# On the Appropriateness of Camera Models

Charles Wiles[1,2] and Michael Brady[1]

[1] Robotics Research Group, Department of Engineering Science, University of Oxford, Parks Road, Oxford, OX1 3PJ, UK.
[2] TOSHIBA, Kansai Research Laboratories, Osaka, Japan.

**Abstract.** Distinct camera models for the computation of structure from motion (sfm) can be arranged in a hierarchy of uncalibrated camera models. Degeneracies mean that the selection of the most appropriate camera from the hierarchy is key. We show how accuracy of fit to the data; efficiency of computation; and clarity of interpretation enable us to compute a measure of the *appropriateness* of a particular camera model for an observed trajectory set and thus automatically select the most appropriate model from our hierarchy of camera models. An elaboration of the idea, that we call the *combined appropriateness* allows us to determine a suitable frame at which to switch between camera models.

## 1 Introduction

A number of different camera models have been suggested for the computation of structure from motion (sfm). The appropriateness of the particular model depends on a number of factors. Usually, known imaging conditions and the required clarity in the computed structure are used to choose the camera model in advance. However, for many tasks (for example motion segmentation), it is neither possible to know in advance what the imaging conditions will be nor is the clarity of the computed structure important. Moreover, under degenerate motion or degenerate structure, particular algorithms for particular models may fail due to degeneracies in the solution. In this paper we propose a new measure called "*appropriateness*", which can be determined directly from observed image trajectories and is robust to gross outliers. The measure can be used to automatically select the most appropriate model from a hierarchy of camera models on the basis of *accuracy*, *clarity* and *efficiency*. The hierarchy of models allows degeneracy to be dealt with by explicitly employing reduced models. Moreover, a measure of *combined appropriateness* [4] enables automatic model switching when different camera models are appropriate for different parts of the sequence.

Previous research has tackled only the forward problem of choosing a camera model when the process of image formation is known. The far more difficult inverse problem of choosing the appropriate camera model to compute structure given only the observed image trajectories has not been tackled. This paper is an initial foray into the problem.

## 2 Rigidity

A camera projects a 3D world point $\mathbf{X} = (X, Y, Z, 1)^T$ to a 2D image point $\mathbf{x} = (x, y, 1)^T$. The *projective* camera model [1] can be written

$$\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \mathbf{P} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \quad (1)$$

This formulation places no restriction on the frame in which $\mathbf{X}$ is defined, but the *rigidity constraint* can be enforced by fixing the frame to the object in such a way that $\mathbf{X}$ is constant for all time. Equation 1 can be rewritten $\lambda \mathbf{x}_i(j) = \mathbf{P}(j)\mathbf{X}_i$, where $\mathbf{x}_i(j)$ is the image coordinate of the $i^{th}$ point in the $j^{th}$ pose frame $\mathbf{P}(j)$.

The observed point trajectories, $\mathbf{z}_i(j) = (x_{ij}, y_{ij})^T$, will generally be perturbed by measurement noise and a solution for $\mathbf{P}(j)$ and $\mathbf{X}_i$ is obtained by minimising the sum of the squared error in the image plane. Thus $2nk$ observations are used to estimate $11k$ pose parameters and $3n$ structure parameters. The computed structure will be unique up to an arbitrary 4×4 projective transformation or *collineation*, $\mathbf{H}$. There are 15 independent elements in $\mathbf{H}^{-1}$, so 15 of the pose parameters can be fixed arbitrarily when solving for sfm. Thus the total number of degrees of freedom in the system for the projective camera model is, $dof_p = 2nk - 11k - 3n + 15$. For a unique solution to be found $dof$ must be positive for all values of $n$ and $k$ (with $k$ greater than 1). Thus $n_{min} \geq 7$ and $k_{min} \geq 2$.

## 3 Degeneracy

Under certain imaging conditions and camera motions, the solution to sfm may be underconstrained or ill-conditioned and thus *degeneracy* must be studied. The causes of degeneracy for the sfm problem fall into four categories: (i) Degenerate structure (critical surfaces); (ii) degenerate motion; (iii) degenerate spatial positioning of features; and (iv) poor preconditioning. Degenerate *structure* occurs when, for example, the sfm problem is solved using the 3D projective camera model and the object being observed is in fact planar. Degenerate *motion* occurs, for example, when using the 3D affine camera model [3] and there is no rotation about a shallow object: effectively only one view of the object is seen. Degenerate *spatial positioning* occurs when the features on an object are *by chance* poorly positioned. For example, two features "close together" may give only one feature's worth of information. Alternatively, when trying to fit the planar projective camera model to four points, there will be a degeneracy in the solution if any three points are collinear. *Poor preconditioning* leads to numerical conditioning problems when using the projective camera models [2].

When degenerate structure or motion are the cause of degeneracy, switching to a reduced camera model allows the cause of degeneracy to be removed explicitly and this has lead to our *hierarchy of camera models* (see Table 1). Not

only is it necessary to use reduced models when analysing real scenes, but it is also desirable, as the reduced models generally can be computed more efficiently and can be more easily interpreted (if for example, the planar projective camera model accurately models the data, then the observed object is known to be a plane). In essence when two camera models both accurately model the same data then the reduced model will be the more "appropriate".

In short, by using *the simplest applicable camera model that accurately models the data* we may expect the following gains: (i) removal of degeneracy and improved numerical conditioning (*accuracy*); (ii) greater computational efficiency (*efficiency*); and (iii) clearer interpretation (*clarity*). We are now faced with two questions, (a) what set of camera models should we choose; and (b) how should we switch between them?

The models in Table 1 are listed with the most complex model, the projective camera model, at the top of our hierarchy. We choose to use only uncalibrated camera models, since calibrated camera models are computationally expensive and often fail to converge to an accurate solution. The *image aspectation* model is applicable when the object is shallow and there is no rotation of the camera around the object such that only one view of the object is ever seen (degenerate motion). This situation occurs frequently in road based scenes, for example when following a car on a highway. Though the image aspectation model is not necessary in theory—it is a special case of the planar affine camera model—it proves valuable in practice, since the structure is interpreted as being either 3D *or* planar. The *planar affine* and *planar projective* models are necessary for dealing with degenerate structure. For example, when only one side of an object is visible (when being overtaken by a lorry, for example) the solutions to the 3D *affine* and 3D *projective* models are underconstrained. The planar projective model is exploited further when finding the ground plane.

## 4 Camera model "appropriateness"

The affine camera can be thought of as an uncalibrated version of the weak perspective camera in the same way as the projective camera model can be thought of as an uncalibrated version of the perspective camera model. We refer to shallow scenes viewed with a long focal length camera as meeting *affine imaging conditions* and deep scenes or scenes viewed with a short focal length camera as meeting *projective imaging conditions.*

We propose the following heuristic metric for computing the "appropriateness" of a model,

$$\Xi = \eta^{\rho_\eta} \kappa^{\rho_\kappa} \nu^{\rho_\nu} \quad , \tag{2}$$

where $\eta$, $\kappa$ and $\nu$ are measures of efficiency, clarity and accuracy respectively and $\rho_\eta$, $\rho_\kappa$ and $\rho_\nu$ are their weightings in the equation. We call $\eta$, $\kappa$ and $\nu$ the $\Xi$-*measures* and each takes a lowest value of 0 (extremely inappropriate) and a highest value of 1 (proper). The weights also take values between 0 and 1 and are defined according to the application.

| | | |
|---|---|---|
| 3D Projective | $\mathbf{P}_p = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \end{bmatrix}$ $\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$ | – Deep object and any motion.<br>– $doa_p = 15$. |
| 3D Affine | $\mathbf{P}_a = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ 0 & 0 & 0 & 1 \end{bmatrix}$ $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$ | – Shallow object and shallow motion containing rotation of the camera around the object.<br>– $doa_a = 12$. |
| Planar Projective | $\mathbf{P}_{pp} = \begin{bmatrix} P_{11} & P_{12} & 0 & P_{14} \\ P_{21} & P_{22} & 0 & P_{24} \\ P_{31} & P_{32} & 0 & P_{34} \end{bmatrix}$ $\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} & P_{14} \\ P_{21} & P_{22} & P_{24} \\ P_{31} & P_{32} & P_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$ | – Deep plane and any motion.<br>– Points at infinity and any motion.<br>– $doa_{pp} = 8$. |
| Planar Affine | $\mathbf{P}_{pa} = \begin{bmatrix} P_{11} & P_{12} & 0 & P_{14} \\ P_{21} & P_{22} & 0 & P_{24} \\ 0 & 0 & 0 & 1 \end{bmatrix}$ $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} & P_{14} \\ P_{21} & P_{22} & P_{24} \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$ | – Shallow plane and shallow motion containing rotation of the camera around the plane.<br>– $doa_{pa} = 6$. |
| Image Aspectation | $\mathbf{P}_{ia} = \begin{bmatrix} C & S & 0 & P_{14} \\ -kS & C & 0 & P_{24} \\ 0 & 0 & 0 & 1 \end{bmatrix}$ $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} C & S & P_{14} \\ -kS & C & P_{24} \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$ | – Shallow plane and shallow motion with no rotation of the camera around the plane.<br>– Shallow object and shallow motion with no rotation of the camera around the object.<br>– $doa_{ia} = 5$. |

**Table 1.** *A hierarchy of uncalibrated camera models for analysing arbitrary scenes containing unknown independent motions.*

The *efficiency measure* $\eta$ depends on the particular algorithm used to implement the model. In general it will be a function of the number of observations. For iterative methods of solution it will also depend greatly on the degenerate nature of the data since as the data becomes more degenerate the accuracy of the initial estimate and the speed of descent will both be poor. However by explicitly incorporating degenerate models into our selection procedure we remove this as a factor in the efficiency measure.

In practice the efficiency measure is only useful in the recursive update of structure and motion (in a batch process, once a solution has been obtained the efficiency of its computation is no longer relevant). In our subsequent analysis we deal only with batch methods and so set $\rho_\eta = 0$.

The *clarity measure* $\kappa$ can be defined by considering the ambiguities in the structure solutions. We choose the following definition: $\kappa = doa_{min}/doa_m$, $\rho_\kappa = 1$, where $doa_m$ is the number of degrees of freedom in the structure solution for a particular model $m$ and $doa_{min}$ is the known minimum value of $doa_m$ for all the models in the hierarchy. The division by $doa_{min}$ scales the clarity measure such that the maximum possible clarity score is $\kappa = 1$.

The ideal *accuracy measure* $\nu$ would be to find how accurately the computed structure fits the real structure taking into account the known ambiguity in the solution. This is achieved by finding the sum of the squared error between the real structure and the estimated structure when it is transformed into the frame of the real structure. The transformation can be computed by finding the elements of $\hat{\mathbf{H}}$ (where $\mathbf{H}$ takes the relevant form for the known structure ambiguity) that minimise the sum of the squared structure error,

$$J_E = \sum_{i=1}^{n} \mid \mathbf{X}_i^E - \hat{\mathbf{X}}_i^E \mid^2 ,$$

where $\mathbf{X}^E$ is the actual Euclidean structure and $\hat{\mathbf{X}}^E$ is the computed structure transformed into the Euclidean frame,

$$\hat{\mathbf{X}}_i^E = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \frac{\hat{\mathbf{H}}\mathbf{X}_i}{\hat{\mathbf{H}}_3^\top \mathbf{X}_i} , \quad \mathbf{H} = \begin{bmatrix} \mathbf{H}_1^\top \\ \mathbf{H}_2^\top \\ \mathbf{H}_3^\top \\ \mathbf{H}_4^\top \end{bmatrix} .$$

We then define $\nu = 1/(1+J_E/\sigma^E), \rho_\nu = 1$, where $\sigma^E$ is the expected standard deviation in the structure error in the Euclidean frame and must be determined in advance. This measure is not robust to outliers in the computed structure and a more robust measure would be to compute

$$J'_E = \sum_{i=1}^{n} f_\sqcup \left( \frac{\mid \mathbf{X}_i^E - \hat{\mathbf{X}}_i^E \mid}{3\sigma^E} \right) , \; f_\sqcup(x) = \begin{cases} 0 \text{ if } -1 < x < 1 \\ 1 \text{ otherwise.} \end{cases} ,$$

and $\nu = 1 - \frac{J'_E}{n}$. Unfortunately, $\mathbf{X}_i^E$ are not known and the prior computation of $\sigma^E$ is difficult. Indeed, the standard deviation in the structure error will depend

on the nature of the relevant motion of the camera and object, and is likely to be anisotropic and different for different strcuture vectors. In essence, the accuracy measure cannot, in general, be computed in this way. However, it is possible to compute the sum of the squared image error (indeed this is what we minimise when solving sfm), but the sum of the squared image error bears no direct relation to the sum of the squared structure error. However, the error between each observation and the projected structure estimate either supports or does not support the hypothesis that each structure estimate is accurate. It is thus possible to compute the *robust image error*,

$$J' = \sum_{j=1}^{k} \sum_{i=1}^{n} f_{\sqcup} \left( \frac{| \mathbf{z}_i(j) - \hat{\mathbf{z}}_i(j) |}{3\sigma} \right) , \tag{3}$$

where $\sigma$ is the standard deviation of the expected image noise and must be determined in advance. $J'$ is simply a count of the number of observed trajectory points that fail to agree with the projected structure estimates. The accuracy is then computed $\nu = 1 - (J'/nk)$.

In theory, the robust image error will be in the range 0 to $nk-(n_{min}(k-1)+n)$ and not 0 to $nk$. This is because in the computation of sfm all $n$ points can be made to agree in at least one frame and at least $n_{min}$ points (where the vlaue of $n_{min}$ depends on the camera model) can be made to agree in the remaining $k-1$ frames. Thus, for small data sets or short time sequences, the accuracy measure can be refined to be $\nu = 1-(J'/dof'_m(n,k))$, where we call $dof'_m(n,k) = nk - (n_{min,m}(k-1)+n)$ the *robust degrees of freedom*. In practice it is only necessary to use the refined measure for very small data sets.

The $\Xi$-equation that we use to measure camera model appropriateness is thus,

$$\Xi = \kappa\nu = \left( \frac{doa_{min}}{doa_m} \right) \left( 1 - \frac{1}{nk} \sum_{j=1}^{k} \sum_{i=1}^{n} f_{\sqcup} \left( \frac{| \mathbf{z}_i(j) - \hat{\mathbf{z}}_i(j) |}{3\sigma} \right) \right) , \tag{4}$$

where $\hat{\mathbf{z}}_i(j)$ are a function of the model, $m$.

## 5 Results

In order to test our measure of appropriateness we selected five trajectory sequences each containing structure, motion and imaging conditions well suited to one of the five camera models. The trajectories were automatically generated by the algorithm described by Wiles and Brady [4, 5] and presegmented so that they are known to be consistent with a rigid motion. We then computed sfm on each test set using each of the five uncalibrated models above and computed the appropriateness, $\Xi$. The results are shown in Table 2(a) and the ideal values are shown in Table 2(b). The ideal values are those that would be generated with perfect data such that the accuracy measure equals one for models above the diagonal. The measured values will be less than or equal to the ideal values since

a small percentage of outliers will cause the accuracy measure to be less than one.

| Sequence | Image Aspectation $doa_{ia} = 5$ | Planar Affine $doa_{pa} = 6$ | Planar Projective $doa_{pp} = 8$ | 3D Affine $doa_a = 12$ | 3D Projective $doa_p = 15$ |
|---|---|---|---|---|---|
| "ambulance" jeep | **0.662** | 0.584 | 0.438 | 0.336 | 0.167 |
| "buggy" front | 0.104 | **0.557** | 0.004 | 0.384 | 0.039 |
| "basement" floor | 0.199 | 0.336 | **0.528** | 0.273 | 0.292 |
| "buggy" | 0.000 | 0.160 | 0.176 | **0.337** | 0.113 |
| "basement" | 0.196 | 0.149 | 0.141 | 0.167 | **0.294** |

(a)

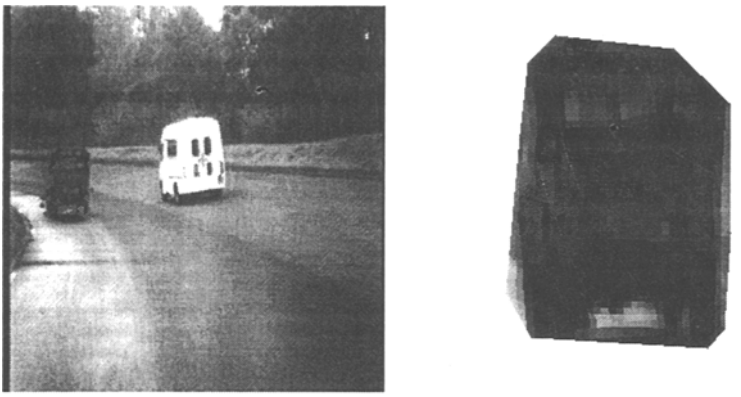
| Imaging conditions appropriate for ... | Image Aspectation $doa_{ia} = 5$ | Planar Affine $doa_{pa} = 6$ | Planar Projective $doa_{pp} = 8$ | 3D Affine $doa_a = 12$ | 3D Projective $doa_p = 15$ |
|---|---|---|---|---|---|
| Image Aspectation | **1.0** | 0.83 | 0.62 | 0.42 | 0.33 |
| Planar Affine | - | **0.83** | 0.62 | 0.42 | 0.33 |
| Planar Projective | - | - | **0.62** | - | 0.33 |
| 3D Affine | - | - | - | **0.42** | 0.33 |
| 3D Projective | - | - | - | - | **0.33** |

(b)

**Table 2.** *Camera model appropriateness values, $\Xi$. (a) Results for a number of sequences. (b) Ideal camera model appropriateness values. The dashes may be any value lower than the diagonal value on that row.*

The "ambulance" sequence consists of the camera following a jeep down a road at a roughly constant distance and an ambulance overtaking the camera on the right. The jeep is a shallow object about which the camera does not rotate and thus the image aspectation camera model should be most appropriate for this object. All the models give accurate solutions for sfm, but the image aspectation camera model has the highest value of $\Xi$ due to its greater clarity. The resulting planar structure is shown in Figure 1(b). The figure was generated by mapping texture from the image onto the computed corner structure.

The "buggy" sequence consists of a model buggy rotating on a turntable. It is viewed from a distance with a camera with a narrow field of view creating a shallow three-dimensional scene. Thus, the affine camera model should be most appropriate. Both the projective and affine camera models should give accurate results, but the projective camera fails to find a solution (due to degenerate imaging conditions). Hence the affine camera model has the highest value of $\Xi$. The resulting structure is shown in Figure 2(c).



(a) (b)

**Fig. 1.** *Planar affine structure computed from the "ambulance" sequence. (a) The image from the last frame. (b) the planar affine structure computed by applying the image aspectation camera model. The $X$ and $Y$ axes of the computed structure are perpendicular and lie in the plane of the paper.*

The subset of the trajectories corresponding to the front of the "buggy" forms a shallow planar scene. Thus, the planar affine camera model should be the most appropriate, and they are. All models except the image aspectation camera model should generate accurate solutions, but again the projective models fail to find a solution (due to degenerate imaging conditions). The two affine models had equal accuracy measures, but the planar model had greater clarity in its solution. Hence the planar affine camera model has the highest value of $\Xi$.

The "basement" sequence consists of a camera translating along the optical axis through a deep three-dimensional scene. Thus, the projective camera model should be most appropriate. Indeed, only the projective camera model gives an accurate solution and hence it has the highest value of $\Xi$. The resulting structure is shown in Figure 2(d).

The subset of the trajectories corresponding to the basement floor forms a deep planar scene. Thus, the planar projective camera model should be the most appropriate. Both the planar projective and projective camera models accurately model the observed trajectories. The sum of the squared image errors is a factor of two smaller for the full projective model due to the greater degrees of freedom in the solution. This demonstrates that, for the purpose of computing appropriateness, the sum of the squared image errors is a poor accuracy measure. However, the robust image error is approximately equal for the two models indicating that on accuracy alone the two models have equal appropriateness. The clarity measure for the planar model is twice as good as the three-dimensional model and hence the planar projective camera model has the highest value of $\Xi$. The resulting structure is shown in Figure 3(b).

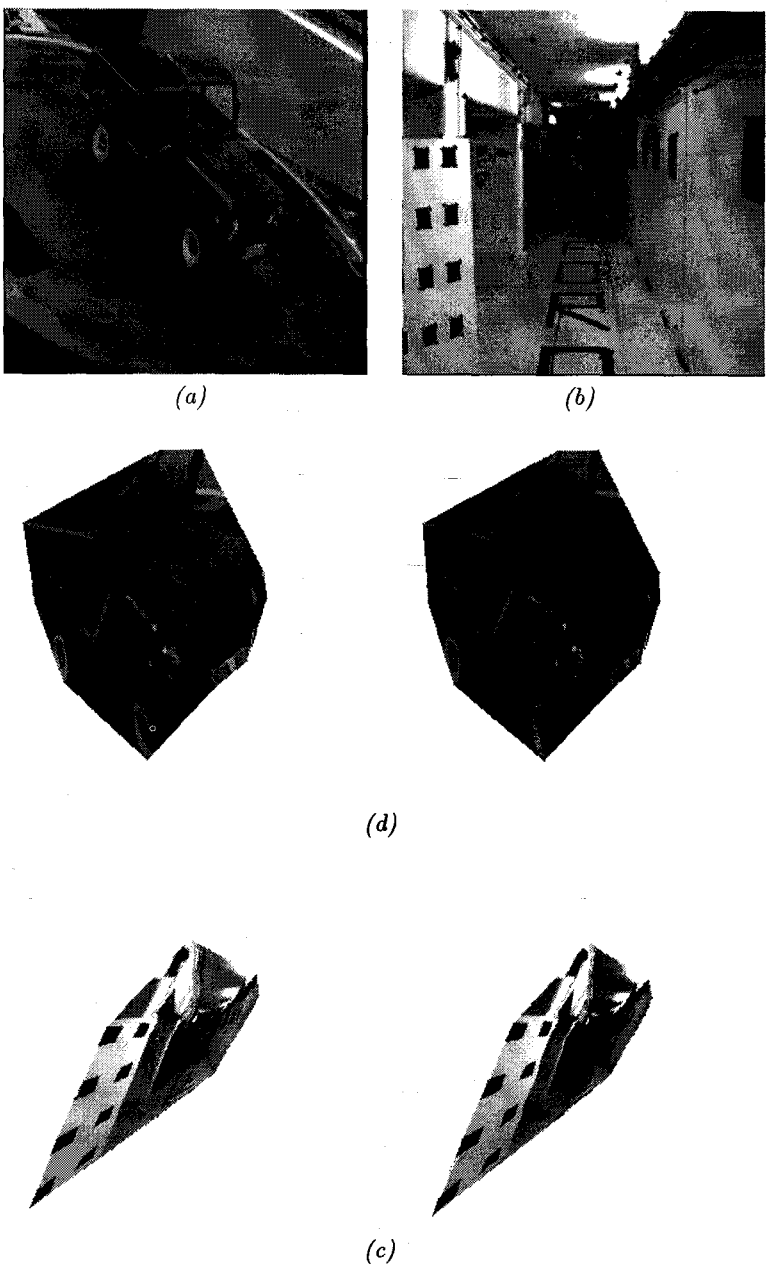
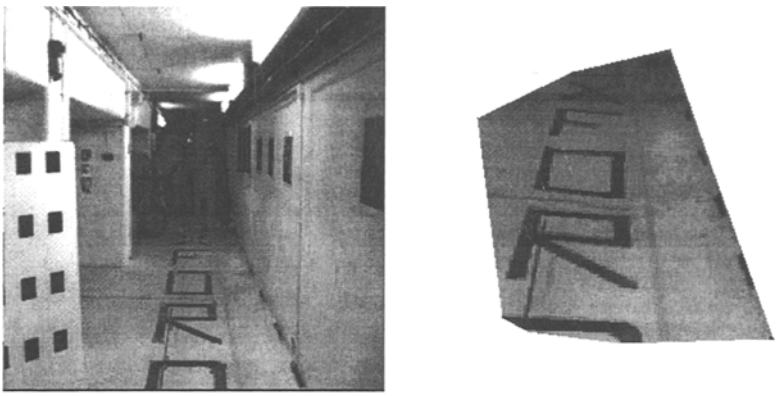



**Fig. 2.** 3D *structure computed from the "buggy" and "basement" sequences. (a) & (b) Images from the last frames. (c) & (d) Cross-eyed stereo pairs showing the structure computed by applying the affine and projective camera models to the sequences respectively.*



(a) (b)

**Fig. 3.** *Planar projective structure computed from the "basement" sequence floor. (a) The image from the last frame. (b) the planar projective structure computed by applying the planar projective camera model. The X and Y axes of the computed structure are perpendicular and lie in the plane of the paper.*

## References


1. O. Faugeras. *Three-Dimensional Computer Vision.* The MIT Press, 1993.
2. R.I. Hartley. In defense of the 8-point algorihtm. In *Proc. 5th International Conference on Computer Vision*, pages 1064–1070, June 1995.
3. J.L. Mundy and A. Zisserman, editors. *Geometric invariance in computer vision.* The MIT Press, 1992.
4. C.S. Wiles. *Closing the loop on multiple motions.* PhD thesis, Dept. Engineering Science, University of Oxford, Michaelmas 1995.
5. C.S. Wiles and J.M. Brady. Closing the loop on multiple motions. In *Proc. 5th International Conference on Computer Vision*, pages 308–313, June 1995.