Independent 3D Motion Detection Through Robust Regression in Depth Layers

Antonis A. Argyros, Manolis I.A. Lourakis,
Panos E. Trahanias and Stelios C. Orphanoudakis
Institute of Computer Science, FORTH
PO Box 1385, Heraklion, Crete 711-10, Greece
and
Computer Science Department, University of Crete
PO Box 1470, Heraklion, Crete 714-09, Greece
{argyros, lourakis, trahania, orphanou}@ics.forth.gr

Abstract

This paper presents a novel method for the detection of objects that move independently of the observer in a 3D dynamic environment. Independent 3D motion detection is formulated as a problem of robust regression applied to visual input acquired by a binocular, rigidly moving observer. The qualitative analysis of images acquired by a parallel stereo configuration yields a segmentation of a scene into depth layers. A depth layer consists of points of the 3D space for which depth variations are small compared to the distance from the observer. Robust regression is applied to each depth layer in order to segment the latter into coherently moving regions. Finally, a combination stage is applied across all layers in order to come up with an integrated view of independent motion in the whole 3D scene. In contrast to other existing approaches for independent motion detection which are based on the ill-posed problem of optical flow computation, the proposed method relies on normal flow fields for both stereo and motion processing. Experimental results show the effectiveness and robustness of the proposed scheme, which is capable of discriminating independent 3D motion in scenes with large depth variations.

1 Introduction

The visual field of a moving observer who is equipped with visual sensors, appears to be moving in a specific manner, depending on the observer's egomotion and the structure of the viewed scene. The problem of independent 3D motion detection can be defined as the problem of locating objects that move independently of the observer in his field of view. The ability to detect independent 3D motion is very important for an observer interacting with a dynamic environment. It is known

[1] that independent motion detection is one of the basic visual competences of most of the biological organisms possessing the sense of vision.

Most of the research efforts towards independent motion detection depend on the accurate computation of the optical flow field. Jain [2] has considered the problem of independent 3D motion detection by an observer pursuing translational motion. In addition to imposing constraints on egomotion, knowledge of the direction of translation is required. Thompson [3] derives various principles for detecting independent motion when certain aspects of the egomotion or of the scene structure are known. However, the practical exploitation of these principles is made difficult by the limiting assumptions they are based on. Bouthemy [4] views motion segmentation as a problem of statistical regularization using Markov Random Field models. The method proposed by Sharma [5] uses the spatiotemporal derivatives of the image intensity function (the so called *normal flow field*), rather than optical flow. However, as in the case of [2], known translational egomotion is hypothesized. Nelson [6] presents two methods for independent motion detection which are also based on the normal flow field. The first of these methods requires *a priori* knowledge of egomotion parameters and assumes upper bounds on the depth of the scene. The second method detects abrupt changes of independent motion rather than independent motion itself.

The method proposed in this paper makes use of the *Least Median of Squares (LMedS)* estimation technique [7]. Initially, images that are acquired by a binocular observer are processed in order to separate the image points into depth layers; each depth layer corresponds to a set of points whose difference in depth is small with respect to their distance from the observer. At a second stage, LMedS is applied to the motion data acquired by the observer at successive time instances. The application of LMedS results in the segmentation of each depth layer into motion inliers and motion outliers. Motion inliers correspond to points moving with a dominant set of 3D motion parameters. Motion outliers correspond to points that do not conform with the dominant motion parameters. Finally, a combination stage is responsible for integrating the information collected through the various layers, yielding the 3D motion segmentation of the scene. Robust regression has also been employed in the past in the problem of motion segmentation [8]. However, since no information on scene structure is used, the method presented in [8] is only applicable in cases of scenes forming a *frontoparallel plane*.

In contrast to other approaches for motion segmentation that use optical flow [2, 3], the proposed method is based on normal flow. The ill-posed correspondence problem is not only avoided for the case of motion, but also for the case of stereo which is treated as the hypothetical motion that would map the position of the left camera to the position of the right camera. Again, normal flow is computed between the two frames of the stereo pair.

The rest of the paper is organized as follows. Section 2 describes the input used by the independent motion detection method. It also gives a brief introduction to robust regression and the LMedS estimation technique, which constitutes a basic building block of the approach. The independent motion detection method is fully described in section 3. In section 4, experimental results are presented and discussed. Finally, section 5 summarizes the contributions of this work.

2 Preliminaries

Issues related to motion representation are briefly reviewed here. The rationale behind the choice to employ the normal flow field in all computations is given. Additionally, a discussion on robust regression and, more specifically, on LMedS is provided, since the later comprises a basic building block of the described scheme.

2.1 Visual motion representation

Consider a 3D coordinate system positioned to the optical center (nodal point) of a camera. Suppose that the camera moves rigidly in its 3D static environment with translational motion $\vec{t} = (U, V, W)$ and rotational motion $\vec{\omega} = (\alpha, \beta, \gamma)$. Under perspective projection, 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]:

$$u = \frac{(-Uf + xW)}{Z} + \alpha\frac{xy}{f} - \beta\left(\frac{x^2}{f} + f\right) + \gamma y \quad (1a)$$

$$v = \frac{(-Vf + yW)}{Z} + \alpha\left(\frac{y^2}{f} + f\right) - \beta\frac{xy}{f} - \gamma x \quad (1b)$$

where f represents the focal length of the imaging system.

Equations (1) describe the 2D *motion field*, which is the projection of the 3D motion of a point on the image plane. The motion field is a purely geometrical concept and is not necessarily identical to the *optical flow* field [10], which describes the apparent motion of brightness patterns resulting from the relative motion between an imaging system and its environment. Verri and Poggio [11] have shown that the motion and optical flow fields are identical in specific cases only. Even in the cases that these two fields are identical, the problem of optical flow estimation is ill-posed [12]. The problem of optical flow computation is often approached using regularization methods, which impose constraints on the solution. Such constraints are related to certain assumptions about the structure of the viewed scene. In practice - especially in the case of independent motion where motion discontinuities exist by definition - these assumptions are quite often violated, resulting in errors in optical flow estimation.

For the above reasons, the proposed method does not rely on optical flow, rather on normal flow, i.e. the projection of optical flow on the direction of the intensity gradient. In order to compute the normal flow field, a sequence of images is modeled as a continuous irradiance function $I(x, y, t)$ of two spatial (x, y) and one temporal (t) variables. Assuming that irradiance is conserved between two consecutive frames, the well known *optical flow constraint equation*, originally developed by Horn and Schunk [10], can be derived:

$$(I_x, I_y) \cdot (u, v) = -I_t \quad (2)$$

where, I_x, I_y and I_t are the spatial and temporal partial derivatives of the image intensity function, respectively, and "$\cdot$" denotes dot product. Equation (2), facilitates the computation of the normal flow field. The latter is not necessarily identical to the normal motion field (the projection of the motion field along the gradient direction), in the same way that the optical flow is not necessarily identical to the motion field. However, normal flow is a good approximation to the normal motion field at points where the image gradient magnitude is large [11]. Normal flow vectors at such points can be used as a robust input to 3D motion analysis.

2.1.1 Normal flow field due to motion

Let (n_x, n_y) be the unit vector in the gradient direction. The magnitude u_{nm} of the normal flow vector is given as $u_{nm} = un_x + vn_y$ which, from eq. (1), yields:

$$\begin{aligned} u_{nm} &= (-n_x f)\frac{U}{Z} + (-n_y f)\frac{V}{Z} + (xn_x + yn_y)\frac{W}{Z} \\ &+ \left\{\frac{xy}{f}n_x + \left(\frac{y^2}{f} + f\right)n_y\right\}\alpha + \left\{-\left(\frac{x^2}{f} + f\right)n_x - \frac{xy}{f}n_y\right\}\beta + (yn_x - xn_y)\gamma \end{aligned} \tag{3}$$

Equation (3) highlights some of the difficulties of the problem of independent motion detection. Each image point (in fact, each point for which a reliable normal flow vector can be computed) provides one constraint on the 3D motion parameters. In the case that only the observer is moving, the above equation holds with the same set of 3D egomotion parameters (U_E, V_E, W_E), $(\alpha_E, \beta_E, \gamma_E)$ at all points. In the case, however, of independent motion, there is at least one more set of motion parameters (U_I, V_I, W_I), $(\alpha_I, \beta_I, \gamma_I)$ which is valid for some of the image points. Furthermore, if no assumptions regarding the depth Z are made, each point introduces an extra independent depth variable. Evidently, the problem cannot be solved if no additional information regarding depth is available.

2.1.2 Normal flow field due to stereo

Consider a stereo configuration, where the optical axes of the two cameras are parallel. A pair of images captured with such a configuration encapsulates information relevant to depth, that manifests itself in the form of *disparities* defined by the displacements of points between images. Since these images are 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 the one camera to the position of the other. This remark enables the analysis of a stereo pair employing motion analysis techniques. Specifically, a translational motion U_s, directly related to the length of the baseline of the stereo configuration, suffices to describe the hypothetical motion. According to eq. (3), a normal flow value u_{ns} due to stereo can be computed at each point, which is equal to

$$u_{ns} = (-n_x f)\frac{U_s}{Z} \tag{4}$$

In practical situations, the computation of normal flow from a pair of stereo images needs further consideration. The computation of normal flow is based on the optical flow constraint equation, which assumes that the motion between consecutive images is in the order of a few pixels. From eq. (4) it can be seen that the magnitude of the stereo normal flow at a point depends on the stereo baseline length (directly related to U_s) and the scene structure (Z), but not on the coordinates of the point on the image plane. If a lower bound for the scene structure and an upper bound for the baseline length can be established, it is ensured that the stereo-equivalent motion (and, therefore, stereo normal flow) can be bounded[1]. Additionally, the magnitude of the optical flow (and consequently of the normal flow) is also a function of the spatial image resolution. Therefore, a proper selection of image resolution can be made, so that the magnitude of normal flow vectors is within valid limits, at the cost of computing coarser depth information.

2.2 Robust regression

Regression analysis (fitting a model to noisy data) is a very important statistical tool. In the general case of a linear model [7], the problem is to estimate the model parameters based on observations of the model that may be contaminated with noise. Traditionally, model parameters are estimated by the popular least squares (LS) method. However, the LS estimator becomes highly unreliable in the presence of outliers, that is observations that deviate considerably from the model describing the rest of the observations. Robust regression methods [7] have been proposed in order to cope with such cases. 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.

The LMedS method, proposed by Rousseeuw [7], comprises one such robust estimation method. 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 may be derived. Based on this estimate, the observations are classified into *model inliers* and *model outliers*. LMedS has a very high breakdown point of 50%, which makes it suitable for the purposes of this work.

3 Independent motion detection

Consider eq. (3) for all normal flows that have been computed from a pair of successive images in time. This relation forms a linear model, in cases where the depth Z and the 3D motion parameters are constant for all points. In terms of LMedS estimation, the outliers of the linear model will be either points for which Z deviates from a dominant depth, or points whose 3D motion is different from the dominant motion, or points where noise was introduced in the computation of normal flow. For the purpose of independent motion detection, we are interested in the second class of points. If we are able to define subsets of observations that correspond to points with (approximately) the same depth and restrict the application of LMedS to each of these subsets, then outliers should be due to independent motion only. The third class of points may easily be discriminated,

[1] n_x does not violate this assumption since it is a normalized value in the range $[0, 1]$.

because it is expected that such points are very few and uniformly distributed over the image plane. It is now possible to delineate the following algorithmic scheme for independent motion detection, for the case of unrestricted 3D egomotion: (a) segment the image points into *depth layers*, (b) for each depth layer, apply LMedS estimation to identify motion outliers, and (c) combine results across all depth layers to get a global 3D motion segmentation. In the following, we provide solutions for the steps of the above, general algorithmic framework.

3.1 Layering of a scene with respect to depth

Let S be the set of all points p_i, $1 \leq i \leq n$, of the image plane for which reliable normal flow values can be computed. Each point $p_i \in S$ corresponds to a point P_i of the 3D world, with a depth Z_i from the observer. Each point p_i may define a *depth layer* L_i, i.e. a subset of S, based on the following relation:

$$L_i = \left\{ p_j : \left| \frac{Z_i - Z_j}{Z_i} \right| < \epsilon \right\} \tag{5}$$

The above relation defines a set of points having depths that differ from Z_i by a small percentage. Each of the layers L_i can be interpreted as a "slice" of the 3D space, that contains 3D points within a range of depths from the observer. The farthest from the observer, the thicker the depth layers become, for the same ϵ. If ϵ is selected to be sufficiently small, then the depth variables within a layer L_i can be considered as constant, equal to some value C_i which depends on layer L_i.

Depth layering can be achieved by appropriate processing of the normal flow values that are computed from a parallel stereo configuration[2]. More specifically, eq. (4) can be written as $g(Z) = \frac{A}{Z}$, where $g(Z) = -\frac{u_{ns}}{n_x}$ is a computable quantity, and $A = U_s f$ is an unknown constant, dependent on the stereo configuration only. Suppose that we want to check if a point p_j belongs to the layer of a point p_i. According to eq. (5), it is required that:

$$\left| \frac{Z_i - Z_j}{Z_i} \right| < \epsilon \Leftrightarrow \left| \frac{\frac{A}{g(Z_i)} - \frac{A}{g(Z_j)}}{\frac{A}{g(Z_i)}} \right| < \epsilon \Leftrightarrow \left| 1 - \frac{g(Z_i)}{g(Z_j)} \right| < \epsilon \tag{6}$$

Since $g(Z_i)$ and $g(Z_j)$ are computable quantities, we can decide whether two points p_i and p_j belong to the same layer or not. Criterion (6) does not depend on the stereo configuration parameter A. Therefore, knowledge of the exact length of the stereo baseline or of the focal length is not required. In practice, depth layering is performed with an iterative scheme. First, a histogram of the function $g(Z)$ is computed. The highest peak of the histogram is determined and the value of the function at this peak becomes the center for the definition of a depth layer. All points which, according to criterion (6), belong to this layer are excluded from subsequent consideration. These steps are repeated until all points of the image are assigned to depth layers. The presented method for depth layering can be characterized as *direct* in the sense that it surpasses the problem of solving for the stereo configuration parameters and tries to extract information about depth based on a specific function of normal flow.

[2] Non parallel (i.e. fixating) stereo configurations can also be used for depth layering, but are not reported here due to space limitations.

3.2 Motion segmentation of a depth layer

Having segmented a scene into depth layers, the goal is now to segment each of these layers based on its 3D motion characteristics. Due to the process of depth layering, it is known that depth differences within a layer are very small compared to the distance from the observer. Robust regression in the form of LMedS can be used in order to estimate the dominant 3D motion parameters in this layer, according to the model of eq. (3). LMedS is actually applied in order to estimate the parameters $(\frac{U}{C_i}, \frac{V}{C_i}, \frac{W}{C_i})$ and (α, β, γ), where C_i is the depth defining the specific depth layer. The application of LMedS will partition the points in a depth layer into model inliers and model outliers. Model inliers correspond to points with a dominant 3D motion. Model outliers correspond to points where either the normal flow values have been corrupted by noise or the underlying 3D motion parameters are not equal to the ones of the dominant motion. Theoretically, up to 50% of outliers can be tolerated. In the case of scenes with at most two rigid motions, motion segmentation can be successfully achieved, since the one or the other motion will dominate and will be estimated by the LMedS regression. In case that there may be more than two rigid motions, the segmentation may be recursively applied to the outliers of the previous robust estimation.

3.3 Integration of results from the various layers

The step of motion segmentation of a layer L_i produces μ motion segments ${M_i}^1$, ${M_i}^2$, $\cdots$, ${M_i}^\mu$, each of which is characterized by a set of parameters $(\frac{U}{C_i}, \frac{V}{C_i}, \frac{W}{C_i})$ and (α, β, γ). In order to come up with a 3D motion segmentation of the whole scene, it should be examined whether two motion segments belonging to different depth layers correspond to the same 3D motion. Unfortunately, the estimated parameters are not pure 3D motion parameters because the translational components of the estimated vectors also include information about depth. Therefore, any direct comparison of the estimated parameter vectors across different depth layers is invalid, unless additional, quantitative information about depth is available. Moreover, the combination of results cannot be achieved on the basis of the inlier or outlier characterization of the scene points, because the dominant motion in one layer may appear as a secondary motion in another layer.

The task of parameter comparison is tackled by reducing the dimensionality of the problem. From each 6-tuple of estimated parameters $(\frac{U}{C_i}, \frac{V}{C_i}, \frac{W}{C_i}, \alpha, \beta, \gamma)$ we derive a 5-tuple $(m_1, m_2, m_3, m_4, m_5) = (\frac{U}{W}, \frac{V}{W}, \alpha, \beta, \gamma)$ by dividing the first two coordinates of the 6-tuple by the third one. This 5-tuple depends only on the 3D motion parameters. Therefore, it forms a basis for deciding whether to merge motion segments residing in different depth layers. The algorithm used for the comparison of the motion segments compares each motion parameter independently. Consider the two motion 5-tuples $({m_1}^a, {m_2}^a, {m_3}^a, {m_4}^a, {m_5}^a)$ and $({m_1}^b, {m_2}^b, {m_3}^b, {m_4}^b, {m_5}^b)$ of motion segments a and b, respectively. These are considered identical iff:

$$\forall i, 1 \leq i \leq 5, \left| \frac{{m_i}^a - {m_i}^b}{max\{{m_i}^a, {m_i}^b\}} \right| < \delta_m \tag{7}$$

where δ_m is a threshold that controls the sensitivity of motion discrimination.

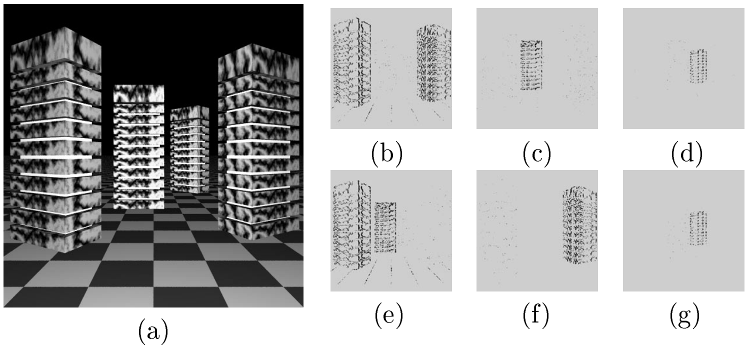


Figure 1: (a) One frame from a synthetic stereoscopic sequence, (b), (c), (d) depth layers, (e), (f), (g) motion segmentation.

The dimensionality reduction employed, affects the discrimination of two motions when both their FOEs[3] and their rotational parameters are identical. However, such cases are not common in practice and, moreover, cannot be tackled without using metric depth information.

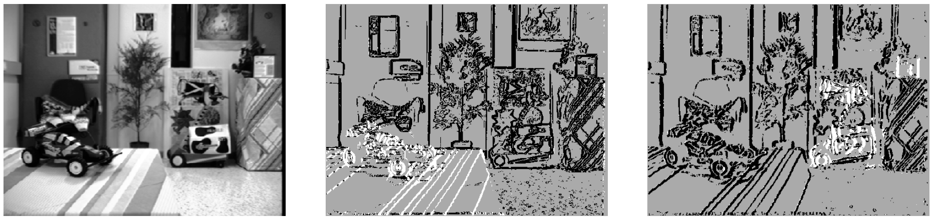
In practical situations, the motion 5-tuples compared are not the estimates provided by LMedS, rather the least squares estimates of the model parameters over the points in the inliers set. This is because in cases where a set of observations has no outliers, least squares estimation gives more accurate estimates of the model parameters. It should be stressed, however, that having already segmented a layer with respect to its motion parameters, all algorithms that can solve the egomotion estimation problem would suffice to accurately estimate the motion parameters of a specific segment and subsequently aid towards 3D motion parameter comparison.

4 Experimental results

The proposed method has been tested with synthetic and real data. The values of the two thresholds δ_m and ϵ (c.f. eqs (5),(7)) were experimentally set.

A first result refers to synthetically generated images. The RAYSHADE [13] ray tracer has been employed to provide a sequence of stereoscopic images. Figure 1(a) shows one frame of the sequence. The composed scene contains 4 artificial "buildings" on a checkered ground. All buildings have the same physical dimensions. The leftmost and rightmost buildings are at the same depth from the observer. The left-middle building is at a larger depth from the observer (compared to the depths of the leftmost and rightmost buildings); the right-middle building is at an even larger depth from the observer. The observer performs a translational motion along the Z axis approaching the scene in view. At the same time, the two buildings in the right half of the image are performing independent motions on their own. The rightmost building performs an independent translational motion along the Y axis and the right-middle building performs a composite translational and rotational motion. Figures 1(b),(c),(d) show the results of depth layering. As

[3] The FOE is the point $(\frac{fU}{W}, \frac{fV}{W})$ on the image plane, which defines the direction of translation.



(a) (b) (c)

Figure 2: (a) A real test image, (b) depth layers, (c) independent motion.

can be verified from this figure, the three different depth layers have been successfully detected and outlined. The points corresponding to each layer have a black color. The first layer corresponds to the two closer objects (leftmost and rightmost buildings), the second to the object in intermediate depth (left-middle building) and the third to the object furthest from the observer (right-middle building). Figures 1(e),(f),(g) show the results of 3D motion segmentation. Two independent motions have been revealed (Figs 1(f),(g)). Egomotion is shown in Fig. 1(e). It can be observed that successful discrimination of all different 3D motions in the scene has been accomplished, although they appear at different depths. Moreover, the 3D motion of the middle-left building has been successfully characterized as being identical to that of the leftmost building.

The method has also been tested using real data. The results obtained in all cases verify the robustness of the method. A sample result refers to the scene of Fig 2(a), which consists of a distant background and a close to the observer foreground. The background consists of a number of static objects as well as an "equipment cart" (right-middle of the scene) and a box (right-top of the scene) that are independently moving between two consecutive image frames. The cart has two tool-racks, each carrying one box. The binocular observer also moves, with unrestricted 3D motion. The image foreground consists of a table on which a toy-car is placed. The depth layering is presented in Fig 2(b). Gray color corresponds to points in the image where normal flow has been rejected as unreliable. For the rest of the points, white color corresponds to the depth layer of the image foreground and black color corresponds to the points of the distant background. The result of 3D independent motion detection is presented in Fig 2(c). In this figure, gray color again corresponds to points where normal flow values have been rejected as unreliable. However, black color now corresponds to the points moving relative to the observer due to his egomotion, while white color corresponds to independently moving points. From Figs 2(b) and 2(c) it can be observed that the method provides correct discrimination of the two depth layers, as well as of the independent motion of the cart (both upper and lower tool-racks) and the box.

5 Summary

In this paper, a method for independent 3D motion detection has been described that combines motion information with stereoscopic information acquired by a parallel stereo configuration. The motivation behind the proposed method is to provide robust 3D motion segmentation by employing the minimum possible assumptions about the external world and the observer. Instead of using optical flow, the normal flow field is used in both stereo and motion domains. Processing of the stereo-pair is limited to the task of scene segmentation into depth layers. Thus, the more general problem of fully recovering scene structure is avoided. LMedS estimation is the basic technique employed. The experimental results obtained, a small sample of which is presented in this paper, demonstrate the robustness and effectiveness of the approach. Therefore, the method may become a powerful tool for an observer navigating in a 3D dynamic environment.

References

[1] A. Horridge. The Evolution of Visual Processing and the Construction of Seeing Systems. In *Proc. Royal Soc., London B 230*, pages 279–292, 1987.

[2] R.C. Jain. Segmentation of Frame Sequences Obtained by a Moving Oberver. *IEEE Transactions on PAMI*, PAMI-7(5):624–629, September 1984.

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

[4] P. Bouthemy and E. Francois. Motion Segmentation and Qualitative Dynamic Scene Analysis from an Image Sequence. *IJCV*, 10(2):157–182, 1993.

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

[6] R.C. Nelson. Qualitative Detection of Motion by a Moving Observer. *IJCV*, 7(1):33–46, 1991.

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

[8] S. Ayer, P. Schroeter, and J. Bigun. Segmentation of Moving Objects by Robust Motion Parameter Estimation over Multiple Frames. In *ECCV*, 1994.

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

[10] B.K.P. Horn. *Robot Vision.* MIT Press, Cambridge, MA, 1986.

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

[12] Y. Aloimonos, I. Weiss, and A. Bandopadhay. Active Vision. *IJCV*, 2:333–356, 1988.

[13] C.E. Kolb. *Rayshade User's Guide and Reference Manual*, 0.4 edition, 1992.