Model Selection for Automated Architectural Reconstruction from Multiple Views

Tomáš Werner, Andrew Zisserman
Visual Geometry Group, University of Oxford

Abstract

We describe progress in automatically fitting a plane plus modelled perturbation surface model to represent architectural scenes. There are two areas of novelty. The first is a method of fitting parametrized models in which the cost function is based on a combination of disparity and gradient extrema, both computed over multiple views. The second is the use of an evaluation criteria for model selection, learnt from training examples. We demonstrate the method on reconstructions of several college scenes from multiple images.

1 Introduction

This[1] paper describes progress in the automated reconstruction of piecewise planar models from multiple images [2, 4, 7, 11, 15]. The work is targetted on architectural scenes.

We investigate two areas here. The first is a method of fitting parametrized model primitives. The model is fitted to a region of an image using two types of information, depth and edges, both computed over multiple views. A set of parametrized model primitives appropriate for the targetted scenes are defined and engineered. The models represent, to a reasonable approximation, indentations such as doors and inset windows, and protrusion such as bay windows, dormer windows, etc.

The second area investigated is that of model selection: given a set of such models, how should the one be selected which best "explains" the scene? The answer proposed is to learn certain characteristics of the types of scene (here architectural) from training images, and use these characteristics to determine the Bayesian probability of each model.

The approach is influenced by three previous papers: first, the Facade modelling system [12] which generated models of excellent visual quality using a representation based on simple geometric primitives (e.g. cuboids, cylinders) from a small set of images. However, model selection and fitting in Facade is entirely manual. Second, the approach of Dick *et al* [4] in which models (e.g. a window) have strong priors (e.g. on height and width) specifying their shape and appearance. Third, as an example of a learning approach, the edge detection method of Konishi and Yuille [6] in which filter responses are learnt at edges in training images in order to determine segmentations.

In this paper we first fit a coarse piecewise planar model to the scene (section 2) and then refine each plane by modelling perturbations from it. This differs from the Facade system (and later in [3]) where perturbations from the plane are computed by dense stereo. Here, the perturbation is defined by the parametrized models. These models include, for

[1]We are very grateful to Frederick Schaffalitzky for computing the projective reconstructions for each set of multiple images, and to helpful discussions with Phil Torr. Funding was provided by a Marie Curie fellowship and EC project Vibes.

Figure 1: Five images of the College set acquired with a hand held low cost Olympus C-820L digital camera. The image size is 1024×768 pixels.

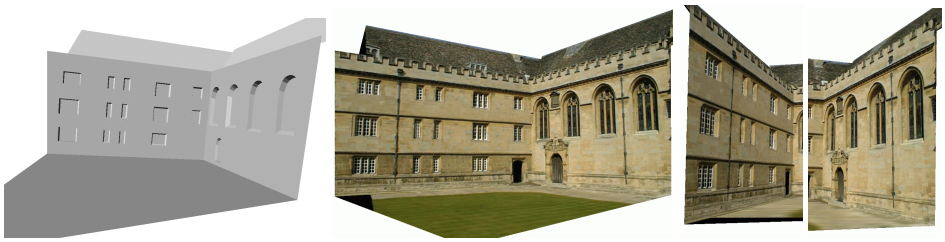


Figure 2: Automatically generated reconstruction computed from the images of figure 1 (respectively: shaded model, and three shots of textured model). Modelling indentations (windows, doors) in the walls is the topic of this paper. Note in the close-ups, the augmented walls look much more realistic than mere planes.

example, rectangular blocks and truncated pyramids which are able to represent unchamfered and chamfered windows respectively, and are defined in section 3. The models are quite generic and do not require strong priors. The fitting approach is described in section 4. A typical example is shown in figure 2.

2 Background – computing piecewise planar shells

This section briefly describes the multiple view calibration, feature matching, and coarse planar model fitting stages that are the pre-requisite of the parametrized model fitting methods. The stages are not novel and details are given in previous publications [15]. They are included so that the paper is relatively self-contained.

The input is a set of (three or more) uncalibrated images of the target scene, and the desired output consists of the following description: a metric reconstruction of the cameras for each view; 3D points and their images; 3D lines and their images; and a coarse piecewise planar representation covering the principal planes of the 3D scene. This is sufficient information to rectify each of the principal planes.

It is assumed that there are three principal directions in the scene (vertical and two orthogonal horizontal), and that they can be retrieved from images of straight lines. This assumption is typically valid for the type of architectural scenes targetted in this work. The method will be illustrated for the five images of a college quad shown in figure 1.

Projective Reconstruction: because the viewpoints are often significantly different in hand held still photographs, wide baseline matching methods [8] are necessary to compute interest point matches between views. We use the method described in [9] which is based on affine invariant descriptors and robust estimation of multiple view geometry (see [5, 13, 14, 16]). The result is a camera corresponding to each image and a set of 3D points, defined up to an unknown projective transformation of 3-space. The RMS reprojection error after bundle adjustment is 0.14 pixels.

Metric Reconstruction: vanishing points corresponding to the three principal directions are computed from image lines in each image independently using a RANSAC approach. The vanishing points are matched across all the images. The projective reconstruction is then upgraded to metric using two constraints: that the three principal directions (which are the pre-images of the vanishing points) are mutually orthogonal; and that the cameras have square pixels. The result is a metric reconstruction of the cameras, 3D points, and three principal directions. The RMS reprojection error after metric bundle adjustment is 0.15 pixels. The reconstructed 3D point cloud is shown in figure 3.

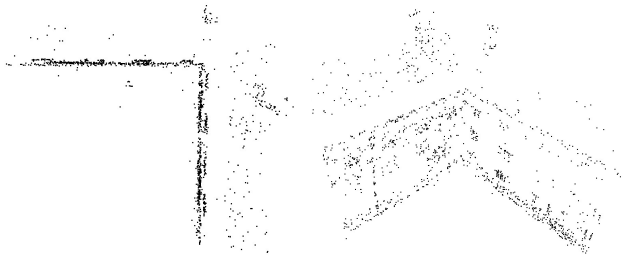


Figure 3: Two views of the cloud of 3D points computed from the images of figure 1 after metric rectification. There are 1331 points.

Line matching: lines are matched over all views using an extended version of the algorithm given in [10]. The original algorithm was limited to image triplets, and here the extension is to any number of views with each view treated equally, i.e. changing the ordering of the views does not affect the matching. The 3D line segments are estimated by minimizing reprojection errors over all views in which the line appears. Deficiencies in the line detection, such as over segmentation (lines erroneously broken) or inconsistent end points in different views, are remedied to some extent by combining the information available from the multiple views. A view of the 3D lines is shown in figure 4a.

Coarse plane fitting: the aim of this stage is to determine the principal planes of the scene, and hence to form a coarse piecewise planar approximation which will be the basis for the subsequent plane plus perturbation model fitting. Given the above stages there is a wealth of geometrical information now available to aid this coarse plane fitting. The principal directions are important in fitting particular planes in turn. For example, to extract a vertical wall two strategies may be applied: (1) sweep (or RANSAC) for a virtual plane in a vertical direction and score the plane's position by the number of 3D points and lines that lie on it; or (2) score the plane's position by computing image cross-correlation between all views using the homography induced by the plane. The polygonal representation of each plane is determined by their mutual intersections (taking into account visibility) and the image outline.

Currently, we first use 3D points and lines to generate vertical wall hypotheses, which are then verified and disambiguated using homography-based image cross-correlation, maximizing the number of explained image points by a simple optimization technique. Then, roof planes are obtained either from diagonal 3D lines belonging to the roof or by sweeping 3D planes around horizontal lines as in [1]. Figure 4b shows the planes fitted, and figure 4c the wireframe outline of their extent.

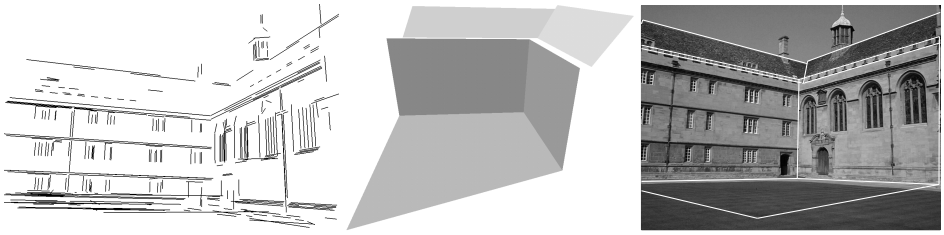


Figure 4: 3D reconstruction computed from the images of figure 1: (a) A view of the 3D lines (there are 380 lines); (b) A view of the 3D coarse piecewise planar model. It includes the five principal scene planes (ground, walls and roofs); (c) The wireframe of the coarse planar model projected onto the second image.

3 Model definition and enumeration

Each plane of the coarse model is now refined by modelling perturbations from the plane. In this section we describe the set of models for these perturbations and their parameters. All that is required to instantiate a model at a point on the plane is a rectangular region and its depth. The description of model instantiation is deferred until section 5.

3.1 Model set

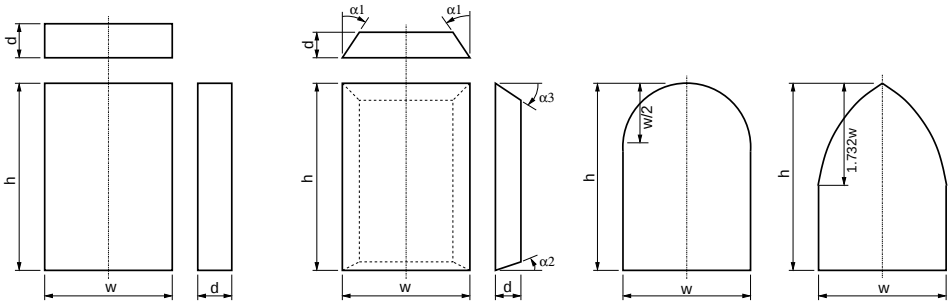


Figure 5: Four instances of the generic indentation models: box, truncated pyramid, circular arch, and gothic arch. The model space is obtained by combining these models and by varying w, h, and α_i. The depth d is known in advance for each model.

The span of the model set is illustrated in figure 5. The simplest instance is that of a rectangular box where the indentation is perpendicular to the host plane. This model is specified by four parameters (the position of the left, right, top and bottom lines on the plane). A truncated pyramid, representing a chamfered window, is obtained by varying three additional angle parameters (α_1, α_2, and α_3). A model representing a window with a circular arch is obtained by replacing the top line by a circular arc. Its chamfer angle is specified by that of the side lines. A gothic arch is obtained in a similar manner.

A particular model is thus specified by the following parameters: the position of the four sides, the chamfer angles, the type of top (straight, circular gothic). Convexity and concavity are distinguished by the sign of the depth d.

3.2 Model enumeration

The quality of the fit is assessed by measuring an objective function based on multiview correspondence, and this is described below in section 4. Here we describe how values of the various parameters are determined, so that a set of candidate models can be enumerated. Models are enumerated (and later fitted) for each boundary separately. We will illustrate the parameter determination for the right vertical side of the window in figure 6a.

In architectural scenes, the outer and inner boundaries of indentations and protrusions almost always correspond to edges. These edges are often very weak and cannot be detected by the usual method of independent edgel detection followed by linking and line fitting. Instead here the horizontal image gradient is aggregated vertically in order to achieve a sufficiently high signal to noise ratio to detect the window edge.

Furthermore, since there are multiple views of each wall, and their correspondence is known from the wall homography, the position of the gradient magnitude extrema from other views are also available (see figure 6b). This means that an edge that is too weak in one view can still be located if it is present in several others. An edge which is located in three or more views is included in the candidate list of parameter values. Typically, 5-15 positions are obtained. A set of parameter values for the left side and bottom boundary are determined in a similar manner.

Determining the set of values for the remaining parameters (e.g. the chamfer angles) is slightly more involved because of mutual dependencies. Chamfer angles are determined to be consistent with the supplied depth and measured image edges. E.g., for the right vertical side the set of possible pairs [side position, α_1] is enumerated to be consistent with the positions of the candidate vertical edges.

There are three possible top boundaries. In the case of a circular or gothic arch, the width of the arch is determined by the width between the left and right sides. In the case of a straight top boundary (a box or truncated pyramid) the set of positions can again be determined independently of the other sides. The set of hypothesized vertical positions of the circular and gothic arches is enumerated in a similar manner as in figure 6b. The gradient is aggregated along the arch curve rather than line segment.

In summary, a set of candidate models specified by their parameter values is enumerated. We now turn to evaluating the model's fit to the image data.

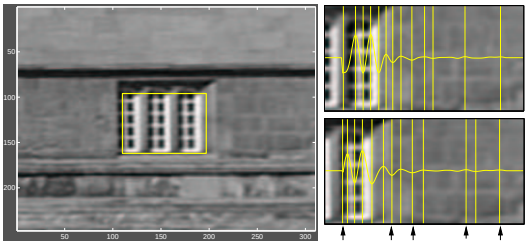


Figure 6: Determining a candidate set of values for the right window boundary. (a) An example window from the first image of the set in figure 1 with the instantiation region indicated as a yellow rectangle. (b) The vertically aggregated image gradient from the corresponding region of images 1 and 2. The edges marked by arrows are consistent across the multiple views and form candidate positions for the window side.

4 Model selection

Section 3 provides a finite set of models, and our task now is to evaluate this set, and select the model which "best explains the image data" over the multiple views. Here "best" means having the highest probability over all the models in the proposed set, and we adopt a learning approach to determine these probabilities. We first describe how a model probability is computed, and then an efficient model selection algorithm for comparing the probabilities across the entire model set.

4.1 Model probability

The probability of each model M_i is determined from the Bayesian rule

$$P(M_i|D) = P(D|M_i)P(M_i)/P(D)$$

where D is the measured data. No prior information on which models are in the scene is assumed, so $P(M_i)$ is equal for all i. The likelihood $P(D|M_i)$ is learnt from ground truth examples in a training image set as described below.

There is, of course, a choice to be made for which data D to use. We use cross-correlation over multiple views as it provides a simple means to measure the model fit over all the views. In more detail, suppose we are computing the cross-correlation between a reference image I_0 image and a second image I for a particular model M_i. The shape of the model M_i induces a transformation T_i from the reference image to I, which maps the points between the views, with some points being occluded. The normalized cross-correlation $c(j,i)$ is then computed for each point j as $c(j,i) = \mathrm{ncc}[I_0(\mathbf{x}_j), I(T_i(\mathbf{x}_j))]$. The probability $p(c)$ of obtaining a correlation c is provided by training images (see below).

It is assumed that the correlation for each point j is independent (the neighbourhoods do not overlap) so

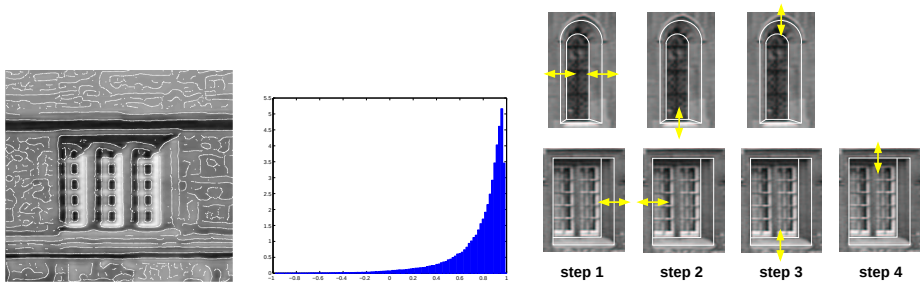
$$P(D|M_i) = \Pi_{\mathrm{points}j} p(c(j,i)) \tag{1}$$

4.2 Learning the correlation probability

Ground truth models are fitted to all the windows of a training image set, i.e. the correct type of model (e.g. a box window or a gothic window) is fitted in each case. The models are instantiated manually (the rectangular region around the window is selected), and an objective function based on multiple view cross-correlation is used to bootstrap the fitting.

Points are selected from edgels computed across the fitted region (including a flange extended over the host wall, see figure 7a). For each point the cross-correlation, given the ground truth model, is measured as described above and a normalized histogram of values computed. Figure 7b shows an example.

If the model were a perfect fit, then the pdf would consist of a single peak at unity. However, the model is at best only an approximation because structures inside the window (e.g. window grills, other arches) are unmodelled. The learnt pdf reflects this approximation.



(a) (b) (c)

Figure 7: (a) The correlation is measured at the edgels shown distributed over the fitted region. (b) The learnt PDF which maps measured correlation (in the range $[0, 1]$) to its corresponding probability. (c) The boundary order followed when evaluating probabilities over the model set. Model fitting differs for windows which do or do not contain a vertical occluded edge.

4.3 Evaluating a model

The model is selected by maximizing expression (1). In addition to all of the models in the set, the plane itself is also included so that evidence for no perturbation from the plane is also considered.

It might appear that fitting a model involves an expensive search over at least a seven dimensional parameter space. However, this is not the case because fitting can be partitioned into a set of sequential and independent low-dimensional searches, as illustrated in figure 7c. For example fitting the top of the model is independent of fitting the bottom.

Furthermore as we run over the set of models the induced transformation T_i is the same for many points and so the cross-correlation need not be re-evaluated. The most important contributions occur where a change in model generates a change in occlusion (e.g. switching between a straight boundary and gothic arch).

5 Model instantiation

Having described the models and their fitting, it only remains to specify how models are instantiated at a point on a plane. This is described in this section.

The aim is to identify promising sites on the plane which are not consistent with the fitted coarse plane model, and thus may be better explained by a perturbation from the plane. The steps involved are illustrated in figure 8.

As the parallax of the window plane with respect to the wall plane can be very small, high accuracy is required to identify off plane regions and to fit models. Here off plane points are identified by sweeping a plane in a direction perpendicular to the wall, and determining a point's depth by an extrema in the cross-correlation of its neighbourhood. Groupings of points at the same depth are obtained by clustering.

The result is a rectangular region with known depth corresponding to each cluster of off-plane points. This provides the instantiation regions for the enumeration and fitting.

6 Results

An overview of the complete algorithm is given in table 1. Results for the image set from figure 1 are shown in figure 2. The results for two other sets (captured with the same

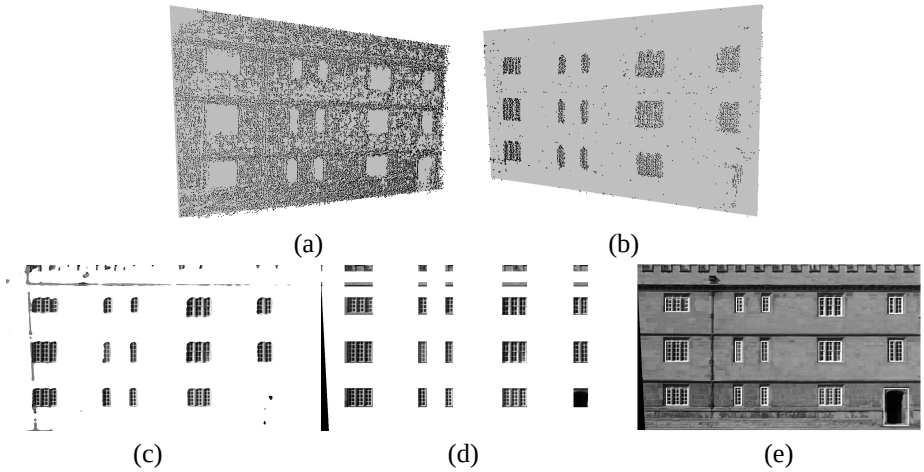


Figure 8: Instantiating models by finding off-plane rectangles. (a) Dense 3D points which lie on the wall plane – the gaps are evident. (b) The points off the plane. (c) Groupings of the off plane points shown as image regions. (d) In the case of a regular window grid, the rectangles can be elegantly obtained as the first eigenimage of the SVD of the score image. This is not used here, but is discussed in section 7. (e) Instantiation regions computed from (c) and shown as superimposed white rectangles.

Olympus C-820L camera) are shown in figures 9 and 10.

The latter was more difficult: the wall profile is more complex than for the College sets – there are several planes on the front wall, the windows sides have a complicated shape, and some of them are close to each other, forming pairs. Since the windows have multiple depths (they are deeper in the rectangular lower region and shallower in the top region), the model set was extended by allowing the depth parameter to change in 3 discreet steps. Note, nested window structure (which we do not address) caused the failure of the algorithm for the window pair in the middle of the wall because it was detected as a single window.

7 Discussion and future work

We have demonstrated that a plane plus modelled perturbation model can be successfully fitted using learnt characteristics. Many variations on this approach are possible, including learning more attributes from the scene (e.g., the distribution of responses to fitting an *incorrect* model), and it should also now be applied to other types of scene.

Currently all of the models are fitted individually. Often though in architectural scenes windows on a facade repeat in design and size, and also are often arranged on a grid with regular spacing. This would correspond to a non-uniform joint prior. An effective method of imposing regularity is illustrated in figure 8. These constraints can be used in two ways: first to make fine adjustments to the fitted models so that for example the tops are aligned to be collinear. Second, a single model can be fitted in all appropriate cases and the parameters optimized over all instances: i.e. the shape parameters are global, but determined from all instances of that model.

Figure 9: Results for the College B image set.

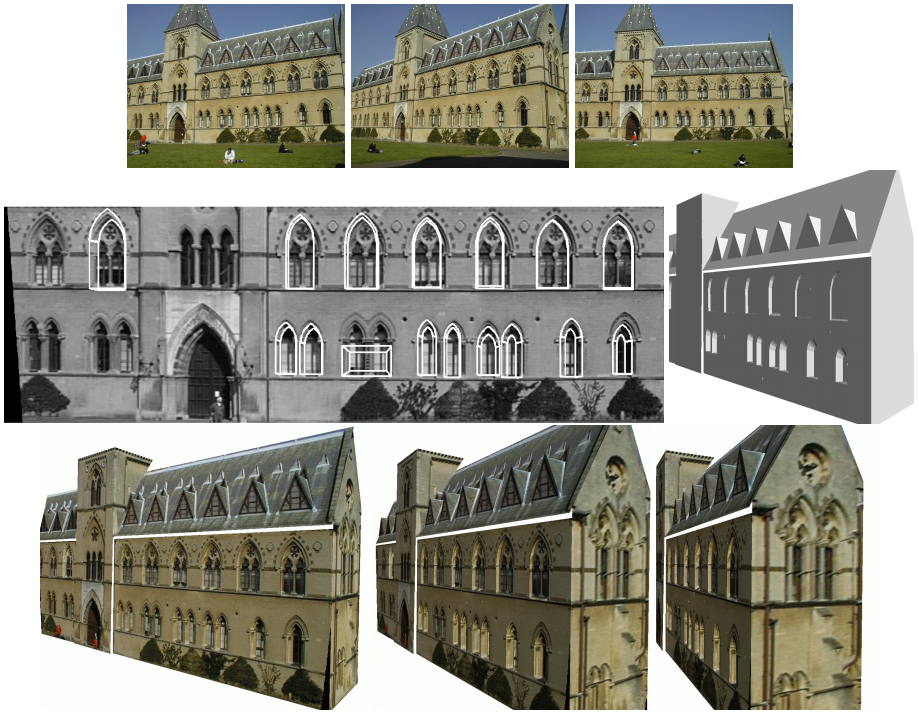


Figure 10: Results for the Library image set.

Given 3+ overlapping views of a static scene:

1. Metric reconstruction of points, lines, and cameras
2. Coarse planar shell model
3. Plane plus modelled perturbation model:
   For each plane of the coarse model:
   - determine regions for instantiating models
   - enumerate the set of models for each region
   - select the highest probability model from the set
4. Model generation
   - texture map from unoccluded image
   - generate VRML model

Table 1: A summary of the main steps of the reconstruction algorithm.

References

[1] C. Baillard, C. Schmid, A. Zisserman, and A. Fitzgibbon. Automatic line matching and 3D reconstruction of buildings from multiple views. In *ISPRS Conference on Automatic Extraction of GIS Objects from Digital Imagery, IAPRS Vol.32, Part 3-2W5*, pages 69–80, Sep 1999.

[2] S. Coorg and S. Teller. Extracting textured vertical facades from controlled close-range imagery. In *Proc. CVPR*, pages 625–632, 1999.

[3] A. R. Dick, P. H. S. Torr, and R. Cipolla. Automatic 3d modelling of architecture. In *BMVC*, 2000.

[4] A. R. Dick, P. H. S. Torr, S. J. Ruffle, and R. Cipolla. Combining single view recognition and multiple view stereo for architectural scenes. In *Proc. ICCV*, pages 268–280. IEEE Computer Society, 2001.

[5] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

[6] S. Konishi and A. L. Yuille. Statistical cues for domain specific image segmentation with performance analysis. In *Proc. CVPR*, 2000.

[7] M. Pollefeys, R. Koch, and L. Van Gool. Self calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *Proc. ICCV*, pages 90–96, 1998.

[8] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *Proc. ICCV*, pages 754–760, Jan 1998.

[9] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?" . In *Proc. ECCV*. Springer-Verlag, 2002.

[10] C. Schmid and A. Zisserman. Automatic line matching across views. In *Proc. CVPR*, pages 666–671, 1997.

[11] H. Tao, H. S. Sawhney, and R. Kumar. A global matching framework for stereo computation. In *Int. Conf. Computer Vision*, pages 532–539. IEEE Computer Society, 2001.

[12] C. Taylor, P. Debevec, and J. Malik. Reconstructing polyhedral models of architectural scenes from photographs. In *Proc. ECCV*. Springer-Verlag, 1996.

[13] P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *IJCV*, 24(3):271–300, 1997.

[14] P. H. S. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15:591–605, 1997.

[15] T. Werner and A. Zisserman. New techniques for automated architecture reconstruction from photographs. In *Proc. ECCV*. Springer-Verlag, 2002.

[16] Z. Zhang, R. Deriche, O. D. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78:87–119, 1995.