Determining the Translational Speed of a Camera from Time-Varying Optical Flow

Anton van den Hengel, Wojciech Chojnacki, and Michael J. Brooks

School of Computer Science, Adelaide University, SA 5005, Australia
{anton,wojtek,mjb}@cs.adelaide.edu.au

Abstract. Under certain assumptions, a moving camera can be self-calibrated solely on the basis of instantaneous optical flow. However, due to a fundamental indeterminacy of scale, instantaneous optical flow is insufficient to determine the magnitude of the camera's translational velocity. This is equivalent to the baseline length indeterminacy encountered in conventional stereo self-calibration. In this paper we show that if the camera is calibrated in a certain weak sense, then, by using time-varying optical flow, the velocity of the camera may be uniquely determined relative to its initial velocity. This result enables the calculation of the camera's trajectory through the scene over time. A closed-form solution is presented in the continuous realm, and its discrete analogue is experimentally validated.

1 Introduction

It is well known that, under certain assumptions, a camera moving smoothly through a stationary environment can be self-calibrated based on instantaneous optical flow. However, because of a fundamental indeterminacy of scale, instantaneous optical flow is insufficient to determine the magnitude of the camera's instantaneous translational velocity.

The aim of this paper is to show that if a point in the static scene is tracked over a period of time as part of time-varying optical flow and if the camera is calibrated in a certain weak sense, then successive translational speeds of the camera evolve in a way that is uniquely determined by the camera's initial translational speed. A closed-form solution is presented in the continuous realm, and its discrete analogue is experimentally validated.

2 Camera and Image Settings

Consider a full-perspective, pinhole camera undergoing smooth motion in a stationary world. Associate with the camera a 3D coordinate frame such that:

- the frame's origin coincides with the camera's optical centre C,
- two basis vectors span the focal plane,
- the other basis vector coincides with the optical axis.

Let $\boldsymbol{v} = [v_1, v_2, v_3]^T$ and $\boldsymbol{\omega} = [\omega_1, \omega_2, \omega_3]^T$ specify the camera's instantaneous *translational velocity* and instantaneous *angular velocity* with respect to the camera frame.

B. Jähne et al. (Eds.): IWCM 2004, LNCS 3417, pp. 190–197, 2007.

Let P be a static point in space. Suppose that the vector connecting C with P has the coordinates $\boldsymbol{x} = [x_1, x_2, x_3]^T$ with respect to the camera frame. As the camera moves, the position of P relative to the camera frame will change and will be recorded in the function $t \mapsto \boldsymbol{x}(t)$. The evolution of the relative position is governed by the equation

$$\dot{\boldsymbol{x}} + \widehat{\boldsymbol{\omega}}\boldsymbol{x} + \boldsymbol{v} = \boldsymbol{0}, \tag{1}$$

where $\widehat{\boldsymbol{\omega}}$ is defined as

$$\widehat{\boldsymbol{\omega}} = \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}$$

(cf. [1]). Let f be the focal length of the camera. Let $\boldsymbol{p} = [p_1, p_2, p_3]^T$ be the vector of the coordinates of the perspective projection of P, through C, onto the image plane $\{\boldsymbol{x} \in \mathbb{R}^3 \colon x_3 = -f\}$, relative to the camera frame. Then

$$\boldsymbol{p} = -f\frac{\boldsymbol{x}}{x_3}. \tag{2}$$

To account for the geometry of the image, we introduce a separate, 2D coordinate frame in the image plane, with two basis vectors aligned with rows and columns of pixels, and with the origin located in one of the four corners of the rectangular image boundary. In the case of rectangular image pixels, it is natural to assume that each of the image frame basis vectors is proportional to one of the two vectors of the camera frame spanning the focal plane. The corresponding proportionality coefficients s_1 and s_2 characterise the pixel sizes in the basis directions, expressed in length units of the camera frame. If an image point has coordinates $\boldsymbol{p} = [p_1, p_2, -f]^T$ and $[m_1, m_2]^T$ relative to the camera and image frames, respectively, and if $\boldsymbol{m} = [m_1, m_2, 1]^T$, then

$$\boldsymbol{p} = \boldsymbol{A}\boldsymbol{m}, \tag{3}$$

where $\boldsymbol{A}$ is a 3×3 invertible matrix called the *intrinsic-parameter matrix*. With $[i_1, i_2]^T$ the coordinates of the principal point (at which the optical axis intersects the image plane) in the image frame, $\boldsymbol{A}$ takes the form

$$\boldsymbol{A} = \begin{bmatrix} s_1 & 0 & -s_1 i_1 \\ 0 & s_2 & -s_1 i_2 \\ 0 & 0 & -f \end{bmatrix}.$$

In particular, if pixels are square with unit length, then $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & -i_1 \\ 0 & 1 & -i_2 \\ 0 & 0 & -f \end{bmatrix}. \tag{4}$$

When pixels are non-rectangular, eq. (3) still applies but $\boldsymbol{A}$ takes a more complicated form accommodating an extra parameter that encodes shear in the camera axes (see [2, Section 3]).

With time t varying, the function $t \mapsto \boldsymbol{p}(t)$ describes the changing position of the image of P in the camera frame, and the function $t \mapsto \dot{\boldsymbol{p}}(t)$ records the rate of change.

Likewise, the function $t \mapsto \boldsymbol{m}(t)$ describes the changing position of the image of P in the image frame, and the function $t \mapsto \dot{\boldsymbol{m}}(t)$ records the corresponding rate of change.

For a given time instant t, any set of length 6 *flow vectors* $[\boldsymbol{m}(t)^T, \dot{\boldsymbol{m}}(t)^T]^T$—each vector corresponding to a point P from a set $\mathcal{P}$ of stationary points in the scene—is termed an *instantaneous true image velocity field*. Similarly, for a given time interval $[a, b]$, any set of trajectories of the form $t \mapsto [\boldsymbol{m}(t)^T, \dot{\boldsymbol{m}}(t)^T]^T$ with $t \in [a, b]$—each trajectory corresponding to a point P from a set $\mathcal{P}$ of stationary points in the scene—is called a *time-varying true image velocity field*. As is customary, we shall identify an image velocity field with an appropriate *observed image velocity field* or *optical flow field* (see [3, Chapter 12]).

The terms "instantaneous optical flow field" and "time-varying flow field" are employed even if the underlying set $\mathcal{P}$ contains a small number of elements. In particular, $\mathcal{P}$ can be reduced to a single point.

The velocities $\boldsymbol{v}$ and $\boldsymbol{\omega}$ are examples of so-called *extrinsic* parameters of the camera. Another example of an extrinsic parameter is the *projective form* $\boldsymbol{\pi}(\boldsymbol{v})$ of $\boldsymbol{v}$, defined as the composite ratio

$$\boldsymbol{\pi}(\boldsymbol{v}) = (v_1 : v_2 : v_3)$$

provided $\boldsymbol{v} \neq \boldsymbol{0}$. As is clear, $\boldsymbol{\pi}(\boldsymbol{v})$ captures the direction of $\boldsymbol{v}$.

A camera for which the intrinsic parameters (encoded in $\boldsymbol{A}$) and the extrinsic parameters $\boldsymbol{v}$ and $\boldsymbol{\omega}$ are known will be referred to as *strongly calibrated*. A camera for which the intrinsic parameters and the extrinsic parameters $\boldsymbol{\pi}(\boldsymbol{v})$ and $\boldsymbol{\omega}$ are known will be termed *weakly calibrated*. Strong calibration is typically performed using equipment external to the camera. In contrast, the process of *self-calibration* is carried out using purely image-based information and results in weak calibration (cf. [1, 4, 5]).

Assuming that $\boldsymbol{v} \neq \boldsymbol{0}$, the *focus of expansion* (FOE), or instantaneous epipole, of the image is the image point whose coordinate representation $\boldsymbol{p}$ in the camera frame is a multiple of $\boldsymbol{v}$—see [6, Subsection 12.3.1 C] for a discussion of certain subtleties involved in this definition. A moment's reflection reveals that an image point $\boldsymbol{p}$ is *not* the FOE if and only if

$$\widehat{\boldsymbol{v}}\boldsymbol{p} \neq \boldsymbol{0}. \tag{5}$$

3 Determining Relative Translational Speed

3.1 Theoretical Result

In reconstructing a scene from instantaneous optical flow seen by a weakly calibrated camera, the magnitude $\|\boldsymbol{v}\|$ of the translational velocity determines the scale of the reconstruction; here, of course, $\|\cdot\|$ denotes the Euclidean length of 3-vectors. It is not possible, however, to recover this velocity magnitude, or *translational speed*, from image-based information and a symbolic value must be chosen to set the reconstruction scale. The fact that any positive value may be selected reflects the inherent scale indeterminacy of the projective interpretation of optical flow [1].

Suppose that a time-varying optical flow field is given over a period of time $[a, b]$. It is *a priori* conceivable that the corresponding scale factor may change in an uncontrollable way from one time instant to another—see comments in [7, Section 9]. However, as we

show below, this indeterminacy can be significantly reduced if a single point in the static scene is tracked over the entire period $[a, b]$. A single trajectory in a time-varying optical flow field suffices to determine the *relative translational speed* $\|\boldsymbol{v}(t)\|/\|\boldsymbol{v}(a)\|$ for all $t \in [a, b]$, with the consequence that the speed $\|\boldsymbol{v}(t)\|$ is uniquely determined for each $t \in [a, b]$ once the initial speed $\|\boldsymbol{v}(a)\|$ is fixed.

The precise formulation of our result is as follows:

Theorem. *Assume that a weakly calibrated camera moves smoothly so that* $\boldsymbol{v}(t) \neq 0$ *for each* $t \in [a, b]$. *Suppose that, for each* $t \in [a, b]$, *a trajectory* $t \mapsto [\boldsymbol{m}(t)^T, \dot{\boldsymbol{m}}(t)^T]^T$ *represents a moving image of a point in the static scene. Suppose, moreover, that, for each* $t \in [a, b]$, $\boldsymbol{m}(t)$ *is not the FOE. Then the relative translational speed* $\|\boldsymbol{v}(t)\|/\|\boldsymbol{v}(a)\|$ *is uniquely determined for all* $t \in [a, b]$. *More specifically, there exists a function* $g\colon [a, b] \to \mathbb{R}$ *such that, for each* $t \in [a, b]$, $g(t)$ *is explicitly expressible in terms of* $\boldsymbol{A}(t)$, $\dot{\boldsymbol{A}}(t)$, $\boldsymbol{\pi}(\boldsymbol{v}(t))$, $\boldsymbol{\omega}(t)$, $\boldsymbol{m}(t)$ *and* $\dot{\boldsymbol{m}}(t)$, *and such that*

$$\frac{\|\boldsymbol{v}(t)\|}{\|\boldsymbol{v}(a)\|} = \exp\left(\int_a^t g(u)\,du\right). \tag{6}$$

Note. The exact form of g will be given in the course of the proof.

Proof. Given $t \in [a, b]$, we first find the value of $\boldsymbol{p}(t)$ by applying (3) to $\boldsymbol{m}(t)$. We next use the equation

$$\dot{\boldsymbol{p}} = \dot{\boldsymbol{A}}\boldsymbol{m} + \boldsymbol{A}\dot{\boldsymbol{m}},$$

obtained by differentiating both sides of (3), to determine $\dot{\boldsymbol{p}}(t)$ from $\boldsymbol{m}(t)$ and $\dot{\boldsymbol{m}}(t)$.

Note that (2) can be equivalently rewritten as

$$\boldsymbol{x} = -\frac{x_3\boldsymbol{p}}{f}. \tag{7}$$

Differentiating both sides of the latter equation, we obtain

$$\dot{\boldsymbol{x}} = \frac{x_3\dot{f} - \dot{x}_3 f}{f^2}\boldsymbol{p} - \frac{x_3}{f}\dot{\boldsymbol{p}}. \tag{8}$$

Substituting (7) and (8) into (1), we find that

$$x_3\big(\dot{f}\boldsymbol{p} - f(\dot{\boldsymbol{p}} + \widehat{\boldsymbol{\omega}}\boldsymbol{p})\big) - \dot{x}_3 f\boldsymbol{p} + f^2\boldsymbol{v} = \boldsymbol{0}. \tag{9}$$

Omitting in notation the dependence upon t, define

$$\begin{aligned}\boldsymbol{k} &= \widehat{\boldsymbol{v}}(\dot{f}\boldsymbol{p} - f(\dot{\boldsymbol{p}} + \widehat{\boldsymbol{\omega}}\boldsymbol{p})),\\ \boldsymbol{l} &= f\widehat{\boldsymbol{v}}\boldsymbol{p}.\end{aligned}$$

Applying $\widehat{\boldsymbol{v}}$ to both sides of (9) and taking into account that $\widehat{\boldsymbol{v}}\boldsymbol{v} = 0$, we see that

$$x_3\boldsymbol{k} - \dot{x}_3\boldsymbol{l} = 0.$$

Hence

$$\frac{\dot{x}_3}{x_3} = \frac{\boldsymbol{l}^T\boldsymbol{k}}{\|\boldsymbol{l}\|^2}. \tag{10}$$

For the last formula to be meaningful the denominators on both sides of (10) have to be non-zero. Without loss of generality we may always assume that $x_3 > 0$, as this assumption reflects the fact that the scene is in front of the camera. On the other hand, since $\boldsymbol{p}$ is not the FOE, it follows from (5) that $\boldsymbol{l} \neq 0$.

If $\boldsymbol{v}$ is multiplied by a non-zero scalar factor, then both $\boldsymbol{k}$ and $\boldsymbol{l}$ are multiplied by the same factor, and consequently $\boldsymbol{l}^T\boldsymbol{k}/\|\boldsymbol{l}\|^2$ does not change. As a result, $\boldsymbol{l}^T\boldsymbol{k}/\|\boldsymbol{l}\|^2$ can be regarded as a function of $\boldsymbol{\pi}(\boldsymbol{v})$—not just $\boldsymbol{v}$—and $\boldsymbol{\omega}$, f, $\dot{f}$, $\boldsymbol{p}$, $\dot{\boldsymbol{p}}$, and treated as known. Define

$$\boldsymbol{q} = \frac{1}{f^2}\left(\frac{\boldsymbol{l}^T\boldsymbol{k}}{\|\boldsymbol{l}\|^2} f\boldsymbol{p} - \dot{f}\boldsymbol{p} + f(\dot{\boldsymbol{p}} + \widehat{\boldsymbol{\omega}}\boldsymbol{p})\right).$$

Clearly, being a composite of known entities, $\boldsymbol{q}$ can be regarded as known. In view of (9) and (10), we have

$$\boldsymbol{v} = x_3\boldsymbol{q}$$

and further

$$\|\boldsymbol{v}\| = |x_3|\,\|\boldsymbol{q}\|. \tag{11}$$

To simplify the notation, let $v = \|\boldsymbol{v}\|$ and $q = \|\boldsymbol{q}\|$. Taking the logarithmic derivative of both sides of (11) and next using (10), we deduce that

$$\frac{\dot{v}}{v} = \frac{\dot{x}_3}{x_3} + \frac{\dot{q}}{q} = \frac{\boldsymbol{l}^T\boldsymbol{k}}{\|\boldsymbol{l}\|^2} + \frac{\dot{q}}{q}. \tag{12}$$

Let

$$g = \frac{\boldsymbol{l}^T\boldsymbol{k}}{\|\boldsymbol{l}\|^2} + \frac{\dot{q}}{q}.$$

Since $\boldsymbol{l}^T\boldsymbol{k}/\|\boldsymbol{l}\|^2$ is known and since q and $\dot{q}$ are known too (both functions being derivable from the known function $\boldsymbol{q}$), one can regard g as known. In view of (12), we finally find that

$$\frac{v(t)}{v(a)} = \exp\left(\int_a^t g(u)\,du\right),$$

which, of course, is the desired formula (6) for the relative translational speed. □

3.2 Computational Aspects

The result given above is applicable only in the case where the camera is weakly calibrated. There are many means by which this information may be recovered, and the method presented here will only be of interest in situations when this form of calibration information is available.

One method by which the required weak calibration of a moving camera may be achieved is as follows. If we assume that $\boldsymbol{A}$ takes the form given in (4) (so that pixels are square and have unit length) with the known principal point $[i_1, i_2]^T$, then the only unknown intrinsic parameters are the focal length and its derivative. These parameters, along with the required instantaneous velocities, can be estimated on the basis of an instantaneous optical flow field comprising at least eight elements $[\boldsymbol{m}_i^T, \dot{\boldsymbol{m}}_i^T]^T$ $(i \geq 8)$. A first step is the estimation, up to a common scalar factor, of two matrices, a symmetric

matrix $\boldsymbol{C}$ and an antisymmetric matrix $\boldsymbol{W}$, entering the *differential epipolar equation for uncalibrated optical flow*

$$\boldsymbol{m}^T \boldsymbol{W} \dot{\boldsymbol{m}} + \boldsymbol{m}^T \boldsymbol{C} \boldsymbol{m} = 0$$

and the *cubic constraint*[1]

$$\boldsymbol{w}^T \boldsymbol{C} \boldsymbol{w} = 0,$$

with $\boldsymbol{w}$ such that $\boldsymbol{W} = \widehat{\boldsymbol{w}}$. Once the composite ratio

$$\boldsymbol{\pi}(\boldsymbol{C}, \boldsymbol{W}) = (c_{11} : c_{12} : c_{13} : c_{22} : c_{23} : c_{33} : w_1 : w_2 : w_3)$$

is estimated, recovering f and $\dot{f}$ proceeds by exploiting explicit formulae that involve $\boldsymbol{\pi}(\boldsymbol{C}, \boldsymbol{W})$ [1]. Estimation of $\boldsymbol{\pi}(\boldsymbol{C}, \boldsymbol{W})$ can be done by applying one of a host of methods available [8].

To show how this approach works in practice, suppose that a set of at least eight points is tracked simultaneously over time. For each $i = 1, \ldots, I$ with $I \geq 8$, let $\{\boldsymbol{m}_i(t_j)\}_{j=1}^{n}$ be a sequence of images of the ith point in the scene, and let $\{\dot{\boldsymbol{m}}_i(t_j)\}_{j=1}^{n-1}$, $\dot{\boldsymbol{m}}_i(t_j) = (t_{j+1} - t_j)^{-1}(m_i(t_{j+1}) - m_i(t_j))$, be the sequence of corresponding image velocities. At each time instant t_j with $j < n$, the camera is first weakly calibrated based on the set $\{[\boldsymbol{m}_i(t_j)^T, \dot{\boldsymbol{m}}_i(t_j)^T]^T\}_{i=1}^{N}$ of all current flow vectors. Then, for each i, a value $g_i(t_j)$ is evolved based on the ith flow vector $[\boldsymbol{m}_i(t_j)^T, \dot{\boldsymbol{m}}_i(t_j)^T]^T$. In absence of information on the reliability of the $g_i(t_j)$, a simple average $\bar{g}(t_j) = I^{-1} \sum_{i=1}^{I} g_i(t_j)$ is next formed. Finally, the current relative translational velocity $\|\boldsymbol{v}(t_j)\|/\|\boldsymbol{v}(t_1)\|$ is updated to

$$\frac{\|\boldsymbol{v}(t_{j+1})\|}{\|\boldsymbol{v}(t_1)\|} = \frac{\|\boldsymbol{v}(t_j)\|}{\|\boldsymbol{v}(t_1)\|} \exp[(\bar{g}(t_j))(t_{j+1} - t_j)].$$

4 Experimental Evaluation

In order to assess the accuracy of the method presented above synthetic testing was carried out. The tests involved generating the equivalent of 5 seconds of video from a moving camera and comparing the estimated and true magnitudes of the instantaneous translational velocity. The 5 second duration was selected as a reasonable period over which it might be expected that all tracked points would remain visible within the boundaries of the image. The video has been generated at the equivalent of 25 frames per second, and the simulated motion of the camera is of the type that might be expected of a real camera.

An optical flow field of 100 elements was generated, and Gaussian noise (mean 0, standard deviation 0.5 pixels) added to both the location and velocity components of each flow vector. Given that optical flow is usually calculated across the majority of the image plane, it seems reasonable to assume that at least 100 flow vectors are available for velocity magnitude estimation. It is important to note also that in testing the method noise has not been added to the calibration information. The weak calibration used is

[1] Note that $\boldsymbol{w}^T \boldsymbol{C} \boldsymbol{w}$ is a cubic polynomial in the entries of $\boldsymbol{C}$ and $\boldsymbol{W}$.

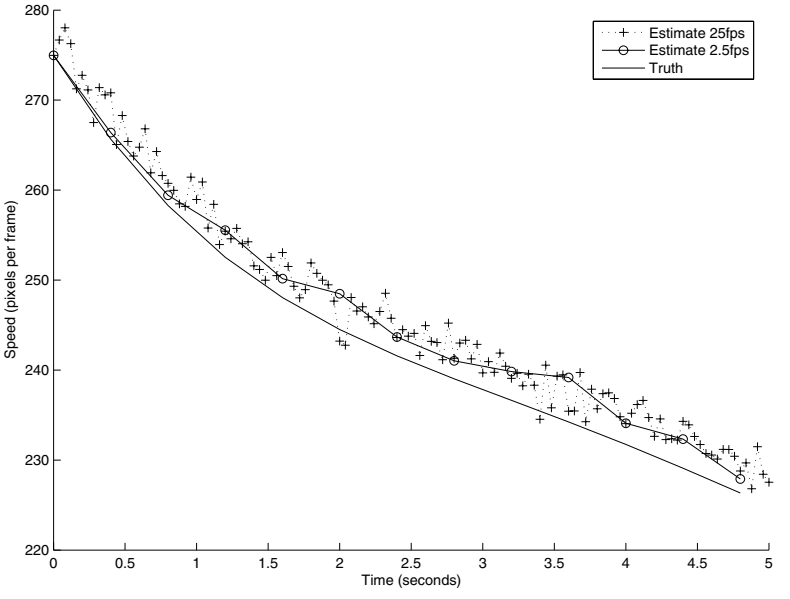

Fig. 1. Speed estimates for 5 seconds of video

thus "true". This reflects the fact that we present here a method of estimating the velocity magnitude, not a method of weak calibration. Figure 1 shows the velocity magnitude estimated using all of the available 25 frames per second, and the magnitude estimated using only 2.5 frames per second. The result based on the lower frame rate is intended to show the degradation in performance in the case where a full weak calibration is not available for each frame of the video. The velocity magnitudes are represented in units of pixels per frame of video (at 25 frames per second).

The graph shows that, despite small variations in the estimated magnitude, on average the method performs very well. The maximum speed error in the 25 frames per second estimate is 0.93%. The decreased frame rate of the 2.5 frames per second test causes a decrease in accuracy as would be expected. The maximum error in the 2.5 frames per second estimate is 1.5%.

5 Conclusion

A novel method was presented for incremental recovery of the magnitude of a camera's translational velocity from time-varying optical flow, relative to an arbitrarily fixed starting value. Results of preliminary synthetic testing confirmed the validity of the approach. Future work will explore how the method may be improved via techniques such as Kalman-like smoothing, use of multiple optical-flow trajectories, and reduction of error accumulation effects.

Acknowledgement

This research was partially supported by the Australian Research Council.

References

1. Brooks, M.J., Chojnacki, W., Baumela, L.: Determining the egomotion of an uncalibrated camera from instantaneous optical flow. Journal of the Optical Society of America A **14** (1997) 2670–2677
2. Faugeras, O.D.: Three-Dimensional Computer Vision: A Geometric Viewpoint. MIT Press, Cambridge, Mass. (1993)
3. Horn, B.K.P.: Robot Vision. MIT Press, Cambridge, Mass. (1986)
4. Faugeras, O.D., Luong, Q.T., Maybank, S.J.: Camera self-calibration: Theory and experiments. In Sandini, G., ed.: Computer Vision—ECCV '92, Second European Conference on Computer Vision. Volume 588 of Lecture Notes in Computer Science., Santa Margherita Ligure, Italy, May 19–22, 1992, Springer-Verlag, Berlin (1992) 321–334
5. Maybank, S.J., Faugeras, O.D.: A theory of self-calibration of a moving camera. International Journal of Computer Vision **8** (1992) 123–151
6. Kanatani, K.: Statistical Optimization for Geometric Computation: Theory and Practice. Elsevier, Amsterdam (1996)
7. Åström, K., Heyden, A.: Multilinear forms in the infinitesimal-time case. In: Proceedings, CVPR '96, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, June 18–20, 1996, IEEE Computer Society Press, Los Alamitos, CA (1996) 833–838
8. Armangué, X., Araújo, H., Salvi, J.: A review on egomotion by means of differential epipolar geometry applied to the movement of a mobile robot. Pattern Recognition **36** (2003) 2927–2944