PAPER *Special Issue on Machine Vision Applications*

Reconstruction of Architectural Scenes from Uncalibrated Photos and Maps*

Ignazio INFANTINO[†], **Roberto CIPOLLA**[††], ***Nonmembers,***
and* Antonio CHELLA**[†], ***Regular Member

SUMMARY We consider the problem of reconstructing architectural scenes from multiple photographs taken from arbitrary viewpoints. The original contribution is the use of a map as a source of geometric constraints to obtain in a fast and simple way a detailed model of a scene. We suppose that images are uncalibrated and have at least one planar structure as a façade for exploiting the planar homography induced between world plane and image to calculate a first estimation of the projection matrix. Estimations are improved by using correspondences between images and map. We show how these simple constraints can be used to calibrate the cameras and recover the projection matrices for each viewpoint. Finally, triangulation is used to recover 3D models of the scene and to visualise new viewpoints. Our approach needs minimal a priori information about the camera being used. A working system has been designed and implemented to allow the user to interactively build a model from uncalibrated images from arbitrary viewpoints and a simple map.

key words: *3D reconstruction, planar homography, self-calibration*

1. Introduction

The aim of this paper is to describe a system able to reconstruct architectural sites from uncalibrated images and using information from the facades and maps of buildings. Many approaches exist to attempt to recover 3D models from calibrated stereo images [16] or uncalibrated extended image sequences [1], [18], [23] by triangulation and exploiting epipolar [15] and trilinear constraints [10]. Other approaches consist of visualisation from image-based representations of a 3D scene. This approach has been successfully used to generate an intermediate viewpoint image given two nearby viewpoints and has the advantage that it is does to need to make explicit a 3D model of the scene [7], [8], [19], [21], [22]. Constructions of 3D model from a collection of panoramic image mosaics and geometrical constraints have also presented [12], [20].

We adopt a simple approach to build a 3D model by exploiting strong constraints present in the scenes to be modelled [3], [5], [14]. The constraints which are used are parallelism and orthogonality, leading to simple and geometrically intuitive methods to calibrate the intrinsic and extrinsic parameter of the cameras and to recover Euclidean models of the scene from only two images from arbitrary positions. We propose an extension to the model presented in [3] dealing with the problem of recovering 3D models from uncalibrated images of architectural scenes viewing façades by exploiting the planar homography induced between world and image and using a simple map. Many map-based approaches are presented in extracting object information from aerial images for analizyng complex urban scenes (for example [17]). We investigate the use of the map in performing reconstruction of buildings and surroundings of a delimitated urban area using few images of it.

In the following, this paper describes our approach for the reconstruction of architectural scene from uncalibrated photos and map. First, we explain the principal steps of the modelling system. We then describe the self-calibration method used and the process of estimation and improvement of projection matrices of the different views. Finally, we include map constraints improving the estimation of the 3D model.

2. The Modelling System

Our modelling system uses one or more pairs of images of a façade. For each pair, the user indicates line, point or plane correspondences between images and the map. The modelling system attempts to use all possible constraints in a consistent and coherent way. The goal is reached by decomposing the process into several linear steps.

For a single view we perform the following steps:

1. Recovering the camera intrinsic parameters and rotation $\mathbf{R}$ from two vanishing points;
2. Recovering camera translation $\mathbf{t}$ from two known points on the map;
3. Rectification of the image based on the homography induced in the image of the planar structure.

In this way we obtain a first approximation of the projection matrix and the texture maps of planar structures [13]. These are directly placed in a 3D model by the map correspondences. To have a better estimation

Manuscript received February 19, 2001.
Manuscript revised June 18, 2001.

[†] The authors are with Department of Computer Science and Automation of the Faculty of Engineering of the University of Palermo and CERE-CNR, Italy.

[††] The author is with the Faculty of Engineering of the University of Cambridge, UK.

[*] Preliminary version of this paper was presented at IAPR Workshop on Machine Vision Application (MVA2000).

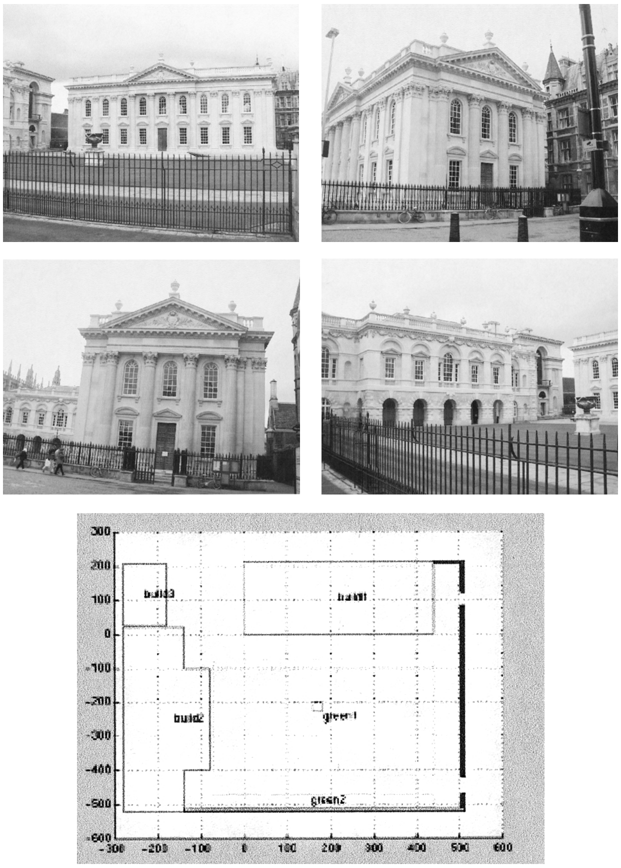


Fig. 1 Four of the ten input images of an architectural scene and its map.

of the projection matrix we use two strategies. The first one exploits two images viewing the same planar structure [22] and it deals with new homography estimation [11] by automatically finding corner correspondences. Decomposition of inter-frame homography matrix gives the new estimation of the projection matrices of two views (Fig. 1).

The second one involves a global refinement using all possible map constraints. It allows reconstruction of points out of the planes used for calibration and rectification. The final steps are:

1. Global optimisation using map constraints (bundle adjustment);
2. Triangulation and raw 3D model;
3. Texture mapping;
4. Export a VRML model.

3. Calibration and Estimation of the Projection Matrix

With respect to a pin-hole camera, perspective projection from Euclidean 3D space to an image can be represented in homogeneous coordinates by a 3×4 camera projection matrix $\mathbf{P}$ [6]:

$$\lambda_{\mathbf{i}} \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \mathbf{P} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix} \tag{1}$$

The projection matrix has 11 degrees of freedom and can be decompose into a 3×3 orientation matrix $\mathbf{R}$ and a 3×1 translation vector $\mathbf{t}$ and a 3×3 camera calibration matrix, $\mathbf{K}$:

$$\mathbf{P} = \mathbf{K}\ [\ \mathbf{R}\ \mathbf{t}\ ] \tag{2}$$

From line correspondences given by the user, we can calculate vanishing points. In the worst case we only have information about two perpendicular direction, but it is enough to estimate the scale factor α of the image plane if we suppose that the principal point coordinates are equal to image centre. Then we calculate the camera calibration matrix $\mathbf{K}$ and the rotation matrix $\mathbf{R}$ for each image [2]. Moreover, we use two point correspondences between the map and the image (or a known length of a line on image) to obtain the translation vector $\mathbf{t}$. From a common planar structure in the pair of images we automatically improve the estimated matrices above by exploiting the homography [6]. Supposing points are on plane Z=0, we have:

$$\lambda_1\ \mathbf{w}_1 = \mathbf{K_1}\ [\mathbf{r_1^1}|\mathbf{r_1^2}|\mathbf{t_1}]\ \mathbf{X^P} \tag{3}$$

$$\lambda_2\ \mathbf{w}_2 = \mathbf{K_2}\ [\mathbf{r_2^1}|\mathbf{r_2^2}|\mathbf{t_2}]\ \mathbf{X^P} \tag{4}$$

where $\mathbf{r_1^1}, \mathbf{r_1^2}$ are the two first columns of $\mathbf{R_1}, \mathbf{r_2^1}, \mathbf{r_2^2}$ are columns of $\mathbf{R_2}$, and $\mathbf{X^P}$ is a 3×1 vector which denote point is on plane Π $(Z = 0)$ in homogeneus coordinates. Let be

$$\mathbf{H_1} = \mathbf{K_1}\ [\mathbf{r_1^1}|\mathbf{r_1^2}|\mathbf{t_1}] \tag{5}$$

$$\mathbf{H_2} = \mathbf{K_2}\ [\mathbf{r_2^1}|\mathbf{r_2^2}|\mathbf{t_2}] \tag{6}$$

and

$$\mathbf{H_{21}} = \mathbf{H_2}\ \mathbf{H_1^{-1}} \tag{7}$$

then

$$\lambda_2\ \mathbf{w}_2 = \lambda_1\ \mathbf{H_{21}}\ \mathbf{w}_1 \tag{8}$$

from which we obtain $\mathbf{w}_2 = [u_2\ v_2]^T$:

$$u_2 = (\mathbf{h_1}\ \mathbf{w_1})/(\mathbf{h_3}\ \mathbf{w_1}) \tag{9}$$

$$v_2 = (\mathbf{h_2}\ \mathbf{w_1})/(\mathbf{h_3}\ \mathbf{w_1}) \tag{10}$$

where $\mathbf{h_1}, \mathbf{h_2}, \mathbf{h_3}$ are rows of $\mathbf{H_{21}}$. By fixing for instance a façade, and using a Harris corners detector we find some correspondences between the two images. For each detected feature found in the first image we calculate where it is exactly in second image by the homography. Only the stronger matches are selected and we suppose them belonging to viewed plane. In this way we have new estimates of correspondences on

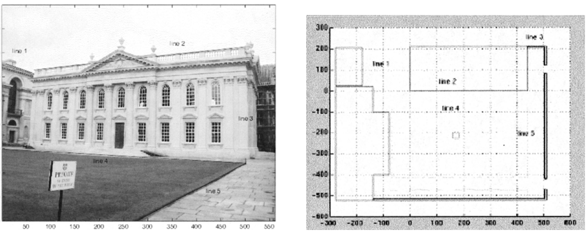


Fig. 2 Map constraint example: line to line correspondences.

second image and we improve estimate of homography. The decomposition of the homography matrix gives a new estimation of the inter-frame projection matrix. (See Fig. 2.)

$$\mathbf{H_{21}} = d_{12} * \mathbf{R_{21}} + \mathbf{t_{21}} \mathbf{n_{21}}^T \tag{11}$$

3.1 Plane Rectification

Points on the world plane are mapped to points on the image plane by a plane to plane homography, also known as a plane projective transformation. It is used for different pourpose such as object recognition, mosaicing and photogrammetry. A homography is described by 3 × 3 matrix H. Once this matrix is determined the back projection of an image point to a point on the world plane is straightforward. The distance between two points on the world plane is computed from the Euclidean distance between their back-projected images. We can obtain a new image of the plane in which the image plane is parallel to it, and same distance from it as the first image:

$$\mathbf{H_1} = \mathbf{K_1}\ [\mathbf{r_1^1} | \mathbf{r_1^2} | \mathbf{t_1}] \tag{12}$$

where $\mathbf{r_1^1}$,$\mathbf{r_1^2}$ are the two first columns of $\mathbf{R_1}$. If $\mathbf{X}_{ca} = -\mathbf{R_1}^{-1} * \mathbf{t_1}$; is camera position in world coordinate system that has $\mathbf{X_a}$ as origin, then distance d_{ca} is:

$$d_{ca} = \sqrt{\mathbf{X_{ca}}^T\ \mathbf{X_{ca}}}; \tag{13}$$

$$\mathbf{H_{par}} = \mathbf{K_1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -d_{ca} \end{bmatrix}; \tag{14}$$

$$\mathbf{H_{rect}} = \mathbf{H_{par}}\ \mathbf{H_1}^{-1} \tag{15}$$

Rectified image is used to obtain texture of the facade in order to construct a more realistic 3D model (Fig. 3).

4. Using Map Constraints

The map gives us important constraints to improve the estimation of the 3D model. We use the following constraints between geometric entities of image and map [21], [22]:

1. point to point correspondences
2. point to line correspondences
3. line to line correspondences

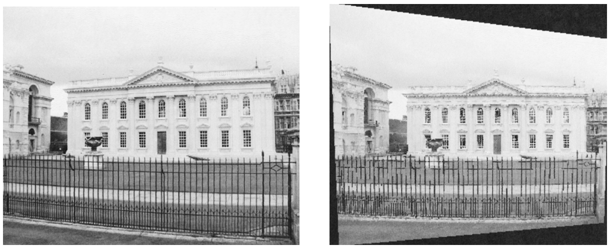


Fig. 3 Example of image rectification in order to have image plane and facade parallel.

For each image the user gives some correspondences to a map and the system tries to improve the projection matrix estimation. Point to point correspondences are strong constraints because they are directly related to the projection matrix:

$$\lambda_\mathbf{k} \mathbf{w_k} = \mathbf{P} \mathbf{X_k} \tag{16}$$

Point to line correspondence is expressed by equation:

$$\mathbf{L} \mathbf{P_{4\times4}^{-1}} \begin{bmatrix} \lambda \mathbf{w} \\ 1 \end{bmatrix} = \mathbf{0_{3\times1}} \tag{17}$$

where $\mathbf{L}$ is a 3×4 matrix of coefficients of 3 dimensional line equations and $\mathbf{P_{4\times4}^{-1}}$ is inverse matrix of projection matrix $\mathbf{P}$ with adding the last row [0 0 0 1]. 3D line through points $\mathbf{X_2}$ and $\mathbf{X_1}$ can be expressed by:

$$(\mathbf{X_2} - \mathbf{X_1}) \wedge \mathbf{X} + (\mathbf{X_2} \wedge \mathbf{X_1}) = 0 \tag{18}$$

Let us denote $a = x_2 - x_1$, $b = z_2 - z_1$, $c = y_2 - y_1$, $d_1 = z_1 y_2 - y_1 z_2$,$d_2 = -z_1 x_2 + x_1 z_2$, $d_3 = y_1 x_2 - x_1 y_2$; the line is described by a pair of equations from

$$\begin{aligned} by - cz + d_1 &= 0 \\ az - bx + d_2 &= 0 \\ cx - ay + d_3 &= 0 \end{aligned} \tag{19}$$

From equations system (17) we eliminate λ and use 2 constraints. Line to line correspondences are expressed as:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \mathbf{w_1} \wedge \mathbf{w_2} = (\mathbf{P} \mathbf{X_3}) \wedge (\mathbf{P} \mathbf{X_4}) \tag{20}$$

where a,b,c are coefficients of line equation $au + bv + c = 0$ on image passing trough generic points $\mathbf{w_1}$ and $\mathbf{w_2}$, and $\mathbf{X_{3i}}$,$\mathbf{X_{4i}}$ are two generic points of line map.

All constraints given by the user are used to define a minimisation problem in order to obtain new estimates of the projection matrices. Decomposing new matrices by QR decomposition we have also new estimates of camera calibration matrix $\mathbf{K}$, rotation matrix $\mathbf{R}$, translation vector $\mathbf{t}$ and camera position $\mathbf{X_c}$ of each image.

4.1 Minimizing Criterion

For point to point correspondence we can use algebraic error between image point given by user and image point which arises using projection matrix to improve. If we have n correspondences we minimize:

$$\sum_i \|\lambda_i \mathbf{w_i} - \mathbf{PX_i}\| \quad i = 1\dots n; \qquad (21)$$

Point to line correspondence minimizing criterion is given by:

$$\sum_i \left\| \mathbf{L_i P^{-1}_{4x4}} \left[\begin{array}{c} \lambda_i \mathbf{w_i} \\ 1 \end{array} \right] \right\| \quad i = 1\dots n; \qquad (22)$$

In the process used to estimate better projection matrix we use to introduce also point to point correspondences which arise from vanishing point calculated. In this way we have always at least two point-to-point correspondences, and we obtain robust reconstruction even user does not introduce others such correspondences. Line to line correspondences minimizing criterion is given by:

$$\sum_{i=1\dots n} \|\mathbf{w_{1i}} \wedge \mathbf{w_{2i}} - \lambda_i[(\mathbf{PX_{3i}}) \wedge (\mathbf{PX_{4i}})]\| \qquad (23)$$

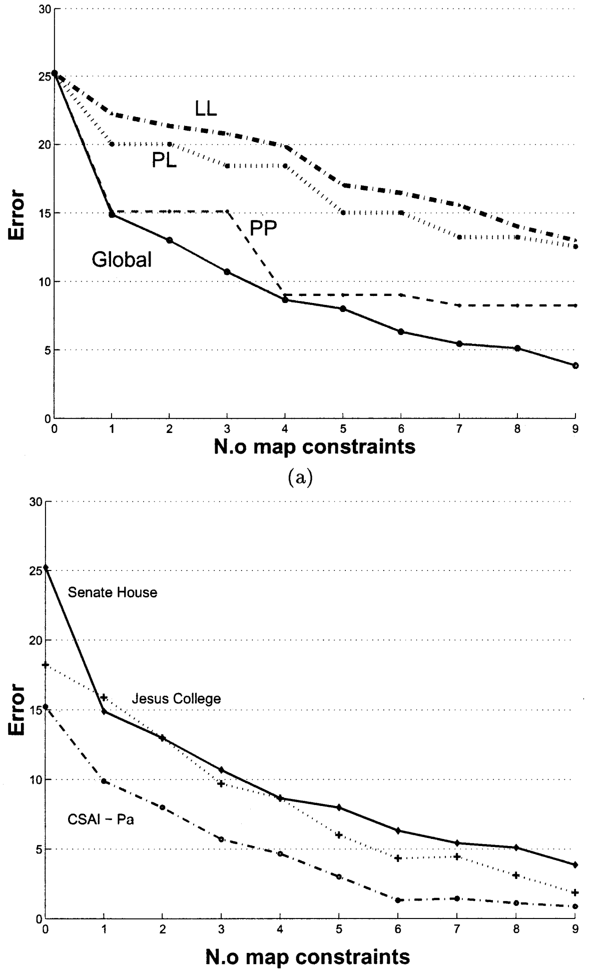
where $\mathbf{w_{1i}}$,$\mathbf{w_{2i}}$ are two generic points of line image and $\mathbf{X_{3i}}$,$\mathbf{X_{4i}}$ are two generic points of line map. All these criterions are used together with different normalisation weights.

4.2 Performance Analysis

Results of the experimental evaluation of the proposed method are shown in Fig. 4. The mean of the reprojection error is plotted against the number of map constraints used to compute the projection matrices. The graph (a) shows the results of different combinations of weights used in minimizing criterion. In particular, it is shown the contribution of each type of correspondences (PP point to point, PL point to line, LL line to line) and the results of the best combination of them ($a_{PP} = 0.5, a_{PL} = 0.25, a_{LL} = 0.25$ have given in every experiment the minimum of the error). The graph (b) shows the results of three different scenes including the Senate House, widely described in the paper. Both show the effectiveness of the method and the accuracy of estimating the projection matrices by introducing map-image correspondences.

5. Rendering and Texturing

The obtained 3D structure is rendered afterwards using a texture mapping procedure and the final model is stored in standard VRML 1.0 format (Fig. 5). In order to preserve the information given by primitive segments, we use a Delaunay triangulation on 2D texture (that is an image used for reconstruction).



(a)

(b)

Fig. 4 Results of the experimental evaluation: the graph (a) shows the results of different combinations of weights used in minimizing criterion; the graph (b) shows the results of three different scenes

6. Conclusion

The techniques presented have been successfully used to interactively build models of architectural scenes from pairs of uncalibrated photographs. Using information only from planar structure such as façades and a simple map, we recover precise projection matrices with only a few point correspondences.

References

[1] P. Beardsley, P. Torr, and A. Zisserman, "3D model acquisition from extended image sequences," Proc. 4th European Conf. on Computer Vision, Cambridge, April 1996; LNCS 1065, vol.II, pp.683–695, Springer-Verlag, 1996.

[2] B. Caprile and V. Torre, "Using vanishing points for camera calibration," IJCV, pp.127–140, 1990.

[3] R. Cipolla, T. Drummond, and D. Robertson, "Camera calibration from vanishing points in images of architectural scenes," Proc. British Machine Vision Conference, Notting-

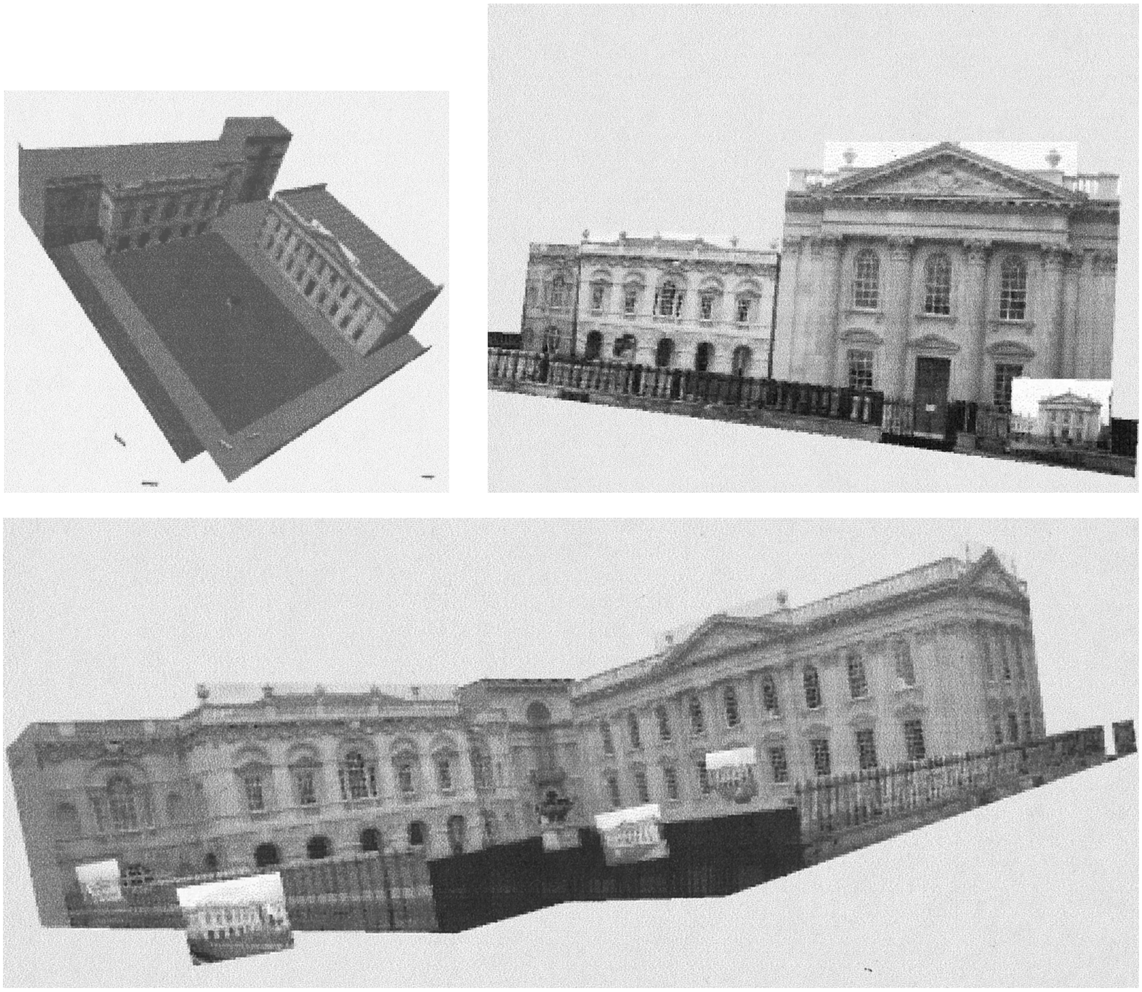


Fig. 5 Some views of VRML model of the scene.

ham, vol.2, pp.382–391, 1999.

[4] A. Criminisi, I. Reid, and A. Zisserman, "Single View Metrology," Proc. 7th International Conference on Computer Vision, pp.434–442, 1999.

[5] P.E. Debevec, C.J. Taylor, and J. Malik, "Modelling and rendering architecture from photographs: A hybrid geometry- and image-base approach," ACM Computer Graphics (Proc. SIGGRAPH), pp.11–20, 1996.

[6] O. Faugeras and F. Lustman, "Motion and structure from motion in a piecewise planar environment," Int. J. Pattern Recognition and Artificial Intelligence, vol.2, no.3, pp.485–508, 1988.

[7] O. Faugeras, S. Laveau, L. Robert, G. Csurka, and C. Zeller, "3-D reconstruction of urban scenes from sequences of images," Computer Vision and Image Understanding, no.69, vol.3, pp.292–309, 1998.

[8] S.J. Gortler, R. Grzeszczuk, R. Szeliski, and M.F. Cohen, "The Lumigraph," ACM Computer Graphics (Proc. SIGGRAPH), pp.31–42, 1996.

[9] R.I. Hartley, "Estimation of relative camera positions for uncalibrated cameras," European Conf. on Computer Vision, pp.579–587, Santa Margherita Ligure, Italy, May 1992.

[10] R.I. Hartley, "Lines and points in three views and the trifocal tensor," International J. Computer Vision, vol.22, no.2, pp.125–140, 1996.

[11] K. Kanatami, N. Ohta, and Y. Kanazawa, "Optimal homography computation with a reliabilty measure," IEICE Trans. Inf. & Syst., vol.E83-D, no.7, pp.1369–1374, July 2000.

[12] S.B. Kang and R. Szeliski, "3-D scene data recovery using omni-directional multibaseline stereo," CVPR'96, pp.364–370, June 1997.

[13] D. Liebowitz and A. Zisserman, "Metric rectification for perspective images of planes," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.482–488, 1998.

[14] D. Liebowitz, A. Criminisi, and A. Zisserman, "Creating architectural models from images," Eurographics'99, vol.18, no.3, 1999.

[15] Q.-T. Luong and T. Viéville, "Canonical representations for the geometries of multiple projective views," Computer Vision and Image Understanding, vol.64, no.2, pp.193–229, 1996.

[16] P.J. Narayanan, P.W. Rander, and T. Kanade, "Constructing virtual worlds using dense stereo," Proc. Sixth IEEE Intl. Conf. on Computer Vision, pp.3–10, Bombay, India, Jan. 1998.

[17] Y. Ogawa, K. Iwamura, and S. Kakumoto, "Exracting object information from arial images: A map-based approach," IEICE Trans. Inf. & Syst., vol.E83-D, no.7, pp.1450–1457, July 2000.

[18] M. Pollefeyes, R. Koch, and L. Van Gool, "Self calibration

and metric reconstruction inspite of varying an unknown internal camera parameters," Proc. Sixth IEEE Intl. Conf. on Computer Vision, pp.90–95, Bombay, India, Jan. 1998.

[19] S.M. Seitz and C.R. Dyer, "Toward image — Based scene representation using view morphing," Proc. Intl. Conf. IEEE Conf. on Pattern Recognition, Vienna, Austria, Jan. 1996.

[20] H-Y Shum, M. Han, and R. Szeliski, "Interactive construction of 3-D models from panoramic mosaics," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.427–433, Santa Barbara, USA, June 1998.

[21] R. Szeliski and S. Heung-Yeung, "Creating full view panoramic image mosaics and environment maps," SIGGRAPH, 1997.

[22] R. Szeliski and P. Torr, "Geometrically constrained structure from motion: Points on planes," European Workshop on 3D Structure from Multiple Images of Large-Scale Environments (SMILE), pp.171–186, June 1998.

[23] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorisation method," International J. Computer vision, vol.9, no.2, pp.137–154, 1990.

Ignazio Infantino obtained a Laurea Degree in Electronic Engineering in 1997 and his Ph.D. in Computer Science in 2001 from the University of Palermo, Italy. During November 1999–July 2001 he was a guest researcher in Computer Vision Group at the Department of Engineering, University of Cambridge, UK. Currently, he is a researcher at Computer Science and Artificial Intelligence Laboratory of the Department of Electrical Engineering of the University of Palermo. His research interests are in the field of computer vision, autonomous robotics, artificial vision, hybrid (symbolic/subsymbolic) systems and knowledge representation.

Roberto Cipolla obtained a B.A. (Engineering) from the University of Cambridge in 1984 and an M.S.E. (Electrical Engineering) from the University of Pennsylvania in 1985. From 1985 to 1988 he studied and worked in Japan he obtained an M.Eng. (Robotics) from the University of Electro-communications in Tokyo in 1988. In 1991 he was awarded a D.Phil. (Computer Vision) from the University of Oxford and from 1991-92 was a Toshiba Fellow and engineer at the Toshiba Corporation Research and Development Centre in Kawasaki, Japan. He joined the Department of Engineering, University of Cambridge in 1992 as a Lecturer and a Fellow of Jesus College. He became a Professor in 2000. His research interests are in computer vision and robotics.

Antonio Chella has received his laurea degree in Electronic Engineering in 1988 and his Ph.D. in Computer Science in 1993 from the University of Palermo, Italy. Currently, he is a Professor of Robotics at the University of Palermo. His research interests are in the field of autonomous robotics, artificial vision, neural networks, hybrid (symbolic/subsymbolic) systems and knowledge representation.