NAVIGATING IN MANHATTAN: 3D ORIENTATION FROM VIDEO WITHOUT CORRESPONDENCES

André T. Martins, Pedro M. Q. Aguiar[1], *Mário A. T. Figueiredo*[2]

Instituto Superior Técnico, Technical University of Lisbon
[1]Institute for Systems and Robotics, [2]Institute for Telecommunications
aftm@mega.ist.utl.pt, aguiar@isr.ist.utl.pt, mtf@lx.it.pt

ABSTRACT

The problem of inferring 3D orientation of a camera from video sequences has been mostly addressed by first computing correspondences of image features. This intermediate step is now seen as the main bottleneck of those approaches. In this paper, we propose a new 3D orientation estimation method for urban (indoor and outdoor) environments, which avoids correspondences between frames. The basic scene property exploited by our method is that many edges are oriented along three orthogonal directions; this is the recently introduced *Manhattan world* (MW) assumption.

In addition to the novel adoption of the MW assumption for video analysis, we introduce the *small rotation* (SR) assumption, that expresses the fact that the video camera undergoes a smooth 3D motion. Using these two assumptions, we build a probabilistic estimation approach. We demonstrate the performance of our method using real video sequences.

1. INTRODUCTION

Applications in areas such as digital video, virtual reality, mobile robotics, and visual aids for blind people, require efficient methods to estimate the 3D orientation of a video camera from the images it captures.

Most current approaches rely on an intermediate step that computes 2D displacements on the image plane. This 2D displacements are represented by either a dense map [1] or a set of correspondences between image feature points [2]. Rigidity assumptions are sometimes used to further constraint the solution. These assumptions lead to a nonlinear inverse problem relating the 3D structure with the 2D displacements, which has been addressed by using general non-linear optimization [3, 4, 5, 6], recursive Kalman-type estimation [4, 7, 8, 9], and matrix factorization [10, 11]. Other authors have used correspondences between line segments [12] or surface patches [13], rather than feature points.

While computing correspondences is feasible in simple cases, it is widely accepted that this is not so when processing real-life video sequences. This has motivated approaches that estimate the 3D structure directly from the image intensity values [14, 15]. These approaches lead to complex time-consuming algorithms.

The key assumption behind all the methods referred above is that some property of the scene, either the brightness pattern or the 3D positions of the features, remains (approximately) constant from frame to frame.

In this paper, we propose a new 3D orientation estimation method for urban (indoor and outdoor) environments, which **i)** does without computing the 2D-motion as an intermediate step, and **ii)** does not rely on scene constancy assumptions. The basic scene property exploited by our method is that many edges are oriented along three orthogonal directions; this is the so-called *Manhattan world* (MW) assumption recently introduced by Coughlan and Yuille [16, 17]. While these authors have used the prior knowledge captured by the MW assumption to build a Bayesian approach to 3D orientation estimation from a *single* image, we use it for *sequences* of images.

The main building blocks of the method herein proposed are: **i)** recent results in the geometry of 3D pose representation [18], which allow a significant reduction of the computational cost of the Bayesian estimation algorithm, and **ii)** a new *small rotation* (SR) model that expresses the fact that the video camera undergoes a smooth 3D motion.

2. CAMERA ORIENTATION IN VIDEO SEQUENCES

2.1. 3D Orientation and Vanishing Points

Let $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ and $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ be the Cartesian coordinate systems of the MW and the camera, respectively. We parameterize the camera orientation with three angles (see Fig. 1): α, the *compass* (azimuth) angle, corresponding to rotation about the $\mathbf{z}$ axis; β, the *elevation* angle above the $\mathbf{xy}$ plane; and γ, the *twist* about the principal axis. One thus denotes the camera orientation by $\mathbf{O}(\alpha, \beta, \gamma)$.

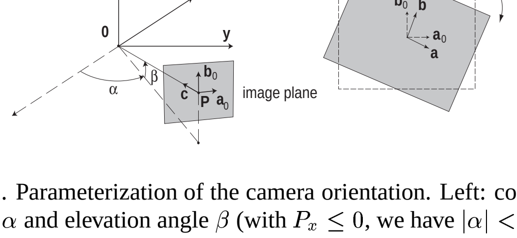

Fig. 1. Parameterization of the camera orientation. Left: compass angle α and elevation angle β (with $P_x \leq 0$, we have $|\alpha| < \pi/2$). Right: twist angle γ represented on the image plane.

This work was partially supported by the (Portuguese) Foundation for Science and Technology, grants POSI/SRI/41561/2001 and POSI/33143/SRI/2000.

For a radial-distortion-free pinhole camera, the three vanishing points, corresponding to the $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ axes, project on the image plane at, respectively,

$$\mathbf{F_x} = f\ \mathbf{R}_\gamma \left(\frac{\tan\alpha}{\cos\beta}, \tan\beta\right)^T \quad (1)$$

$$\mathbf{F_y} = f\ \mathbf{R}_\gamma \left(-\frac{\cot\alpha}{\cos\beta}, \tan\beta\right)^T \quad (2)$$

$$\mathbf{F_z} = f\ \mathbf{R}_\gamma \left(0, -\cot\beta\right)^T, \quad (3)$$

where $(\cdot)^T$ denotes transpose, f is the focal length, and $\mathbf{R}_\gamma$ is the *twist matrix*,

$$\mathbf{R}_\gamma = \begin{bmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{bmatrix}. \quad (4)$$

Estimating the orientation $\mathcal{O}(\alpha, \beta, \gamma)$ can thus be achieved by locating the vanishing points on the image plane [2].

2.2. Equiprojective Orientations

Let $\mathcal{F} = \{\mathbf{F_x}, \mathbf{F_y}, \mathbf{F_z}\}$ be the set of vanishing points. Since these vanishing points are mutually indistinguishable, it is the set $\mathcal{F}$, rather than each element *per se*, which provides the solution to the orientation estimation problem.

In [18], we have introduced the concept of *equiprojective orientations*: two orientations are termed *equiprojective* if and only if they lead to the same set of vanishing points. We have proved that *equiprojectivity* is an equivalence relation and that each corresponding equivalence class $\mathcal{E}_\mathcal{F}$ contains 12 elements, except for singular cases. The general solution to the orientation estimation problem is thus a 12-element set $\mathcal{E}_\mathcal{F}$ of possible camera orientations. We have also provided closed-form expressions for the solution set given any particular element; this turns out to be very important in reducing the space to be searched for a solution. Finally, we have proved that any equivalence class $\mathcal{E}_\mathcal{F}$ contains at least one orientation $\mathcal{O}(\alpha, \beta, \gamma)$ verifying

$$\alpha \in \left]-\frac{\pi}{4}, \frac{\pi}{4}\right], \quad \beta \in \left]-\frac{\pi}{4}, \frac{\pi}{4}\right], \quad \text{and} \quad \gamma \in \left]-\varphi, \varphi\right], \quad (5)$$

where

$$\varphi = \frac{\pi}{2} - \arctan\frac{\sqrt{2}}{2} \approx 54.7^\circ.$$

This means that, for any camera orientation, it is always possible to find one vanishing point inside the shaded area shown in Fig. 2.

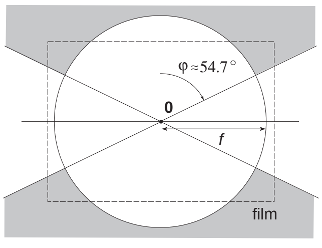

Fig. 2. Area of the image plane where it is guaranteed that there exists at least one vanishing point.

2.3. Small Rotations Model

Let us now assume that the camera is moving and acquiring a sequence of frames $\{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$. It is clear that the variation of the camera orientation only depends on the rotational component of the motion. Let $\mathcal{O}_k(\alpha_k, \beta_k, \gamma_k)$ denote the orientation at the k-th frame. In typical video sequences, the camera orientation evolves in a smooth continuous way. We formalize this property by introducing the *small rotations* (SR) model, next described.

Let $\mathcal{R}_k(\rho_k, \mathbf{e}_k)$ be the rotational component of the camera motion between the k-th and $(k+1)$-th frames, where ρ_k and $\mathbf{e}_k$ denote the angle and the axis of rotation, respectively. The SR model states, independently of $\mathbf{e}_k$, that there exists a small fixed angle ξ such that $|\rho_k| \leq \xi$. Here, we take $\xi = 5^\circ$, which implies that for a sampling rate of 12.5 Hz the rotation angle is less than 62.5° in each second; this is an intuitively reasonable assumption.

As a consequence of the SR assumption, we can show that the variations of the compass, elevation and twist angles between consecutive frames are bounded as follows:

$$|\alpha_{k+1} - \alpha_k| \leq \xi_\alpha \quad (6)$$

$$|\beta_{k+1} - \beta_k| \leq \xi \quad (7)$$

$$|\gamma_{k+1} - \gamma_k| \leq \xi, \quad (8)$$

where

$$\xi_\alpha = \arccos\left(1 - \frac{\cos|\beta_{k+1} - \beta_k| - \cos\xi}{\cos\beta_k \cos\beta_{k+1}}\right),$$

if $|\beta_k + \beta_{k+1}| \leq \pi - \xi$, and $\xi_\alpha = \pi/2$, otherwise.

For an orientation $\mathcal{O}_k$ in the region defined by (5), it is guaranteed that $|\beta_k + \beta_{k+1}| \leq \pi/4 + \pi/4 + \xi \leq \pi - \xi$. In particular, with $\xi = 5^\circ$, we have $\xi_\alpha \leq \arccos(2\cos 5^\circ - 1) \approx 7.1^\circ$. This enables a significant reduction of the search space for the estimate of $\mathcal{O}_{k+1}$, given the previous orientation $\mathcal{O}_k$.

3. ORIENTATION ESTIMATION

Our goal is to estimate the sequence of orientations $\{\mathcal{O}_1, \ldots, \mathcal{O}_N\}$ from the observed sequence of images $\{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$. This is done in a probabilistic estimation framework using the MW and SR assumptions.

3.1. Probabilistic Formulation

The MW assumption states that all images contain many edges consistent with the $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ axes [16, 17]. Hence, we use the statistics of the image intensity gradient $\nabla\mathbf{I}_k$ to extract this information. Non-relevant pixels are previously excluded by non-maxima suppression followed by thresholding of the gradient magnitude. Finally, the gradient magnitude is quantized, which allows building a discrete probabilistic model for edge strength. These pre-processing yields, for each image, a set of pairs $\mathbf{E_u} = (E_\mathbf{u}, \phi_\mathbf{u})$, where $E_\mathbf{u}$ is the quantized edge strength, and $\phi_\mathbf{u}$ the gradient direction, for each relevant pixel $\{\mathbf{u}\} = \{(u_1, u_2)\}$. The probabilistic formulation is built from the following elements:

Pixel classes: each pixel $\mathbf{u}$ has a label $m_\mathbf{u} \in \{1, ..., 5\}$ which indicates that $\mathbf{u}$ belongs to: (1) an edge consistent with the $\mathbf{x}$ axis; (2) an edge consistent with the $\mathbf{y}$ axis; (3) an edge consistent with the $\mathbf{z}$ axis; (4) an edge not consistent with either the $\mathbf{x}$, $\mathbf{y}$ or $\mathbf{z}$ axes; (5) the set of non-edge pixels. These classes have *a priori* probabilities $\{P_m(m_\mathbf{u}),\ m_\mathbf{u} = 1, ..., 5\}$.

Factorization assumption: at each pixel, the gradient magnitude and direction are independent. Moreover, the gradient magnitude is independent of the orientation and of the pixel location, given the class label. Thus,

$$P(\mathbf{E}_\mathbf{u}|m_\mathbf{u},\mathcal{O},\mathbf{u}) = P(E_\mathbf{u}|m_\mathbf{u})P(\phi_\mathbf{u}|m_\mathbf{u},\mathcal{O},\mathbf{u}), \qquad (9)$$

where $P(E_\mathbf{u}|m_\mathbf{u}) = P_{\text{on}}(E_\mathbf{u})$, if $m_\mathbf{u} \neq 5$, and $P(E_\mathbf{u}|m_\mathbf{u}) = P_{\text{off}}(E_\mathbf{u})$, if $m_\mathbf{u} = 5$. Here, $P_{\text{on}}(E_\mathbf{u})$ and $P_{\text{off}}(E_\mathbf{u})$ are the probability mass functions of the quantized gradient magnitude *conditioned* on whether pixel $\mathbf{u}$ is *on* or *off* an edge, respectively.

Decoupling: from (3), it is clear that the vanishing point $\mathbf{F_z}$ does not depend on the compass angle α. Thus, any estimation procedure can be decoupled into two stages: the first ignores information from the edges consistent with the $\mathbf{x}$ and $\mathbf{y}$ axes, and uses the remaining edges to estimate β and γ, a problem with 2 degrees of freedom (dof); the second stage estimates α, given the estimates of β and γ, which is a 1-dof problem. This decoupling avoids the 3-dof joint estimation of the three angles.

Gradient direction pdf: let $U(\cdot)$ be a uniform pdf on $]-\frac{\pi}{2},\frac{\pi}{2}]$. In the first stage (estimation of β and γ) we only consider edges consistent with the $\mathbf{z}$ axis ($m_\mathbf{u} = 3$), thus

$$P(\phi_\mathbf{u}|m_\mathbf{u},\beta,\gamma,\mathbf{u}) = \begin{cases} P_{ang}(\epsilon_\mathbf{u}) & \text{if } m_\mathbf{u} = 3 \\ U(\phi_\mathbf{u}) & \text{otherwise,} \end{cases} \qquad (10)$$

where

$$P_{ang}(t) = \begin{cases} (1-\epsilon)/(2\tau) & \text{if } t \in [-\tau,\tau] \\ \epsilon/(\pi - 2\tau) & \text{if } t \in\,]-\pi/2,-\tau[\, \cup\,]\tau,\pi/2], \end{cases}$$

$\epsilon = 0.1$, and $\tau = 4°$. In (10), $\epsilon_\mathbf{u} = \phi_\mathbf{u} - \theta_\mathbf{z}(\beta,\gamma,\mathbf{u})$ (mod-π) is the difference between the measured gradient direction and that that would be ideally observed at pixel $\mathbf{u}$, given β and γ.

In the second stage (estimating α with fixed $\widehat{\beta}$ and $\widehat{\gamma}$),

$$P(\phi_\mathbf{u}|m_\mathbf{u},\alpha,\widehat{\beta},\widehat{\gamma},\mathbf{u}) = \begin{cases} P_{ang}(\epsilon_\mathbf{u}) & \text{if } m_\mathbf{u} \in \{1,2,3\} \\ U(\phi_\mathbf{u}) & \text{if } m_\mathbf{u} \in \{4,5\}\,, \end{cases} \qquad (11)$$

where $\epsilon_\mathbf{u}$ has a similar meaning as in (10), now for $m_\mathbf{u} \in \{1,2,3\}$.

Joint likelihood: the joint likelihood for the relevant data of the k-th frame, $\{\mathbf{E}_\mathbf{u}\}$ (we omit the index k, for economy) is obtained by marginalizing (summing) over all possible models at each relevant pixel, and assuming independence among data from different pixels:

$$P(\{\mathbf{E}_\mathbf{u}\}|\mathcal{O}_k) = \prod_{\mathbf{u}} \sum_{m_\mathbf{u}=1}^{5} P(E_\mathbf{u}|m_\mathbf{u})\; P(\phi_\mathbf{u}|m_\mathbf{u},\mathcal{O}_k,\mathbf{u})\; P_M(m_\mathbf{u})\,. \qquad (12)$$

where $P(\phi_\mathbf{u}|m_\mathbf{u},\mathcal{O}_k,\mathbf{u})$ represents (10) and $\mathcal{O}_k$ stands for (β,γ), in the first stage, while $P(\phi_\mathbf{u}|m_\mathbf{u},\mathcal{O}_k,\mathbf{u})$ represents (11), and $\mathcal{O}_k$ stands for $(\alpha,\widehat{\beta},\widehat{\gamma})$, in the second stage.

3.2. Locating the Estimates

As explained above, the estimation procedure is decoupled into two stages. In the first stage, we find MAP estimates $\widehat{\beta}_k$ and $\widehat{\gamma}_k$,

$$\left(\widehat{\beta}_k,\widehat{\gamma}_k\right) = \arg\max_{\beta,\gamma} \log P(\{\mathbf{E}_\mathbf{u}\}_k|\beta,\gamma) + \log P(\beta,\gamma) \qquad (13)$$

where $P(\beta,\gamma)$ is a prior. In the first frame, this prior is flat over the entire domain $\beta \in\,]-45°,45°]$ and $\gamma \in\,]-54.7°,54.7°]$, where (5) guarantees the existence of a solution. For the other frames, $k > 1$, the prior expresses two assumptions: **i**) small rotations (see (7) and (8)), and **ii**) smoothness of the motion. These assumptions are formalized by taking $P(\beta,\gamma)$ as a truncated bivariate Gaussian with mean $[\widehat{\beta}_{k-1},\,\widehat{\gamma}_{k-1}]^T$, defined over the region $\beta \in\,]\widehat{\beta}_{k-1}-\xi,\widehat{\beta}_{k-1}+\xi]$ and $\gamma \in\,]\widehat{\gamma}_{k-1}-\xi,\widehat{\gamma}_{k-1}+\xi]$. The variance of this Gaussian allows controlling the tradeoff between the smoothness of the estimated sequence of angles and the accuracy of this estimates.

Given $\widehat{\beta}_k$ and $\widehat{\gamma}_k$, we then estimate the compass angle α_k:

$$\widehat{\alpha}_k = \arg\max_{\alpha} \log P(\{\mathbf{E}_\mathbf{u}\}|\alpha,\widehat{\beta}_k,\widehat{\gamma}_k) + \log P(\alpha).$$

For the first frame, the prior $P(\alpha)$ is flat over $]-45°,45°]$. For $k > 1$, as above, the prior is a truncated Gaussian with mean $\widehat{\alpha}_{k-1}$, defined over $]\widehat{\alpha}_{k-1}-\xi_\alpha,\widehat{\alpha}_{k-1}+\xi_\alpha]$.

If a given estimate $\widehat{\mathcal{O}}_k(\widehat{\alpha}_k,\widehat{\beta}_k,\widehat{\gamma}_k)$ is located outside of the minimal region defined in (5), we replace it by an equiprojective orientation inside that region. As explained in the last paragraph of subsection 2.3, this allows $|\xi_\alpha|$ to be less than 7.1°, hence keeping a small search space.

A consequence of the fact that we sometimes replace an estimate $\widehat{\mathcal{O}}_k$ by an element of its equiprojective equivalence class $\mathcal{E}(\widehat{\mathcal{O}}_k)$, is that the resulting sequence of estimates may not verify the SR assumption. Thus, as a final step, we pick an orientation from each equivalence class $\{\mathcal{E}(\widehat{\mathcal{O}}_k)\}$, such that the resulting sequence satisfies the SR model.

4. EXPERIMENTS

We tested our algorithm with outdoor MPEG-4 compressed video sequences, acquired with a hand-held camera. The sequences are of low quality due to some radial distortion and several over and underexposed frames. Our algorithm estimated successfully the orientations for the vast majority of the sequences tested. Typical running time for each (288×360)-pixels frame is less than 10 seconds, on a 1.5 GHz Pentium IV, using a straightforward non-optimized MATLAB implementation.

Figs. 3 and 4 show some frames of two sequences with superimposed cubes indicating the estimated orientations of the MW axes. Notice that the algorithm is able to estimate the correct orientation, despite the presence of many edges not aligned with the MW axes (*e.g.*, people in Fig. 4).

Fig. 5 plots the estimates of the orientation angles, for these two sequences. Note that the estimates on the right hand plot are slightly noisier than those on the left, due to the worse quality of the image sequence. Obviously, we can control the smoothness of these estimates by adjusting the prior variances referred in Subsection 3.2; here, these variances are the same for both sequences and all the three angles. Of course, there is a tradeoff between smoothness and the ability to accurately follow fast camera rotations.

5. CONCLUSIONS

We have proposed a probabilistic approach to estimating camera orientation from video sequences of urban scenes. The method avoids standard intermediate steps such as feature detection and

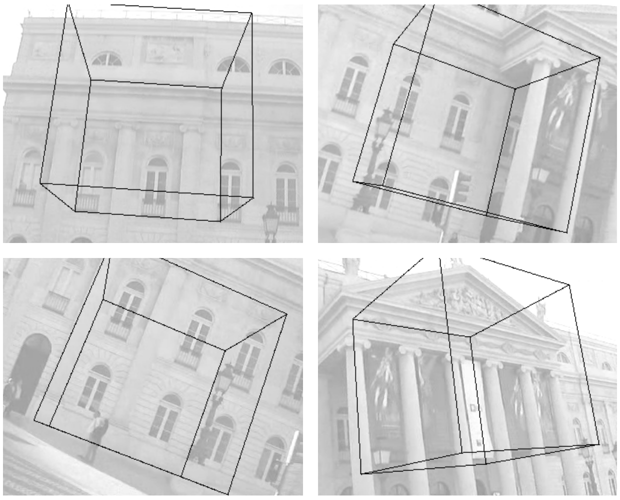

Fig. 3. Frames 20, 30, 40, and 50 of a video sequence with superimposed cubes representing the estimated orientation of the MW axes.

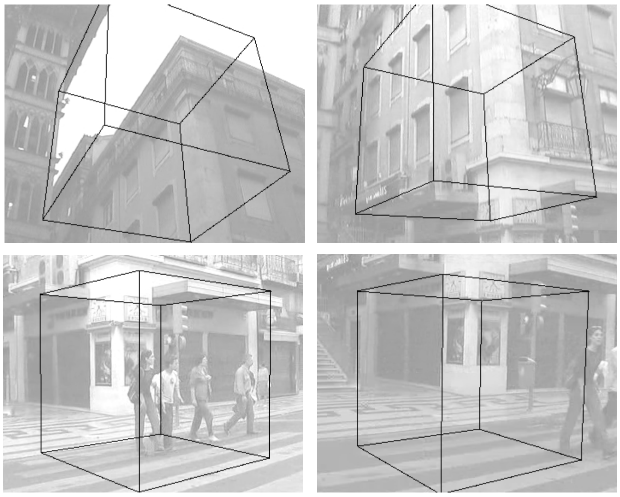

Fig. 4. As in Fig. 3, for frames 110, 130, 150, and 170 of another vide sequence.

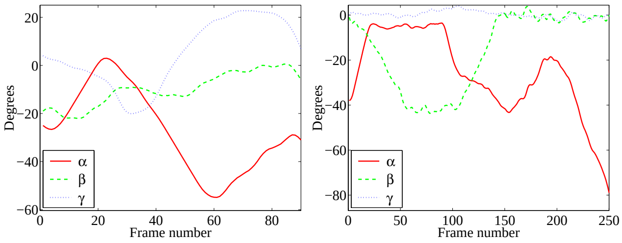

Fig. 5. Camera angle estimates, for the sequences of Figs. 3 and 4.

correspondence, or edge detection and linking. Experimental results show that the method is able to handle low-quality video sequences, even when many spurious edges are present.

6. REFERENCES

[1] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[2] O. Faugeras. *Three-Dimensional Computer Vision*. MIT Press, 1993.

[3] M. Spetsakis and Y. Aloimonos. A multi-frame approach to visual motion perception. *Int. Journal of Computer Vision*, 6(3):245–255, 1991.

[4] T. Broida and R. Chellappa. Estimating the kinematics and structure of a rigid object from a sequence of monocular images. *IEEE TPAMI*, 13(6), 1991.

[5] J. Weng, N. Ahuja, and T. Huang. Optimal motion and structure estimation. *IEEE TPAMI*, 15(9):864–884, 1993.

[6] R. Szeliski and S. Kang. Recovering 3D shape and motion from image streams using nonlinear least squares. *Jounal of Visual Comm. and Image Representation*, 5(1), 1994.

[7] A. Azarbayejani and A. Pentland. Recursive estimation of motion, structure, and focal length. *IEEE TPAMI*, 17(6), 1995.

[8] S. Soatto, P. Perona, R. Frezza, and G. Picci. Recursive motion and structure estimations with complete error characterization. In *Proc. of IEEE CVPR*, pages 428–433, 1993.

[9] J. Thomas, A. Hansen, and J. Oliensis. Understanding noise: The critical role of motion error in scene reconstruction. In *Proc. of IEEE ICCV*, pages 325–329, 1993.

[10] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. Journal of Computer Vision*, 9(2), 1992.

[11] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *European Conference on Computer Vision*, Cambridge, UK, 1996.

[12] L. Quan and T. Kanade. A factorization method for affine structure from line correspondences. In *Proc. of IEEE CVPR*, San Francisco, CA, USA, 1996.

[13] P. Aguiar and J. Moura. Three-dimensional modeling from two-dimensional video. *IEEE Trans. on Image Processing*, 10(10):1541–1551, 2001.

[14] B. Horn and E. Weldon Jr. Direct methods for recovering motion. *Int. Journal of Computer Vision*, 2(1):51–76, 1988.

[15] G. P. Stein and A. Shashua. Model-based brightness constraints: On direct estimation of structure and motion. *IEEE TPAMI*, 22(9):992–1015, 2000.

[16] J. Coughlan and A. Yuille. Manhattan world: Compass direction from a single image by Bayesian inference. In *IEEE ICCV*, Corfu, Greece, 1999.

[17] J. Coughlan and A. Yuille. The manhattan world assumption: Regularities in scene statistics which enable Bayesian inference. In *NIPS*, Denver CO, USA, 2000.

[18] A. Martins, P. Aguiar, and M. Figueiredo. Equivalence classes for camera orientation. Technical Report, Instituto Superior Técnico, Technical University of Lisbon, 2002.