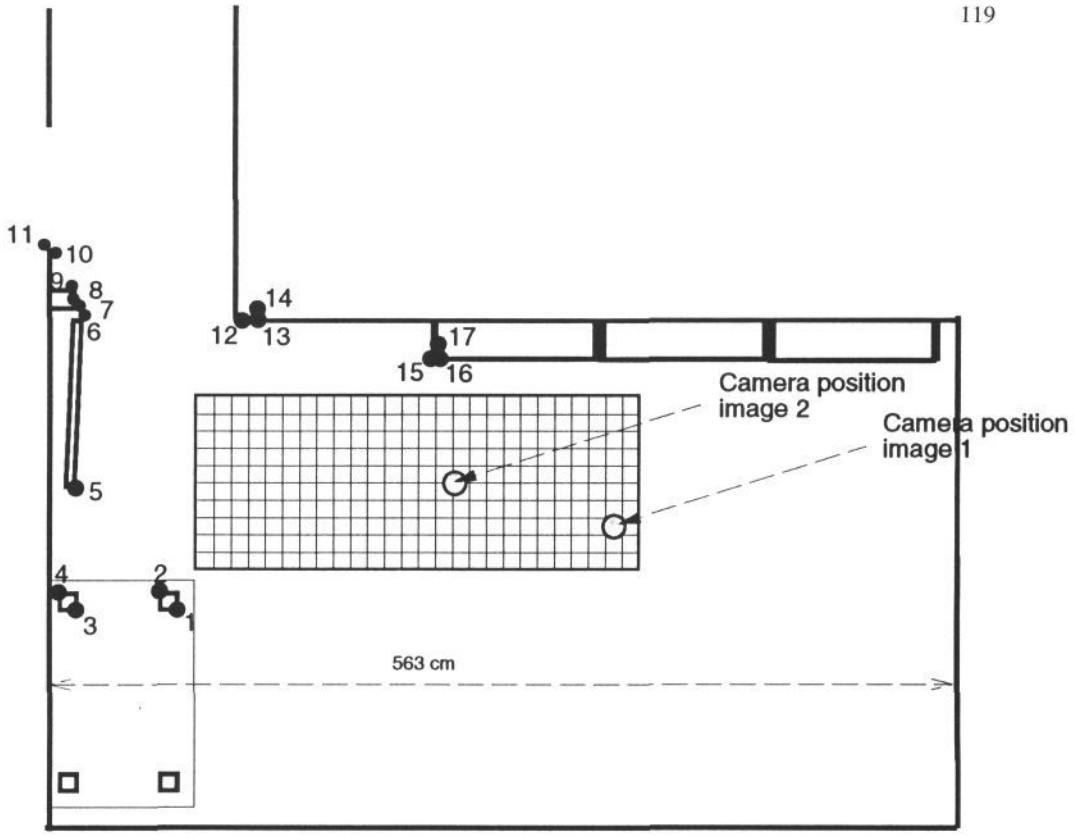


Figure 1: Map of the model environment with model features. The area and sampling grid for computing triple cross ratio indexes is indicated and the two camera positions for testing position estimation are marked by the shaded circles

The model in our case is a list of X, Y coordinates of features in the environment that could give rise to vertical lines in the image. The environment is part of a room with a door opening into a corridoor. A map of the room with 17 model features is shown in fig 1

The index space was constructed in the following way:

1. For X_c, Y_c camera positions chosen on a grid with 10 cm spacing within the rectangle in the map fig. 1 and all combinations of 6 features in the model, ordered clockwise as seen from the camera position, 3 cross ratios were computed by selection od 3 sets of 4 points.
2. The 3 cross ratios were used as indexes to a 3-dimensional index table. The table was divided into bins chosen as cubes with sides 0.02. Whenever a cross ratio triplet indexed a bin of the table, the 6 model features $j_1 \ldots j_6$ together with the position X_c, Y_c were recorded in that bin.

The indexing for feature matching works as follows:

1. All combinations of 6 horizontal coordinates of extracted vertical lines $i_1 \ldots i_6$ are used to compute cross ratio triplets.
2. The cross ratio triplets are used to index into the table. For each model group of 6 features $j_1 \ldots j_6$ found in the bin, an association table of image and model features is incremented one step at positions $(i_1, j_1) \ldots (i_6, j_6)$

3.2 Position Estimation

Pose estimation in the general 3-D case involves determining the 6 parameters of rotation and translation describing camera position. This requires at least 3 image points and their corresponding 3 model points. Methods have been presented based on selecting point triplets from image and model and for each triplet pair compute a tentative pose. Correct triplet pairs will then give clusters in pose parameter space. [7] Since this method is based on selecting groups of features from both image and model the complexity will in general be higher than using indexing where selection is only of features in the image. In the case where a priori information about possible image model matchings is available the complexity of this method can be reduced.

Given matched image and model features from indexing, pose estimation is in general a non linear over determined problem provided more than 3 features have been matched. With a calibrated camera both rotation and translation parameters of the camera can be obtained. In the case of no calibration information we can still obtain camera position parameters X_c, Y_c, Z_c by using the relation between cross ratios and camera position of eq. 5 From eq. 7 we see that given cross ratios and model coordinates, the position parameters are related by a second order polynomial equation. Solving these equations is a non-trivial matter. In order to avoid complicated non-linear equation solving the index tables were equipped with the information about position corresponding to each triple cross ratio entry. Every triplet cross ratio therefore indexes a number of positions corresponding to those from which this triplet was computed from model data. Using only index table positions corresponding to correctly matched features the indexed positions can be accumulated in a matrix and large values of this matrix should correspond to probable camera positions.

Note that camera rotation is not obtained by this method but requires some form of calibration of the camera.

4 Experimental Results

In order to test model indexing and pose estimation the camera was placed in two different positions indicated in fig. 1. For each position vertical edge segments were extracted by a very simple and robust procedure based on averaging image intensity in the vertical direction for each horizontal image coordinate. In order to cope with slight variations from ideally vertical edges, the image was divided into 7 horizontal stripes. In each stripe vertical edges were extracted independently and subsequently linked across stripes based on horizontal proximity.

Figs. 2 and 3 show extracted edge segments and resulting number of associations between image and model features using the indexing method described earlier. The size of the number at position i, j of this table can be seen as a measure of support for the hypothesis that image feature i should be matched to

model feature j. Finding the best association for each image feature will in general be combinatorially very complex. By simple thresholding several hypothetical matches can be discarded at the outset however. From figs. 2 and 3 it can be seen that for most image features the correct model feature dominates the row in the association table. Another factor reducing the complexity of finding associations is the fact that feature ordering from left to right will in general be invariant over all positions. This imposes constraints on possible pairwise matchings of image and model features. The main factors determining the complexity of finding correct associations will be the presence of clutter lines that are not modelled and the disappearance of modelled lines from the image. Work is presently going on in testing the robustness of the method against extra clutter and disappearing model features.

Every cross ratio triplet associated with correctly matched features can be used to index the look up table containing the position from which this triplet was computed. These positions are accumulated and figs. 4 and 5 shows the largest accumulated positions obtained using correctly matched features together with correct camera positions. Note that the grid spacing is 10 cm so the accuracy averaged largest accumulated positions is around 10 - 20 cm. From the spread of the accumulated positions we see that position uncertainty is largest in the direction to the features and substantially less in the orthogonal direction. This effect is more pronounced in position 2. This can be expected since especially in position 2 the image mapping should be close to parallel and the features are concentrated in a rather small angular sector, which means that perspective will be very small, i.e. cross ratios will not vary substantially in the direction to the imaged features.

References

[1] Y. Lamdan, J.T. Schwartz, and H. J. Wolfson, Object recognition by affine invariant matching. In: Proc. CVPR-88, pp. 335-344, (1988)

[2] C.A. Rothwell, A.P. Zisserman, J.L. Mundy, D.A. Forsyth, Efficient Model Library Access by Projectively Invariant Indexing Functions, In: Proc. CVPR-92, pp. 109-114 , (1992)

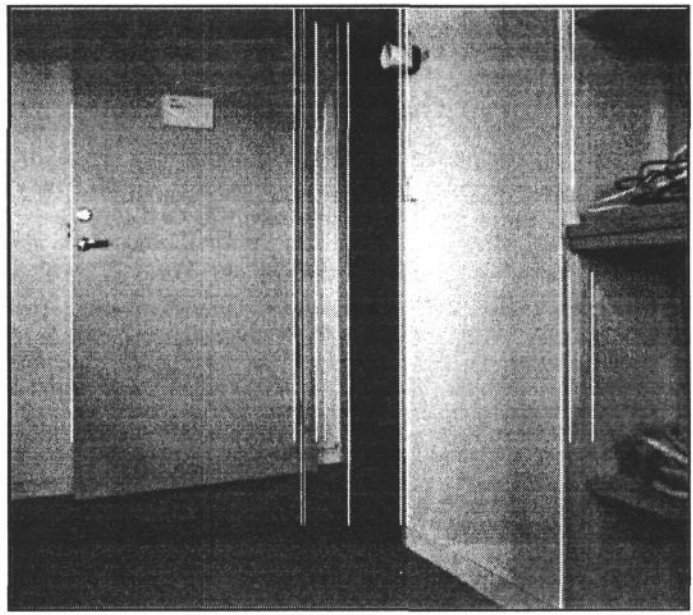
[3] D.T. Clemens and D.W. Jacobs, Space and time bounds on indexing 3-D models from 2-D images, IEEE Trans. on Pattern Analysis and Machine Intelligence, 13, pp 1007-1017. (1991)

[4] D. Jacobs, Space efficient 3D model indexing, IU-workshop pp. 717-725 (1992)

[5] J. B. Burns, R. S. Weiss and E.M. Riseman, View variation of point-set and line-segment features, IEEE Trans. on Pattern Analysis and Machine Intelligence, 15, pp 51-68, (1993)

[6] K. Astrom, A Correspondence Problem in Laser Guided Navigation, Proc. Symposium on Image Analysis, Uppsala, Sweden,pp. 141 - 144, 10-11 March (1992)

[7] G. Stockman, Object recognition and localization via pose clustering, CVGIP 40, 361-387, (1987)



Vertical lines extracted: Image 1

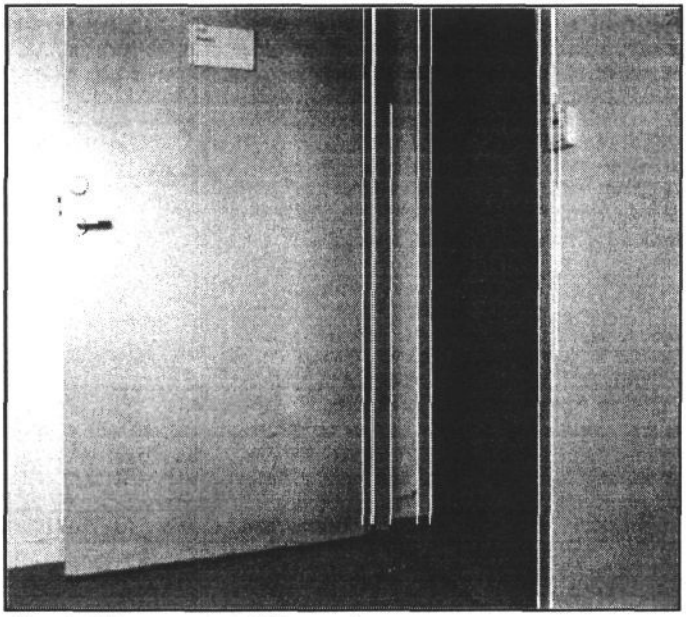
	1	12	98	391	610	<1042>	4	41	23	17	6	2	12	41	32	43	218	45
i	2	193	271	144	266	231	<2022>	243	383	127	35	4	36	27	26	47	272	302
m	3	234	214	223	197	93	826	<2253>	462	320	61	6	16	20	13	210	227	116
a	4	111	362	136	311	353	392	935	<2586>	874	71	33	45	38	26	160	838	85
g	5	316	728	939	1098	612	756	675	644	<3003>	446	278	69	66	56	146	768	574
e	6	115	590	407	663	1347	319	209	464	1311	475	374	138	75	76	99	781	558
	7	120	115	60	45	662	143	92	132	460	<2371>	411	78	100	56	109	42	33
f	8	58	119	124	144	834	11	11	10	417	1535	<2090>	279	107	155	52	75	9
e	9	405	289	1001	192	500	1651	1063	374	171	308	216	<4888>	590	219	856	133	478
a	10	192	203	490	595	514	329	1235	534	197	866	249	856	<3256>	1041	564	552	271
t	11	0	135	255	955	701	19	310	1642	297	301	817	336	1592	<3307>	410	1003	132
u	12	134	101	107	13	129	2050	1109	813	186	514	424	1410	252	287	<1985>	1223	608
r	13	1	56	92	168	63	84	692	1215	1296	639	500	393	256	346	569	<1244>	416
e	14	0	0	0	0	35	22	20	16	9	189	105	167	39	14	274	286	215
	15	0	0	22	26	86	76	92	83	795	1645	551	1068	491	220	419	472	<935 >

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

model feature

1-4 outside picture

Figure 2: Number of associations between image feature (i) and model feature (j) from indexing. Correct associations are marked with < >. Image feature 6 which is a shadow edge and feature 13 do not have any corresponding model features.



Vertical lines extracted: Image 2

i		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
m	1	15	125	536	735	<1415>	9	7	6	10	3	0	3	22	14	39	133	43
a	2	16	180	821	1244	<1916>	10	9	9	16	5	0	6	30	20	60	216	48
g	3	89	130	44	48	57	<3571>	310	274	11	17	10	12	5	5	36	69	47
e	4	101	30	197	86	6	1367	<3479>	494	352	79	49	10	0	0	88	24	61
	5	13	127	15	188	68	261	1167	<3703>	305	39	75	28	12	10	22	108	2
f	6	30	3	38	116	12	11	77	30	<3590>	513	273	34	19	18	4	5	32
e	7	7	2	45	38	130	5	20	41	86	<2502>	337	70	42	7	109	14	10
a	8	0	5	0	46	189	0	0	0	93	1338	<2376>	90	61	55	34	70	2
t	9	0	0	31	2	25	136	83	66	125	19	31	<3607>	229	172	137	36	14
u	10	0	0	0	11	3	6	16	14	45	74	7	515	<1276>	531	77	39	24
r	11	0	0	0	11	4	7	14	11	45	109	6	500	<1409>	659	79	44	28
e	12	0	0	0	12	3	4	6	4	5	86	4	245	645	<505>	46	24	20
	13	0	0	0	1	1	0	0	0	3	14	3	9	25	27	13	17	12

model feature

1-4 outside picture

15-17 outside picture

Figure 3: Number of associations between image feature (i) and model feature (j) from indexing. Correct associations are marked with < >. Image features 1 and 2 and features 10 and 11 are the same edge split into two. Feature 13 is an extra edge not present in the model.

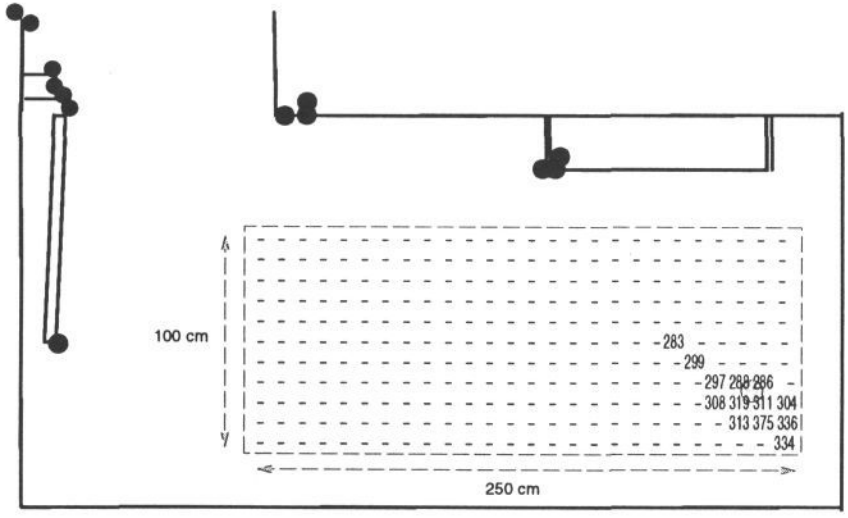


Figure 4: 13 largest numbers accumulated positions from the index table representing relations between cross ratio triplets and x-y position. The grey circle is the correct camera position for image 1.

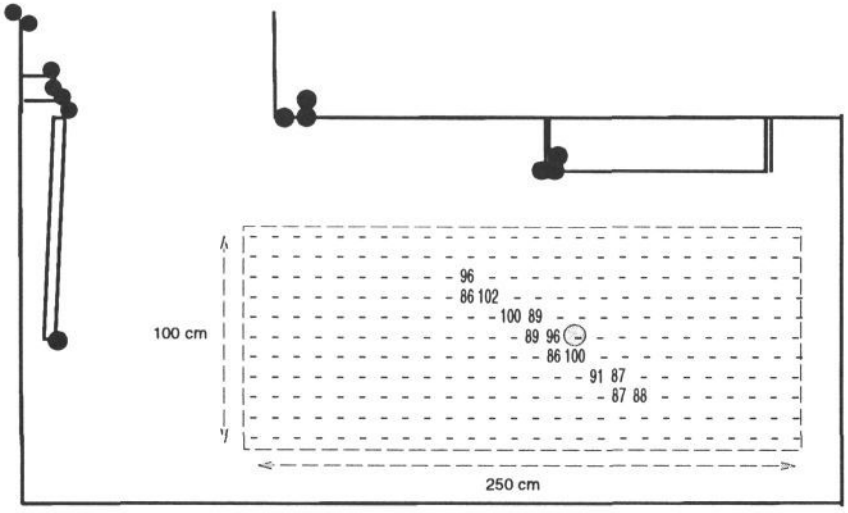


Figure 5: 13 largest numbers accumulated positions from the index table representing relations between cross ratio triplets and x-y position. The grey circle is the correct camera position for image 2.