VHS to VRML: 3D Graphical Models from Video Sequences

Andrew Zisserman and Andrew Fitzgibbon and Geoff Cross
Dept. of Engineering Science, University of Oxford,
19 Parks Road, Oxford OX1 3PJ, UK
e-mail: {az,awf,geoff}@robots.ox.ac.uk

Abstract—**We describe a method to completely automatically recover 3D scene structure together with a camera for each frame from a sequence of images acquired by an unknown camera undergoing unknown movement. Previous approaches have used calibration objects or landmarks to recover this information, and are therefore often limited to a particular scale. The approach of this paper is far more general, since the "landmarks" are derived directly from the imaged scene texture. The method can be applied to a large class of scenes and motions, and is demonstrated here for sequences of interior and exterior scenes using both controlled-motion and hand-held cameras.**

We demonstrate two applications of this technology. The first is the construction of 3D graphical models of the scene; the second is the insertion of virtual objects into the original image sequence. Other applications include image compression and frame interpolation.

I. Introduction

The goal of this work is to obtain 3D scene structure and camera projection matrices from an uncalibrated sequence of images. The structure and cameras form the basis for a number of applications and two of these will be illustrated in this paper. The first application is building 3D graphical models from an image sequence acquired by a hand-held camcorder. This enables texture mapped models of isolated objects, building interiors, building exteriors etc to be obtained simply by videoing the scene, even though with a camcorder the motion is unlikely to be smooth, and is unknown *a priori*. The second application is to use the camera which is estimated for each frame of the sequence in order to insert virtual objects into the original real image sequence [15]. An 'augmented reality' facility of this type is of use for post-production in the film industry.

To obtain the structure and cameras we employ Structure and Motion recovery results from the photogrammetry and computer vision literature, where it has been shown that there is sufficient information in the perspective projections of a static cloud of 3D points and lines to determine the 3D structure as well as the camera positions *from image measurements alone*. In our approach these points and lines are obtained automatically from features in the scene, and their correspondence established across multiple views. Establishing this correspondence is a significant part of the problem.

The core of the system is shown in figure 1. This automatic process can be thought of, at its simplest, as converting a camcorder to a sparse range sensor. Together with more standard graphical post-processing such as triangulation of sparse 3D point and line sets, and texture mapping from images, the system becomes a "VHS to VRML" converter (VRML is the Virtual Reality Modeling Language, a standard for the interchange of 3D scenes).

The key advantage of the approach we adopt is that no information other than the images themselves is required *a priori*: more conventional photogrammetry techniques require calibration objects or 3D control points to be visible in every frame.

A. Background

Although the general framework for uncalibrated structure from motion has been in place for some time [4], [11], [14] only recently have general acquisition systems come near to becoming a reality. This is because a combination of image processing, projective geometry for multiple views [9], [20], [22], and robust statistical estimation [26], [28] has been required in order to succeed at automating structure and motion algorithms [2], [12].

Tomasi and Kanade [24] demonstrated that 3D models could be built from an uncalibrated sequence, but employed a simplified projection model. The work described in this paper is an improvement in three respects: first, the most general projection model is used — significant perspective effects (giving rise to vanishing points etc) will degrade the Tomasi and Kanade results; second, their system uses a simple point tracker to find matches and does not employ robust statistics and rigid geometry for tracking—this severely limits camera motions and scene types; third, their method is optimal when points are visible in all images, and that restriction does not apply to our work.

B. The scope of the approach

The limitations of the approach of this paper are: first, that the images must be sufficiently "interesting"—if the scene has no significant texture (to be defined more precisely later), then the feature based methods we use will have too few 2D measurements to work with; and second,

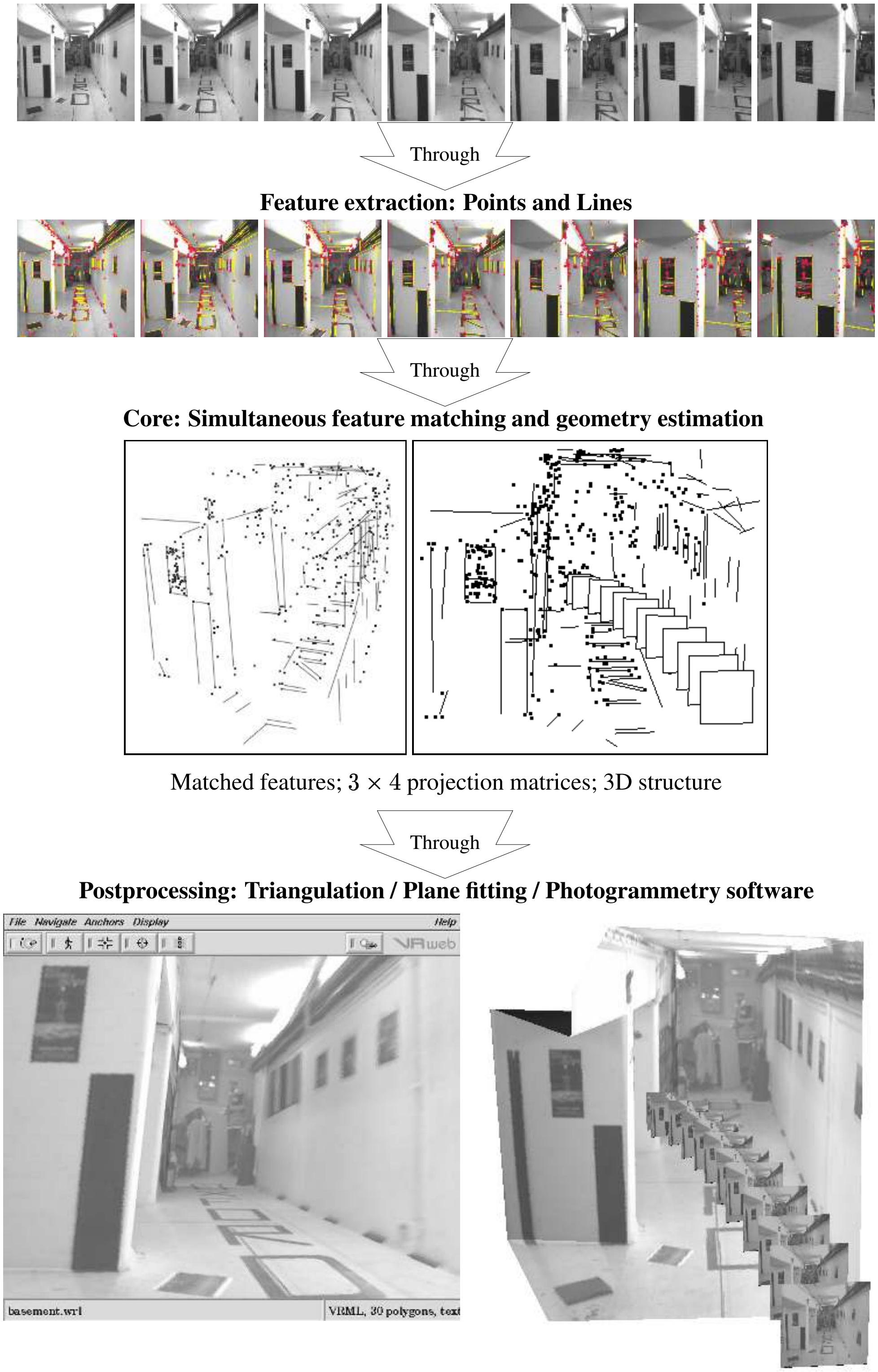

Through
Feature extraction: Points and Lines
Through
Core: Simultaneous feature matching and geometry estimation
Matched features; 3 × 4 projection matrices; 3D structure
Through
Postprocessing: Triangulation / Plane fitting / Photogrammetry software
File Navigate Anchors Display
Help
VRweb
basement.wrl
VRML, 30 polygons, text

Fig. 1. Overview of the VHS to VRML process. Six of the eleven frames from an input video sequence are shown at the top. The images are acquired by a camera mounted on a vehicle moving down a corridor. A three dimensional reconstruction of points and lines in the scene and cameras (represented by their image planes) is computed automatically from the images. A texture mapped triangulated graphical model is then constructed. Left: a rendering of the scene from a novel viewpoint different from any in the sequence. Right: VRML model of the scene with the cameras represented by their image planes (texture mapped with the original images from the sequence).

that the camera motion between images needs to be relatively small, in particular rotation about the optical axis should be limited—otherwise the cross-correlation techniques used to match the features between images will fail. Happily, this restricted motion is the typical motion between frames of a video sequence, and the system is tuned for such data. We also require that the 3D scene be largely static, although smaller independently moving objects—shadows, passing cars and the like—are excised automatically by the use of robust estimation techniques.

The advantage of a video sequence, where the distance between camera centres (the baseline) for successive frames is small, is that evaluating correspondences between successive images is simplified because the images are similar in appearance. The disadvantage is that the 3D structure is estimated poorly due to the small baseline. However, this disadvantage is ameliorated by tracking over many views in the sequence so that the effective baseline is large.

II. Review: the correspondence problem over multiple views

In this section we rehearse the method for establishing correspondences throughout the sequence, and thence compute the scene and camera reconstruction.

Under rigid motion there are relationships between corresponding image points which depend only on the cameras and their motion relative to the scene, but not on the 3D structure of the scene. These relationships are used to guide matching. The relationships include the epipolar geometry between view pairs, represented by the fundamental matrix [4], [11]; and the trifocal geometry between view triplets, represented by the trifocal tensor [9], [20], [22]. These relationships, and image correspondences consistent with the relations, can be computed automatically from images, and this is described below.

Geometry guided matching, for view pairs and view triplets, is the basis for obtaining correspondences, camera projection matrices and 3D structure. The triplets may then be sewn together to establish correspondences, projection matrices and structure for the entire sequence [2], [6]. The correspondence method will be illustrated on the corridor sequence of figure 1.

A. Matching for view pairs

Correspondences are first determined between all consecutive pairs of frames as follows. An interest-point operator [8] extracts point features ("corners") from each frame of the sequence. Putative correspondences are generated between pairs of frames based on cross-correlation of interest point neighbourhoods and search windows. Matches are then established from this set of putative correspondences by simultaneously estimating epipolar geometry and matches consistent with this estimated geometry. The estimation algorithm is robust to mismatches and is described in detail in [25], [27], [28]. This basic level of tracking is termed the F-Based Tracker ("F" for fundamental matrix).

Typical results. Typically the number of corners used in a 768×576 image of an indoor scene is about 500, the number of seed matches is about 200, and the final number of matches is about 250. Using corners computed to sub-pixel accuracy, the average distance of a point from its epipolar line is $\sim$0.2 pixels.

The robust nature of the estimation algorithms means that it is not necessary to restrict putative correspondences to nearest neighbours or even the highest cross-correlation match, as the rigidity constraint can be used to select the best match from a set of candidates. Typically the radius of the search window for candidate matches is 10–20% of the image size, which adequately covers image point motion for most sequences.

B. Matching for view triplets

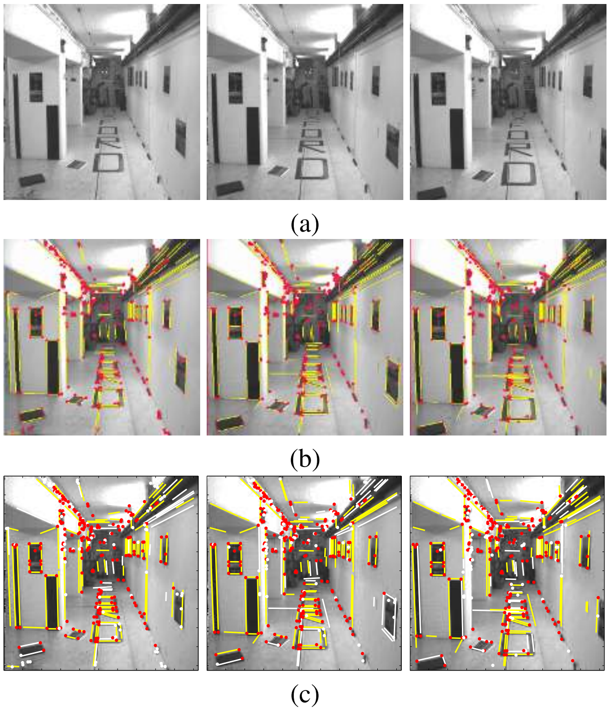
Correspondences are then determined between all consecutive triplets of frames. The 3-view matches are drawn from the 2-view matches provided by the F-Based Tracker. Although a proportion of these 2-view matches are erroneous (outliers), many of these mismatches are removed during the simultaneous robust estimation of the trifocal tensor and consistent matches [26]. The trifocal geometry provides a more powerful disambiguation constraint than epipolar geometry because image position is completely determined in a third view, given a match in the other two views, whereas image position is only restricted to a line by the epipolar geometry between two views.

The output at this stage of matching consists of sets of overlapping image triplets. Each triplet has an associated trifocal tensor and 3-view point matches. The camera matrices for the 3-views may be instantiated from the trifocal tensor [10], and 3D points instantiated for each 3-view point match by minimizing reprojection error over the triplet.

Typical results. Typically the number of seed matches over a triplet is about 100 corners. The final number of matches is about 180. Using corners computed to sub-pixel accuracy, the typical distance of a corner from its transferred position is $\sim$ 0.2 pixels. An example is shown in figure 2a.

C. Matching lines over view triplets

Line matching is notoriously difficult over image pairs as there is no geometric constraint equivalent to the fundamental matrix for point correspondences. However, over 3 views a geometric constraint is provided by the trifocal tensor computed as above from point correspondences.



(a)

(b)

(c)

Fig. 2. **Image triplet processing**: The workhorse of the system, converting a passive, uncalibrated, camera into a sparse range sensor. (a) The first three images of the corridor sequence. (b) Point (white) and line (grey) features extracted from the sequence. (c) features matched across these three views.

Line segments are matched over the triplet in two stages. First, given the trifocal tensor and putatively corresponding lines in two images, the corresponding line in the third image is determined. A line segment should be detected at the predicted position in the third image for a match to be instantiated. Second, the match is verified by a photometric test based on correlation of the line's intensity neighbourhood. The point to point correspondence for this correlation is provided by the computed epipolar geometry. Details are given in [18].

Typical results. Typically there are 200 lines in each image and a third of these are matched over the triplet. The line transfer error is generally less than a pixel. In practice the two stages of verification eliminate all but a couple of mismatches. An example is shown in figure 2b.

D. *Matching for sequences*

Correspondences are extended over many frames by merging 3-view point matches for overlapping triplets [6], [12]. For example a correspondence which exists across the triplet 1-2-3 and also across the triplet 2-3-4 may be extended to the frames 1-2-3-4, since the pair 2-3 overlaps for the triplets. The camera matrices and 3D structure are then computed for the frames 1-2-3-4. This process is extended by merging neighbouring groups of frames until camera matrices and correspondences are established throughout the sequence. At any stage the available cameras and structure can be used to guide matching over any frame of the sequence. The initial estimate of 3D points and cameras for a sequence is refined by a hierarchical bundle adjustment [6], [21]. Finally, the projective coordinate system is transformed to Euclidean (less overall scale) by autocalibration [5], [16].

In this manner structure and cameras may be computed automatically for sequences consisting of hundreds of frames. Examples are given in the following section.

III. Results

Some example sequences are shown in figure 3, each of which particularly exercise different aspects of the system. First the sequences are discussed, with the points of note being identified, and then two applications of the system are presented, with reference to the example sequences.

A. *Corridor sequence*

A camera is mounted on a mobile vehicle for this sequence. The vehicle moves along the floor turning to the left. The forward translation in this sequence makes structure recovery difficult, due to the small baseline for triangulation. In this situation, the benefit of using all frames in the sequence is significant. Figure 1 shows the recovered structure.

B. *Dinosaur sequence*

In this sequence, the model dinosaur is rotated on a turntable so that effectively the camera circumnavigates the object. Feature extraction is performed on the luminance component of the colour signal. No reliable lines are extracted on this object so only points are used. In this case, the additional constraint that the motion is known to be circular is applied, resulting in improved structure fidelity. Although the angle of rotation was known to be precisely $10° \pm 0.005°$, this information was not supplied to the system in order to gain a measure of accuracy. The recovered RMS difference from $10°$ was $0.04°$, or approximately 1 milliradian. Figure 4 shows the recovered point structure and cameras.

C. *Castle sequence*

This sequence is taken with a standard SLR camera, by a cameraman walking around the grounds of a Belgian castle. The images are digitized to PAL resolution. There is significant lighting variation between the first and final frames, and the sequence contains non-rigid components (passing pedestrians and moving trees). Figure 5 shows that structure and motion are successfully recovered despite these impediments.

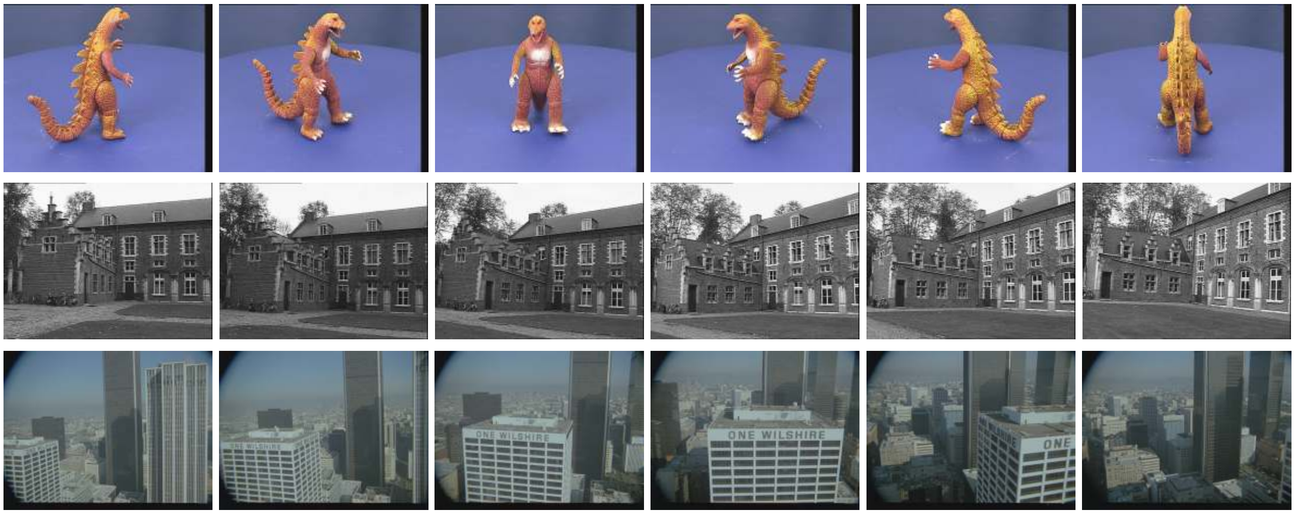


Fig. 3. **Example sequences**: **Dinosaur**, fixed camera, object on turntable (36 frames); **Castle**, hand-held camera (25 frames); **Wilshire**, camera in helicopter (350 frames).

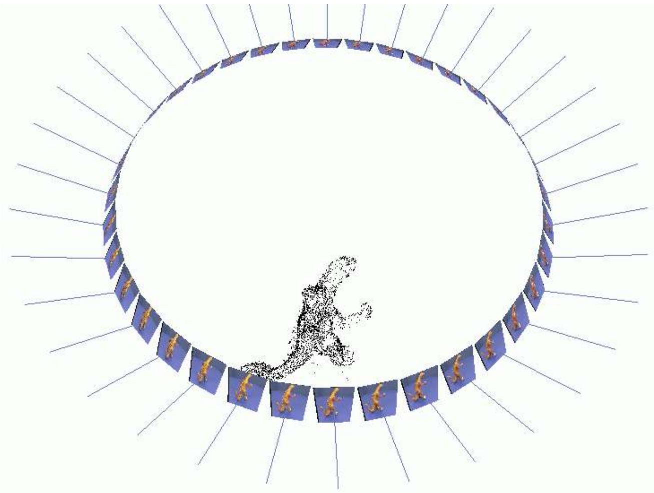


Fig. 4. **Dinosaur:** 3D point structure and camera positions for the Dinosaur sequence.

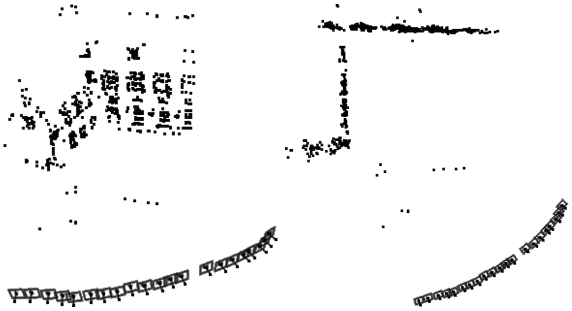


Fig. 5. **Castle:** Computed cameras and 3D point structure. The plan view shows the accuracy of the self calibration.

D. *"Wilshire" sequence*

The final sequence is a helicopter shot of Wilshire Boulevard, Los Angeles. In this case reconstruction is hampered by the repeated structure in the scene—many of the feature points (for example those on the skyscraper windows) have very similar intensity neighbourhoods, so correlation-based tracking produces many false candidates. However, the robust geometry-guided matching (§II-A) successfully rejects the incorrect correspondences. Figure 6 shows the structure.

IV. Applications

The previous sections have described the core camera-and-structure recovery system, and we now develop two applications which use this information.

A. *Construction of Virtual-Reality Models*

Having the complete point and line structure, we now describe how to convert the sparse 3D features into a form suitable for graphical rendering.

To produce triangulated structure for the polyhedral examples in this paper, planes are automatically extracted from the 3D data using the RANSAC technique: random 3-point subsets of the data are selected to define planes, and the number of 3D points which are less than a user-specified distance from each plane are counted. The plane with the greatest number of consistent points is stored, and the data points which were consistent with it removed from the structure. Repeating this process extracts the largest planes from the dataset, and the process is terminated when the required number of planes have been found.

The RANSAC procedure, by its nature, will ignore small-scale structure in the data, but is an ideal starting

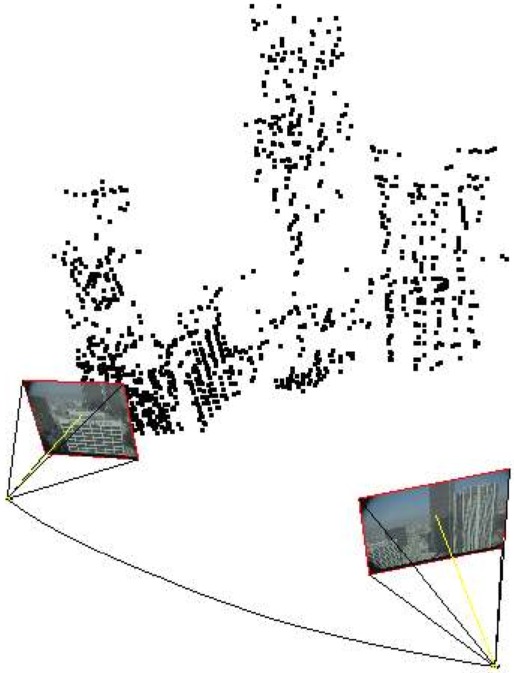


Fig. 6. **Wilshire:** 3D points and cameras for 350 frames of a helicopter shot. Cameras are shown for just the start and end frames for clarity, with the camera path plotted between.

point for photogrammetric techniques such as the Debevec *et al.* architectural system [3].

The planes are textured by selecting (automatically) the image from the sequence which is most fronto-parallel to that plane, and then texture mapping from the appropriate polygonal image region. As the texture mapping from the image to the plane is via an affine transformation, it is necessary to first warp the image to remove any projective distortion. Again this correction is automatic. Figure 1 shows the final texture-mapped model.

For the non-polyhedral objects, the surface extraction problem is more difficult, mainly due to the sparsity of the data. However, the dinosaur sequence is easily approached by segmenting the (blue) background and intersecting the generalized cones formed by the occluding contours (or silhouettes), and results are shown in figure 7.

B. Augmented reality

Because the system automatically determines the camera position for each view, it is possible to render computer-generated objects as if they are part of the scene. This process, generally known as *augmented* reality is of great importance in cinematic special effects. Figures 8 and 9 demonstrate this process on two of the example sequences. In figure 8, planar surfaces are identified in the 3D structure, and then an image is transformed via the implied 2D perspective transformation such that it appears to be attached to the plane. Figure 9 demonstrates the use of the recovered 3D structure of the scene for depth-keying. The cage around the object is rendered into a Z-buffer which is initialized using the 3D model, so that the bars behind the object are correctly occluded, and those in front correctly occlude the object.

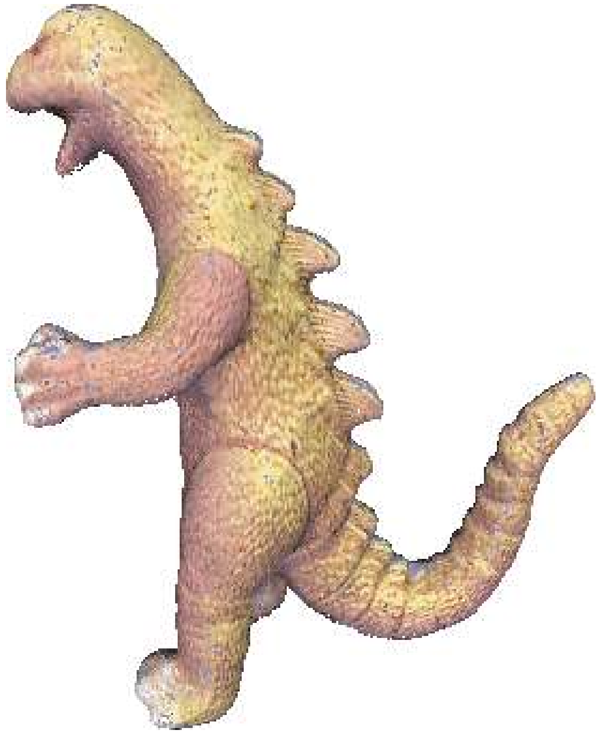


Fig. 7. **Dinosaur**: Reconstruction from occluding contours.

V. FUTURE DEVELOPMENTS

We have presented a system that will take sequences of images from an uncalibrated camera or cameras, and will automatically recover camera positions and 3D point and line structure from these sequences. We are currently extending the core system to include space curves [19], improve the automatic plane extraction [1], and cope with wide baselines between frames [17].

The system can be used as a pre-process to a number of applications. We have demonstrated VRML construction and Augmented Reality here. Other applications include: building a lumigraph or light field rendering [7], [13], image sequence compression, and frame interpolation.

Acknowledgements

We are grateful for permission to use the castle sequence supplied by the University of Leuven, the dinosaur sequence supplied by the University of Hannover, and the Wilshire sequence supplied by Oxford Metrics.

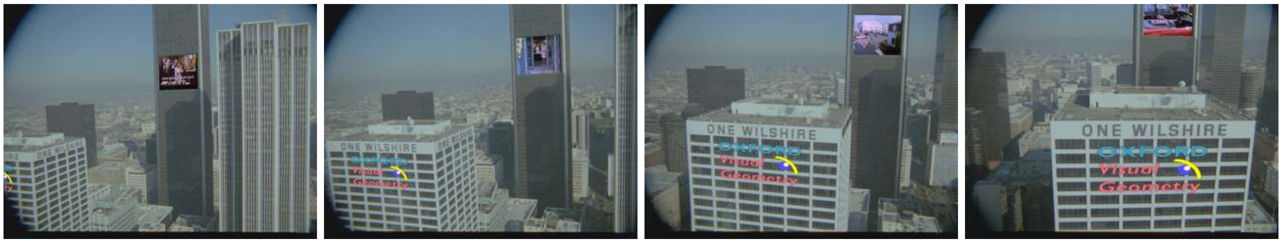


Fig. 8. **Wilshire**: Augmented reality. Planar surfaces are identified in 3D, then 2D homographies are computed which map the augmenting images onto the planes. Note that images appear to be rigidly attached onto two skyscrapers.

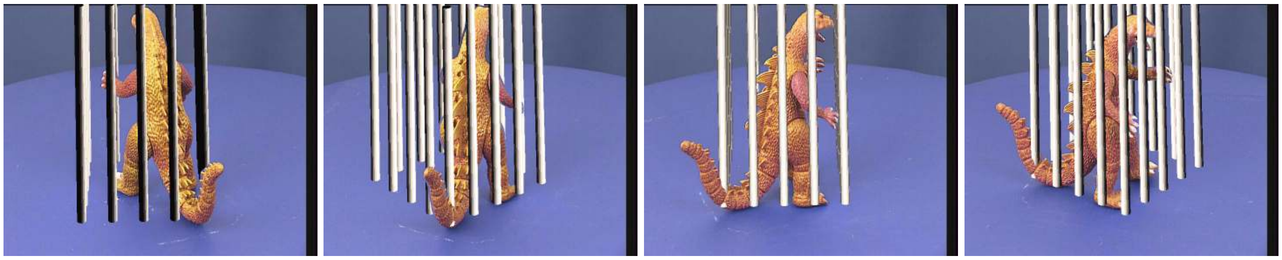


Fig. 9. **Dinosaur**: Augmented reality. The recovered 3D structure is used to depth-key the cage, which then correctly occludes the model.

Financial support was provided by the EC ACTS project VANGUARD.

REFERENCES

[1] C. Baillard and A. Zisserman. Automatic reconstruction of piecewise planar models from multiple views. In *Proc. CVPR*, June 1999. (to appear).

[2] P. Beardsley, P. Torr, and A. Zisserman. 3D model acquisition from extended image sequences. In *Proc. ECCV*, LNCS 1064/1065, pages 683–695. Springer-Verlag, 1996.

[3] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Proceedings, ACM SIGGRAPH*, pages 11–20, 1996.

[4] O. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *Proc. ECCV*, LNCS 588, pages 563–578. Springer-Verlag, 1992.

[5] O. Faugeras, Q. Luong, and S. Maybank. Camera self-calibration: Theory and experiments. In *Proc. ECCV*, LNCS 588, pages 321–334. Springer-Verlag, 1992.

[6] A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *Proc. ECCV*, pages 311–326. Springer-Verlag, June 1998.

[7] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. In *SIGGRAPH96*, 1996.

[8] C. J. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conf.*, pages 147–151, 1988.

[9] R. I. Hartley. A linear method for reconstruction from lines and points. In *Proc. ICCV*, pages 882–887, 1995.

[10] R. I. Hartley. Lines and points in three views and the trifocal tensor. *IJCV*, 22(2):125–140, 1997.

[11] R. I. Hartley, R. Gupta, and T. Chang. Stereo from uncalibrated cameras. In *Proc. CVPR*, 1992.

[12] S. Laveau. *Géométrie d'un système de N caméras. Théorie, estimation et applications.* PhD thesis, INRIA, 1996.

[13] M. Levoy and P. Hanrahan. Light field rendering. In *SIGGRAPH96*, 1996.

[14] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[15] P. Milgram, S. Shumin, D. Drascic, and J. Grodski. Applications of augmented reality for human-robot communication. In *International Conference on Intelligent Robots and Systems Proceedings, Yokohama, Japan*, pages 1467–1472, 1993.

[16] 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.

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

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

[19] C. Schmid and A. Zisserman. The geometry and matching of curves in multiple views. In *Proc. ECCV*, pages 394–409. Springer-Verlag, June 1998.

[20] A. Shashua. Trilinearity in visual recognition by alignment. In *Proc. ECCV*, volume 1, pages 479–484, May 1994.

[21] C. Slama. *Manual of Photogrammetry.* American Society of Photogrammetry, Falls Church, VA, USA, 4th edition, 1980.

[22] M. E. Spetsakis and J. Aloimonos. Structure from motion using line correspondences. *IJCV*, 4(3):171–183, 1990.

[23] M. E. Spetsakis and J. Aloimonos. A multi-frame approach to visual motion perception. *IJCV*, 16(3):245–255, 1991.

[24] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *IJCV*, 9(2):137–154, November 1992.

[25] P. H. S. Torr and D. W. Murray. Statistical detection of independent movement from a moving camera. *Image and Vision Computing*, 1(4):180–187, May 1993.

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

[27] P. H. S. Torr and A. Zisserman. Robust computation and parameterization of multiple view relations. In *Proc. ICCV*, pages 727–732, January 1998.

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