TR-189, ICS-FORTH, March 1997.

Independent 3D Motion Detection Based on Depth Elimination in Normal Flow Fields

Antonis A. Argyros†, ‡

†CVRL/ICS/FORTH
POBox 1385 Heraklion, Crete, GR-711-10, Greece
E-mail: {argyros,orphanou}@ics.forth.gr

Stelios C. Orphanoudakis †

‡CVAP/NADA/KTH
Stockholm, S-100 44, Sweden
E-mail: argyros@nada.kth.se

Abstract

This paper considers a specific problem of visual perception of motion, namely the problem of visual detection of independent 3D motion. Most of the existing techniques for solving this problem rely on restrictive assumptions about the environment, the observer's motion, or both. Moreover, they are based on the computation of optical flow, which amounts to solving the ill-posed correspondence problem. In this work, independent motion detection is formulated as robust parameter estimation applied to the visual input acquired by a binocular, rigidly moving observer. Depth and motion measurements are combined in a linear model. The parameters of this model are related to the parameters of self-motion (egomotion) and the parameters of the stereoscopic configuration of the observer. The robust estimation of this model leads to a segmentation of the scene based on 3D motion. The method avoids the correspondence problem by employing only normal flow fields. Experimental results demonstrate the effectiveness of this method in detecting independent motion in scenes with large depth variations, without any constraints imposed on observer motion.

1. Introduction

The visual perception of motion has been the subject of many research efforts due to its fundamental importance for many visually assisted tasks. Independent 3D motion detection (IMD) is an important motion perception capability of a seeing system. In a world where changes of state are often more important than the states themselves, the perception of independent motion provides a rich input to attention, informing a seeing system about dynamic changes in the environment.

In the case of a static observer, the problem of independent motion detection can be treated as a problem of *change detection* [8]. The situation is much more complicated when the observer moves relative to the environment. In this case, even the static parts of the scene appear to be moving in a way that depends on the motion of the observer and on the structure of the viewed scene. The case of a moving observer, is also of great interest because biological and some man-made visual systems are usually in constant motion.

In the case of a moving observer, IMD has been often approached as a problem of segmenting the 2D motion that is computed from a temporal sequence of images. Wang and Adelson [16] estimate affine models for optical flow in image patches. Patches are then combined in larger motion segments based on a k-means clustering scheme that merges two patches if the distance of their motion parameters is sufficiently small. Nordlund and Uhlin [11] estimate the parameters of an affine model of 2D motion, assuming that the estimation of the model parameters will not be affected considerably by the presence of small independently moving objects. IMD is then achieved by determining the points where the residual between the measured and the predicted flow is large. The basic problem of the methods that employ 2D models is that they assume scenes where depth variations are small compared to the distance from the observer. However, in real scenes depth variations may be large and, therefore, the discontinuities that are detected by the 2D methods are not only due to motion, but also due the structure of the scene.

Solutions to the problem of IMD have also been provided using 3D models. Employing 3D models makes the problem more difficult because extra variables are introduced regarding the depths of scene points. This in turn requires certain assumptions to be made made in order to provide additional constraints for the problem. Most of the methods depend on the accurate computation of a dense optical flow field or on the computation of a sparse map of feature correspondences. Wang and Duncan [17] present an iterative method for recovering the 3D motion and structure of in-

dependently moving objects from a sparse set of velocities obtained from a pair of calibrated, parallel cameras. Other assumptions that are commonly made by existing methods are related to the motion of the observer, to the structure of the scene in view, or both. Sharma and Aloimonos [13] and Clarke and Zisserman [4] have considered the IMD problem for an observer pursuing restricted translational motion. Adiv [1] performs segmentation by assuming planar surfaces undergoing rigid motion, thus introducing an environmental assumption. Thompson and Pong [14] derive various principles for detecting independent motion when certain aspects of the egomotion or of the scene structure are known. However, the practical exploitation of the underlying principles is limited because of the assumptions they are based on and other open implementation issues. Argyros et al [2] present a method that uses stereoscopic information to segment an image into depth layers, in an effort to decompose the 3D problem into a set of 2D ones. The method provides reliable results at each depth layer, but there are certain limitations regarding the integration of results from the various depth layers. In Argyros et al [3], qualitative functions of depth estimated from stereo and motion are extracted in image patches. Comparison of these functions leads to conclusions regarding the number of 3D motions in a patch. The method is reliable and computationally efficient, but the resulting map of independently moving objects is coarse.

In order to overcome the limitations of existing methods, this paper proposes a new method for IMD. The method relies on the computation of normal flow, the component of motion in the direction of the image gradient, which is less informative compared to optical flow but can be more accurately computed from a temporal sequence of images. Based on the choice of normal flow to represent visual motion, the method exploits stereoscopic information in order to eliminate the depth variable from 3D motion equations. However, knowledge on the parameters of the stereo configuration (i.e. extrinsic calibration) is not required. The method assumes an observer that moves rigidly with unrestricted translational and rotational egomotion. Independent motion can be rigid or non-rigid.

The rest of this paper is organized as follows. Section 2 presents the input used by the proposed method and issues related to robust regression, which constitutes a basic building block of the proposed method. Section 3 presents the method itself. Section 4 presents experimental results from applying the method to real-world image sequences. Finally, section 5 concludes the paper with an overview of its main contributions.

2. Preliminaries

Before proceeding with the description of the proposed method, issues related to motion representation are discussed. In addition, a brief discussion on robust regression methods is provided, since they constitute a building block of the proposed IMD method.

2.1. Visual motion representation

Consider a coordinate system $OXYZ$ at the optical center (nodal point) of a pinhole camera, such that the axis OZ coincides with the optical axis. Suppose that the camera is moving rigidly with respect to its 3D static environment with translational motion (U, V, W) and rotational motion (α, β, γ), as shown in Fig. 1. Under perspective projection,

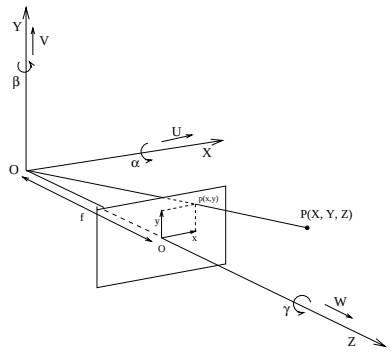


Figure 1. The camera coordinate system.

the equations relating the 2D velocity (u, v) of an image point $p(x, y)$ to the 3D velocity of the projected 3D point $P(X, Y, Z)$ are [9]:

$$\begin{aligned} u &= \frac{(-Uf + xW)}{Z} + \alpha\frac{xy}{f} - \beta\left(\frac{x^2}{f} + f\right) + \gamma y \\ v &= \frac{(-Vf + yW)}{Z} + \alpha\left(\frac{y^2}{f} + f\right) - \beta\frac{xy}{f} - \gamma x \end{aligned} \quad (1)$$

Equations (1) describe a 2D motion vector field, which relates the 3D motion of points to their 2D projected motion on the image plane. The motion field is a purely geometrical concept and it is not necessarily identical to the optical flow field [6], which describes the motion of brightness patterns observed because of the relative motion between the imaging system and the viewed scene. Even in the cases that these two fields are identical, the computation of the optical flow field requires special conditions (such as smoothness) to be satisfied for a unique solution to exist. This is because the computation of optical flow requires the recovery of two unknowns (u, v) at a certain point, while, at each point, only one constraint can be derived without any smoothness assumptions. This constraint is the well known *optical flow constraint equation*, originally developed by Horn and Schunk [7]:

$$I_x u + I_y v + I_t = 0 \quad (2)$$

In eq. (2) I_x, I_y and I_t are the two spatial and the temporal derivatives of the image intensity function. This equation gives only one local constraint on the flow values. In order to get a second constraint, the methods that aim at recovering optical flow typically assume a smooth flow field. However, this assumption does not always hold because of depth discontinuities, independent 3D motion etc.

For the above reason, the proposed IMD method does not rely on the computation of optical flow, but rather on the normal flow field, the projection of the optical flow field in the direction of image gradients. The normal flow field is not necessarily identical to the *normal motion field* (the projection of the motion field along the image gradient), in the same way that the optical flow is not necessarily identical to the motion field [15]. It has been shown, however, that normal flows are reliable in points where the image gradient has a large magnitude. Normal flow vectors at such points can be used as a robust input to 3D motion perception algorithms.

2.2. Robust regression

The aim of robust regression methods [12] is to estimate the parameters of a linear model based on data sets containing outliers, i.e. observations that deviate considerably from the model describing the rest of the observations. The main characteristic of robust estimators is their high breakdown point, which may be defined as the smallest amount of outlier contamination that may force the value of the estimate outside an arbitrary range.

A variety of robust estimators have been used in computer vision. The RANSCAC method [5] is probably the most popular one, but its reported breakdown point is small compared to other robust estimators. Meer et al [10] provide an interesting review of the use of robust regression methods in computer vision.

The LMedS method, proposed by Rousseeuw [12], is a robust estimator with a breakdown point of 50%. Qualitatively, LMedS tries to find a set of model parameters such that the model best fits the *majority* of the observations. Once LMedS has been applied to a set of observations, a standard deviation estimate can be derived, which enables the identification of model outliers. The high breakdown point of LMedS makes it suitable for the purposes of this work.

3. Proposed method

Consider a stereoscopic observer that is moving with unrestricted motion in 3D space. Due to this motion, a reliable normal flow vector can be computed at each point where the image intensity gradient is large. Let (n_x, n_y) be the unit vector in the gradient direction. The magnitude u^M of the normal flow vector is given by:

$$u^M = un_x + vn_y \tag{3}$$

which, by substitution from eq. (1), yields:

$$\begin{aligned} u^M &= -n_x f\frac{U}{Z} - n_y f\frac{V}{Z} + \left(xn_x + yn_y\right)\frac{W}{Z} \\ &+ \left\{\frac{xy}{f}n_x + \left(\frac{y^2}{f} + f\right)n_y\right\}\alpha \\ &- \left\{\left(\frac{x^2}{f} + f\right)n_x + \frac{xy}{f}n_y\right\}\beta + (yn_x - xn_y)\gamma \end{aligned} \tag{4}$$

Equation (4) highlights some of the difficulties of the IMD problem when employing normal flow. Each image point (in fact, each point at which the intensity gradient has a significant magnitude and, therefore, a reliable normal flow vector can be computed) provides one constraint on the 3D motion parameters. For each 3D motion k present in the scene (either egomotion or independent motion), one set of unknown motion parameters (U_k, V_k, W_k), $(\alpha_k, \beta_k, \gamma_k)$ is introduced. Furthermore, if no assumption is made regarding the depth Z, each point introduces one independent depth variable. Thus, n computed normal flow vectors and m 3D motions result in n available constraints with $n + 6m$ unknowns. Evidently, the problem cannot be solved without any additional information on depth.

Consider now the geometry of a typical stereo configuration of a fixating pair of cameras. A pair of images captured with such a configuration contains information relevant to depth, that manifests itself in the form of *disparities* defined by the displacements of points between images. Since the stereoscopic pair of images can be acquired simultaneously, there is no dynamic change in the world that can be recorded by them. It can easily be observed that a stereo image pair is identical to the sequence that would result from a hypothetical (ego)motion that brings one camera to the position of the other[1]. This observation enables the analysis of a stereo pair based on motion analysis techniques. The hypothetical motion that transforms the position of one camera to the other is simpler than the one described by the general motion model of eq. (1). Fig. 2 shows the motion that maps the position of the left camera that of the right camera. Evidently, there is no rotation around the X and Z axes, and no translation along the Y axis. Thus, the translational and rotational component of the imaginary motion can be written as $(U_s, 0, W_s)$ and $(0, \beta_s, 0)$, respectively. Furthermore, in most practical situations, the translation W_s along the Z axis is negligible compared to the rest of the terms. W_s is usually two orders of magnitude smaller than U_s. In fact, for

[1] Regardless of how a pair of images is captured, i.e. by a binocular system configuration or by a moving camera, these images can be considered as views of the same scene from different viewpoints.

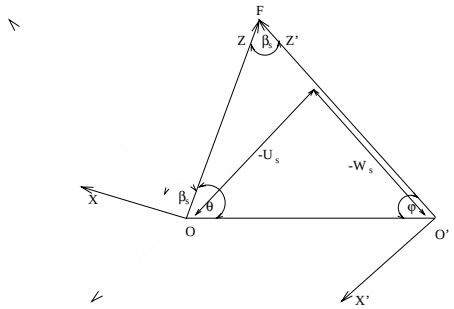


Figure 2. The parameters of the motion that transforms the position of the left camera to the position of the right camera (top view of the stereo configuration).

special stereo configurations (e.g. a right angled one) it can be shown that W_s is exactly equal to zero. Consequently, at each image point, a normal flow value u^S due to stereo may be computed as:

$$u^S = -n_x f \frac{U_s}{Z} - \left\{ \left(\frac{x^2}{f} + f \right) n_x + \frac{xy}{f} n_y \right\} \beta_s \quad (5)$$

In practical situations, the computation of normal flow from a stereoscopic pair of images needs further consideration. The computation of normal flow is based on the optical flow constraint equation, which does not hold if the two images differ too much. Moreover, normal flow is computed from discrete images through spatial and temporal differentiation with small masks. Issues related to the computation of normal flow due to stereo are considered in Argyros et al [2].

By solving eq. (5) for Z, we obtain:

$$Z = \frac{-n_x f U_s}{u^S - \left[\left(\frac{x^2}{f} + f \right) n_x + \frac{xy}{f} n_y \right] \beta_s} \quad (6)$$

The computation of normal flow involves the computation of the partial image derivatives I_x and I_y, which define the normalized vector (n_x, n_y) in the gradient direction. If, for the computation of both stereo and motion normal flow fields, these derivatives are computed in the same reference frame, then n_x and n_y are the same for both eqs. (4) and (6). Therefore, the substitution of eq. (6) into eq. (4) results in the following equation:

$$\begin{aligned} u^M &= u^S \frac{U}{U_s} - \left\{ \left(\frac{x^2}{f} + f \right) n_x + \frac{xy}{f} n_y \right\} \left(\frac{U\beta_s}{U_s} - \beta \right) \\ &+ \frac{n_y}{n_x} u^S \frac{V}{U_s} - \frac{n_y}{n_x} \left\{ \left(\frac{x^2}{f} + f \right) n_x + \frac{xy}{f} n_y \right\} \frac{V\beta_s}{U_s} \\ &- \frac{(xn_x + yn_y) u^S}{n_x f} \frac{W}{U_s} \\ &+ \frac{(xn_x + yn_y)\{(\frac{x^2}{f} + f) n_x + \frac{xy}{f} n_y\}}{n_x f} \frac{W\beta_s}{U_s} \\ &+ \left\{ \frac{xy}{f} n_x + \left(\frac{y^2}{f} + f \right) n_y \right\} \alpha + (yn_x - xn_y)\gamma \end{aligned} \quad (7)$$

Equation (7) is linear in the variables $\phi_1 = \frac{U}{U_s}$, $\phi_2 = \frac{U\beta_s}{U_s} - \beta$, $\phi_3 = \frac{V}{U_s}$, $\phi_4 = \frac{V\beta_s}{U_s}$, $\phi_5 = \frac{W}{U_s}$, $\phi_6 = \frac{W\beta_s}{U_s}$, $\phi_7 = \alpha$, and $\phi_8 = \gamma$. These variables are expressions involving the 3D motion parameters and the stereo configuration parameters. LMedS estimation can be applied to a set of observations of the model of eq. (7) as a means to estimate the parameters ϕ_i, $1 \leq i \leq 8$. LMedS will provide estimates $\hat{\phi}_i$ of the parameters ϕ_i and a segmentation of the image points into model inliers and model outliers. Model inliers, which are compatible with the estimated parameters $\hat{\phi}_i$, correspond to image points that move with a dominant set of 3D motion parameters. A point may belong to the set of outliers if at least one of the following holds:

1. The quantities u^S and/or u^M for this point have been computed erroneously.
2. The 3D motion parameters for this point are different compared to the 3D motion parameters describing the majority of points.

The points of the first class will, in principle, be few and sparsely distributed over the image plane. This is because only reliable normal flow values are considered. The second class of points is essentially the class of points that are not compatible with the dominant 3D motion parameters. Thus, in the case of two rigid motions in a scene, the inlier/outlier characterization of points achieved by LMedS is equivalent to a dominant/secondary 3D motion segmentation of the scene. In the case that more than two rigid motions are present in a scene, the correctness of 3D motion segmentation depends on the spatial extent of the 3D motions. If there is one dominant 3D motion (in the sense that at least 50% of the total number of points move with this motion), LMedS will be able to handle the situation successfully. This is because of the high breakdown point of LMedS, which tolerates an outlier percentage of up to 50% of the total number of points. The inliers will correspond to the dominant motion (egomotion) and the set of outliers will contain all secondary (independent) motions. A recursive application of LMedS to the set of outliers may further discriminate the rest of the motions. The recursive application of LMedS should be terminated when the remaining points become fewer than a certain threshold. There are two reasons for this. First, if the number of points becomes too small, then the number of constraints provided by eq. (7) becomes small and the discrimination between inliers and outliers is subject to errors. Second, at each recursive application of LMedS, the set of outliers does not contain only

points that correspond to a motion different than the dominant one, but also points where normal flows have not been computed accurately.

3.1. Postprocessing

According to the proposed method for independent motion detection, points are characterized as being independently moving or not based on their conformance to a general rigid 3D model of egomotion. The characterization is made at the point level, without requiring any environmental assumptions, such as smoothness, to hold in the neighborhood of each point. In order to further exploit information regarding independent motion, it is often considered preferable to refer to connected, independently moving areas rather than to isolated points. There are three main reasons why the points of a motion segment do not form connected regions. First, the normal flow field is usually a sparse field, because normal flow values are considered unreliable in certain cases (e.g. in points with a small gradient value). Second, there is always the possibility of errors in measurements of normal flow and, therefore, some points may become model inliers (or outliers) because of these errors and not due to their 3D motion parameters. Finally, normal flow is a projection of the optical flow onto a certain direction. Infinitely many other optical flow vectors have the same projection on this direction. Consequently, a normal flow vector may be compatible with the parameters of two different 3D motions, and therefore a number of point misclassifications may arise.

We overcome the problem of disconnected motion segments by exploiting the fact that, in the above cases, misclassified points are sparsely distributed over the image plane. A simple majority voting scheme is used. At a first step, the number of inliers and outliers is computed in the neighborhood of each image point. The label of this point becomes the label of the majority in its neighborhood. This allows isolated points to be removed. In the resulting map, the label of the outliers is replicated in a small neighborhood in order to group points of the same category into connected regions.

3.2. Egomotion estimation

Besides the inlier/outlier characterization, LMedS provides estimations $\hat{\phi}_i$ for the parameters ϕ_i of the linear model of eq. (7). Each of the model parameters ϕ_i corresponds to expressions involving the 3D motion parameters (U, V, W) and (α, β, γ) of the observer and the stereo configuration parameters (U_s, β_s). Thus, the observer is able to relate his own motion parameters to the parameters of his stereo configuration, i.e. to parameters of his own body[2]. Moreover, the estimated parameters $\hat{\phi}_i$ can also be used to provide quantitative knowledge regarding the 3D motion parameters of the observer. More specifically, the following relations hold:

$$x_0 = \frac{\phi_1 f}{\phi_5},\ y_0 = \frac{\phi_3 f}{\phi_5}$$

$$\alpha = \phi_7,\ \beta = \phi_2 - \phi_1 \frac{\phi_4}{\phi_5},\ \gamma = \phi_8$$

where, x_0 and y_0 are the coordinates of the FOE (i.e. the point where the direction of translation intersects the infinitely large image plane). Similarly, an estimation of the vergence angle of the stereo configuration is possible:

$$\beta_s = \frac{\phi_4}{\phi_3}$$

4. Experimental results

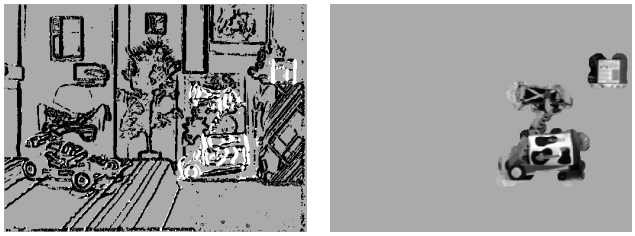
The experimental evaluation of the proposed method has been based on real-world image sequences that have been acquired by TALOS, the mobile robotic platform of the Computer Vision and Robotics Laboratory (CVRL) of FORTH. Several experiments have been conducted to test the proposed method. It should be stressed that during the course of all the experiments the exact values for the intrinsic camera parameters and the stereo configuration parameters were unknown.

A sample result refers to the "cart" image sequence. One frame of the "cart" sequence (right image of the stereo pair at time t) is shown in Fig. 3. In this sequence, a binocular

Figure 3. One frame of the "cart" sequence.

observer with parallel cameras performs a translational motion with U and W components as well as with a rotational β component. The horizontal translation is the motion that dominates. The scene contains a distant background and a foreground close to the observer. The background contains two independently moving objects: A cart that translates in the opposite direction of the observer (middle of the scene) and a small box (to the right of the scene) that translates in the same direction with the observer, but with different velocity. The foreground of the scene contains a table on which there is a toy car. Both objects are stationary. Figure 4 illustrates the results of 3D motion segmentation of the "cart" sequence

[2]For example, $\hat{\phi}_1$ relates the horizontal component of the instantaneous translational 3D motion to the baseline length.



(a) (b)

Figure 4. 3D motion segmentation for the "cart" sequence (a) before and, (b) after post-processing.

by using the proposed method. Figure 4(a) shows the intermediate segmentation results (after LMedS estimation). Black color corresponds to egomotion and white color corresponds to independent motion. Gray color corresponds to points where no decision can be made, due to low values of image gradients and, therefore, lack of normal flow vectors. It can be observed that the largest concentration of white (i.e. independently moving) points is on the regions of the independently moving objects. The points that are not identified as independently moving, although they belong to an independent motion, are mainly those belonging to horizontal edges. This is because the model of eq. (7) does not hold for $n_x = 0$, which is the case of vertical gradients or, equivalently, horizontal edges ($(n_x, n_y) = (0, 1)$). Figure 4(b) presents the results of Fig. 4(a) after postprocessing, which eliminates isolated outliers (inliers) in large populations of inliers (outliers) and, in the resulting map, dilates the label of remaining outliers in a small neighborhood. In Fig. 4(b), areas that are detected as independently moving appear with the intensities that they have in the original image, while all other areas are masked out. It can be seen that after this type of postprocessing the bodies of the cart and the box have been successfully identified as independently moving.

5. Conclusions

Artificial seeing systems should be able to operate in environments that contain both stationary and moving objects. The perception of independent 3D motion is crucial because it provides useful information on where attention should be focused and, probably, maintained. In this paper, IMD was based on motion and structure information that an observer acquires while moving in 3D space. The proposed method employs 3D motion models and is able to perform satisfactorily even in scenes with considerable depth variations. The method relies on normal flow fields, thus avoiding the ill-posed correspondence problem. Unrestricted rigid egomotion was assumed for the observer. Ongoing research aims at exploiting the proposed method in the general context of a robot navigating in 3D space, where the cooperation among various visually-guided behaviors and issues such as real-time performance are of central importance.

References

[1] G. Adiv. Determining Three Dimensional Motion and Structure from Optical Flow Generated by Several Moving Objects. *IEEE Trans. on PAMI*, 7(4):384–401, July 1985.

[2] A. Argyros, M. Lourakis, P. Trahanias, and S. Orphanoudakis. Independent 3D Motion Detection Through Robust Regression in Depth Layers. In *Proc. BMVC '96, Edinburgh, UK*, Sep. 9-12 1996.

[3] A. Argyros, M. Lourakis, P. Trahanias, and S. Orphanoudakis. Qualitative Detection of 3D Motion Discontinuities. In *Proc. IROS '96, Tokyo, Japan*, Nov. 4-8 1996.

[4] J. C. Clarke and A. Zisserman. Detection and Tracking of Independent Motion. *Image and Vision Computing*, 14:565–572, 1996.

[5] M. Fischler and R. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *CACM*, 24:381–395, 1981.

[6] B. Horn. *Robot Vision*. MIT Press, Cambridge, MA, 1986.

[7] B. Horn and B. Schunck. Determining Optical Flow. *Artificial Intelligence*, 17:185–203, 1981.

[8] Y. Hsu, H. Nagel, and G. Rekkers. New Likelihood Test Methods for Change Detection in Image Sequences. *CVGIP*, 26:73–106, 1984.

[9] H. Longuet-Higgins and K. Prazdny. The Interpretation of a Moving Retinal Image. In *Proc. Royal Society*, pages 385–397. London B, 1980.

[10] P. Meer, A. Mintz, and A. Rosenfeld. Robust Regression Methods for Computer Vision: A Review. *IJCV*, 6(1):59–70, 1991.

[11] P. Nordlund and T. Uhlin. Closing the Loop: Detection and Pursuit of a Moving Object by a Moving Observer. *Image and Vision Computing*, 14:267–275, 1996.

[12] P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. John Wiley and Sons Inc., New York, 1987.

[13] R. Sharma and Y. Aloimonos. Early Detection of Independent Motion from Active Control of Normal Image Flow Patterns. *IEEE Transactions on SMC*, SMC-26(1):42–53, February 1996.

[14] W. Thompson and T. Pong. Detecting Moving Objects. *IJCV*, 4:39–57, 1990.

[15] A. Verri and T. Poggio. Motion Field and Optical Flow: Qualitative Properties. *IEEE Trans. on PAMI*, 11(5):490–498, May 1989.

[16] J. Wang and E. Adelson. Representing Moving Images with Layers. *IEEE Trans. on Image Processing*, 3(5):625–638, Sep. 1994.

[17] W. Wang and J. H. Duncan. Recovering the Three-Dimensional Motion and Structure of Multiple Moving Objects from Binocular Image Flows. *Computer Vision and Image Understanding*, 63(3):430–440, May 1996.