FORTH-ICS / TR-206 August 1997

Independent 3D Motion Detection Using Residual Parallax Normal Flow Fields †

Manolis I. A. Lourakis, Antonis A. Argyros and Stelios C. Orphanoudakis

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 a dense optical flow field, which amounts to solving the ill-posed correspondence problem. In this work, independent motion detection is formulated as a problem of robust parameter estimation applied to the visual input acquired by a rigidly moving observer. The proposed method automatically selects a planar surface in the scene and the residual planar parallax normal flow field with respect to the motion of this surface is computed at two successive time instants. The two resulting normal flow fields are then combined in a linear model. The parameters of this model are related to the parameters of self-motion (egomotion) and their robust estimation leads to a segmentation of the scene based on 3D motion. The method avoids a complete solution to the correspondence problem by selectively matching subsets of image points and by employing normal flow fields. Experimental results demonstrate the effectiveness of the proposed method in detecting independent motion in scenes with large depth variations and unrestricted observer motion.

† A shorter version appears in the Proceedings of the 6th IEEE International Conference on Computer Vision (ICCV'98), Bombay, India, January 4-7 1998.

Independent 3D Motion Detection Using Residual Parallax Normal Flow Fields

Manolis I. A. Lourakis, Antonis A. Argyros and Stelios C. Orphanoudakis

Computer Vision and Robotics Laboratory
Institute of Computer Science (ICS)
Foundation for Research and Technology --- Hellas (FORTH)
Science and Technology Park, Heraklio, Crete
POBox 1385, GR-711-10 Greece
http://www.ics.forth.gr/proj/cvrl
E-mail: lourakis@ics.forth.gr, Tel: +30 (81) 391704, Fax: +30 (81) 391601

Technical Report FORTH-ICS / TR-206 --- August 1997

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 a dense optical flow field, which amounts to solving the ill-posed correspondence problem. In this work, independent motion detection is formulated as a problem of robust parameter estimation applied to the visual input acquired by a rigidly moving observer. The proposed method automatically selects a planar surface in the scene and the residual planar parallax normal flow field with respect to the motion of this surface is computed at two successive time instants. The two resulting normal flow fields are then combined in a linear model. The parameters of this model are related to the parameters of self-motion (egomotion) and their robust estimation leads to a segmentation of the scene based on 3D motion. The method avoids a complete solution to the correspondence problem by selectively matching subsets of image points and by employing normal flow fields. Experimental results demonstrate the effectiveness of the proposed method in detecting independent motion in scenes with large depth variations and unrestricted 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* [12, 28]. 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 most man-made visual systems are usually in continuous motion.

In the case of a moving observer, IMD has often been approached as a problem of segmenting the 2D motion field that is computed from a temporal sequence of images. Wang and Adelson [33] 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 [20] estimate the parameters of an affine model of 2D motion, assuming that the estimation of the model parameters will not be considerably affected by the presence of small independently moving objects. IMD is then achieved by determining the points where the deviation of the measured from 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 can be large and, therefore, 2D methods may detect discontinuities that are not only due to motion, but also due to 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 regarding the depths of scene points are introduced. This in turn requires certain assumptions to be 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 [34] present an iterative method for recovering the 3D motion and structure of independently moving objects from a sparse set of velocities obtained from a pair of calibrated, parallel cameras. Lobo and Tsotsos [16] use a constraint defined with respect to three collinear image points to estimate the egomotion from a dense optical flow field and then detect independently moving objects having small spatial extent. Other assumptions that are commonly made by existing methods are related to the motion of the observer, to the structure of the viewed scene, or both. Sharma and Aloimonos [24] and Clarke and Zisserman [7] 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 [30] 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. Sinclair [26] assumes that surfaces are localy planar and describes the motion of 3D points in terms of their angular velocity relative to the camera. His method detects independent motion that violates the epipolar constraint and recovers the orientations of the normals of planar patches. 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 Argyros et al [4], the combination of depth and motion information extracted by a binocular observer permits the elimination of depth from the motion equations. This leads to a linear model involving the 3D motion parameters and the problem of IMD is then solved by estimating the linear model with robust regression methods. Although [2, 3, 4] avoid any assumptions related to the egomotion or the scene structure and do not require the correspondence problem to be solved, their main disadvantage is that they assume that normal flow can be computed from a pair of stereo images, an assumption that is valid in special cases only.

In order to overcome the limitations of existing methods, this paper proposes a new method for IMD. This method is based on two key observations. The first is that, although an accurate solution to the correspondence problem in the general case is very difficult, the problem can be solved with satisfactory accuracy in special cases, such as those involving corners or points belonging to a planar surface. The second observation is that the *residual parallax field* that remains after the registration of the images of a planar surface in two frames is an epipolar field. The proposed method exploits the information contained in the *normal residual field*, the component of residual motion in the direction of the image gradient. This field is less informative compared to the full residual flow, but can be more accurately computed from a temporal sequence of images. The combination of two such residual normal flow fields allows the elimination of the depth variables from the 3D motion equations, which in turn leads to the derivation of a model that is linear in the 3D motion parameters. IMD is then handled by applying a robust estimator to solve for the parameters of the linear model. Points that conform to the estimated model are labeled as moving due to the motion of the observer, while points that are characterized as outliers during the estimation process are labeled as independently moving. The proposed method assumes an observer that moves rigidly with unrestricted translational and rotational egomotion. Independent motion can be rigid or non-rigid and no calibration information is necessary.

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 approach. Section 3 develops a technique for identifying the dominant planar surface in a scene. The estimation of the motion of the dominant plane is outlined in Section 4. Section 5 discusses the decomposition of rigid image motion in terms of the motion of a planar surface and a residual parallax field. The proposed method for IMD is detailed in Section 6. Experimental results from the application of the method on real-world image sequences are presented in Section 7. Finally, the paper is concluded in Section 8.

2 Preliminaries

Before proceeding with the description of the proposed method, issues related to motion representation are discussed. In addition, a brief discussion of 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, 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 [17]:

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

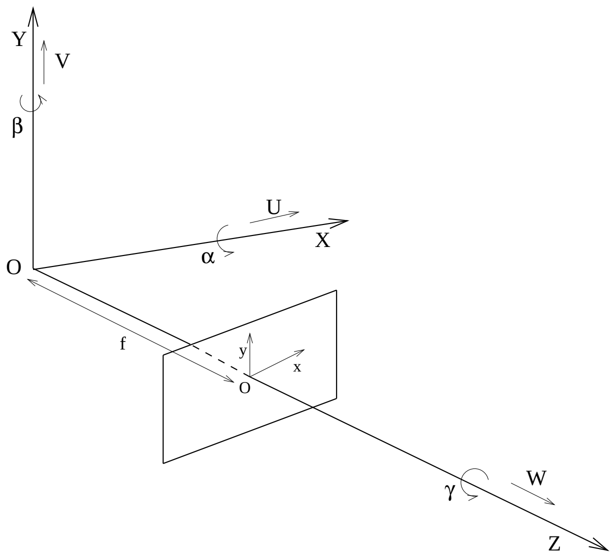


Figure 1: The camera coordinate system.

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

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 [10], 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 [18]. 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 given by the well known *optical flow constraint equation*, originally developed by Horn and Schunk in [11]:

$$I_x u + I_y v = -I_t \tag{2}$$

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 expresses the *aperture problem*, according to which local measurements supply only one 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 dense optical flow, but rather on the combination of the optical flow of a planar surface and the *normal flow* field for the whole image. As it will be shown in Section 4, once a planar surface in the scene has been identified, the problem of estimating its optical flow is a well-posed problem. On the other hand, the normal flow field is the projection of the optical flow field in the direction of image gradients. It can be shown that 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 [31]. However, normal flow is a good approximation to normal motion at points where the image gradient has a large magnitude [31]. Normal flow vectors at such points can be used as a robust input to 3D motion perception algorithms.

2.2 Robust Regression

Regression analysis (fitting a model to noisy data) is a very important subfield of statistics. In the general case of a linear model given by the relation $y_i = x_{i1}\theta_1 + \ldots + x_{ip}\theta_p + e_i$, the problem is to estimate the parameters θ_k, $k = 1, \ldots, p$, from the observations y_i, $i = 1, \ldots, n$, and the explanatory variables x_{ik}. The term e_i represents the error in each of the observations. In classical applications of regression, e_i is assumed to be normally distributed with zero mean and unknown standard deviation. Let $\hat{\theta}$ be the vector of estimated parameters $\hat{\theta}_1, \ldots, \hat{\theta}_p$. Given these estimates, predictions can be made for the observations as $\hat{y}_i = x_{i1}\hat{\theta}_1 + \ldots + x_{ip}\hat{\theta}_p$. Thus, a residual between the observation and the value predicted by the model may be defined as $r_i = y_i - \hat{y}_i$.

Traditionally, $\hat{\theta}$ is estimated by the least squares (LS) method, which is popular due to its low computational complexity. LS involves the solution of a minimization problem, namely:

$$Minimize \sum_{i=1}^{n} {r_i}^2 \tag{3}$$

and achieves optimal results if the underlying noise distribution is Gaussian with zero mean. However, in cases where the noise is not Gaussian, the LS estimator becomes unreliable. The LS estimator becomes highly unreliable also in the presence of outliers, that is observations that deviate considerably from the model representing the rest of the observations. One criterion for measuring the tolerance of an estimator with respect to outliers is its *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. As an example, LS has a breakdown point of 0%, because a single outlier may have a substantial impact on the estimated parameters.

The *Least Median of Squares* (LMedS) estimator was originally proposed by Rousseeuw [22] and is able to handle data sets containing large portions of outliers. LMedS involves the solution of a non-linear minimization problem, namely:

$$Minimize\{median_{i=1,\dots,n} {r_i}^2\} \tag{4}$$

Qualitatively, LMedS tries to estimate a set of model parameters that best fit the *majority* of the observations, while LS tries to estimate a set of model parameters that best fit all the observations. The above statement gives an indication of the difference in the behavior of the two estimators. The presence of some outliers in a set of observations will not influence LMedS estimation, as long as the majority of the data fit into the particular model. More formally, LMedS has a breakdown point of 50%, a characteristic which makes it particularly attractive for the purposes of this work. Another important property of LMedS is that it adapts automatically to the noise levels of the observations. The better the estimated model fits to the observations, the smaller the median residual is and, therefore, the finer the outlier detection becomes.

LMedS minimization is solved by a search in the space of possible estimates generated by the data. Since this search space is usually prohibitively large, a Monte-Carlo type of speedup technique is employed, in which a certain probability of error is tolerated.

3 Dominant Plane Extraction

The traditional approach for identifying planar regions using two images of a scene has been to recover the depth of each point in the field of view and then segment the resulting depth map into planes. This process however, involves computations that are numerically unstable and requires difficult problems such as point correspondence and camera calibration to be solved. To avoid these difficulties, Sinclair et al [27] have proposed a method for identifying coplanar sets of corresponding points, using simple results from projective geometry. Based on [27], the dominant plane in a scene is extracted using a method which is briefly outlined in the following subsections.

3.1 The Invariants of Five Coplanar Points

A well known result from projective geometry [19] is that groups of five corresponding coplanar points give rise to two projective invariants [1]. Those invariants are expressed by

$$I_1 = \frac{|M_{124}||M_{135}|}{|M_{134}||M_{125}|} \quad , \quad I_2 = \frac{|M_{241}||M_{235}|}{|M_{234}||M_{215}|}, \tag{5}$$

where $|M_{ijk}|$ denotes the determinant of the matrix whose columns are the vectors x_i, x_j, x_k formed by the homogeneous coordinates of three image points, i.e. $M_{ijk} = (x_i, x_j, x_k)$. Both I_1 and I_2 degenerate when any three of the five points are collinear. To test whether a set of five points imaged in two views satisfy the above invariants, a statistical test based on the variance in the values of the invariants is employed. This variance is estimated from the variances in the positions of the points in an image. The reader is referred to [27] for more details.

3.2 The Plane Homography

Another important concept from projective geometry is the *plane homography* (also known as plane projectivity or plane collineation) $\mathbf{H}$, which relates two uncalibrated views of a plane in three dimensions. Each plane Π in the world defines a nonsingular matrix $\mathbf{H}$ which transforms the projection $\mathbf{m}$ of a point in Π to the corresponding $\mathbf{m}'$, according to the equation [8, 21]:

$$\mathbf{m}' = \mathbf{H}\mathbf{m}. \tag{6}$$

In the previous equation, $\mathbf{m}$ and $\mathbf{m}'$ are homogeneous 3×1 vectors and $\mathbf{H}$ is a 3×3 matrix. Since $\mathbf{H}$ is defined with respect to projective (homogeneous) coordinates, it is defined up to an unknown scale factor. This implies that $\mathbf{H}$ has 8 degrees of freedom and since each pair of corresponding coplanar points provides 2 constraints, 4 pairs of corresponding coplanar points in general position (no three points are collinear) suffice for estimating it.

Assuming a set of N pairs of corresponding coplanar points, the plane homography $\mathbf{H}$ that they define can be estimated as follows: The N equations of the type shown in Eq. (6) can be written more compactly as $\mathbf{Ah} = 0$, where $\mathbf{A}$ is a $(2 * N) \times 9$ matrix and $\mathbf{h}$ a 9×1 vector. The plane homography is then estimated from the solution of the following minimization problem:

$$min_{\mathbf{h}} ||\mathbf{Ah}||^2 \quad subject\ to \quad ||\mathbf{h}||^2 = 1, \tag{7}$$

[1] Projective invariants are quantities that remain unchanged after projective transformations.

where $|| \ ||$ denotes the vector 2-norm. The solution to this problem is known to be the eigenvector of the matrix $\mathbf{A}^T\mathbf{A}$ that corresponds to the smallest eigenvalue, where $\mathbf{A}^T$ is the transpose of $\mathbf{A}$. Similar to what noted in [9, 36], $\mathbf{A}^T\mathbf{A}$ is inhomogeneous in image coordinates, and, therefore, ill-conditioned. To improve its condition number and to derive a more stable linear system, the coordinates of the set of corresponding points are normalized by a pair of linear transformations $\mathbf{L}$ and $\mathbf{L}'$ as follows: $\mathbf{L}$ defines a translation of the points in the first image, such that their centroid is brought to the origin of the coordinate system, followed by an isotropic scaling that maps the average point coordinates to $(1, 1, 1)$. $\mathbf{L}'$ is defined similarly for points in the second image. These transforms result in a more stable system, from which a homography matrix $\hat{\mathbf{H}}$ can be estimated. $\mathbf{H}$ is then computed from $\hat{\mathbf{H}}$ as $\mathbf{L}'^{-1}\hat{\mathbf{H}}\mathbf{L}$.

Since the set of normalized matching pairs that is given as input to the estimation process is very likely to contain errors, care must be taken so that these errors do not corrupt the computed estimate. Thus, instead of using all N points to estimate $\mathbf{H}$, the LMedS estimator is employed to find an estimate that is consistent with the majority of the matched points. Using a predetermined number of iterations, LMedS picks random samples of matching pairs and computes an estimate of $\mathbf{H}$ from each of them. The estimate that yields the smallest median error is returned as the plane homography which best fits the set of matched points.

3.3 Iterative Algorithm for the Extraction of Planes

Based on the above discussion, an iterative method for extracting the dominant plane can now be described. First, the SUSAN corner detector [29] is used to extract a set of corners from a pair of images. Corners are distinct image features that can be accurately localized and correspond to 3D scene elements appearing in consecutive images. Here it is assumed that the two images have been acquired from considerably different locations in the 3D space. Such an image pair can be captured either by the two cameras of a binocular system, or by the single camera of a monocular system at two instants that are far apart in time. The extracted corners are then matched using a similarity criterion based on normalized cross-correlation. The matching algorithm is based on that proposed in [35]. A random sample consisting of five pairs from the set of matched corners is then formed. If the selected corners satisfy the invariants in Eq. (5), they are likely to belong to the same plane. To ensure that the selected corners are sufficiently far apart so that the invariants and the corresponding plane homography are not swamped by noise, a bucket-based sampling technique similar to that discussed in [36] is employed. Next, the plane homography corresponding to the selected corners is computed as described previously. To verify that the five selected points lie on the same plane, the estimated plane homography is used to find more coplanar points. For every corner in one image, the plane homography can predict the location of the corresponding corner in the second image. If this location is sufficiently close to the true location of the matching corner, the corner in question is assumed to be coplanar with the corners in the selected sample. If the number of coplanar points identified during this step is above a threshold, the method concludes that a plane has indeed been found. The corresponding plane homography is then re-estimated using the whole set of coplanar points and this set is removed from further consideration. The sampling process iterates until either the number of corners that have not been assigned to a plane drops below a threshold or a predetermined number of iterations is completed.

When the iterative algorithm terminates, a set of planes along with their homographies have been computed. The application of Eq. (6) to each point in the first view warps the second view with respect to the first and registers the image of the corresponding plane in the two views. Change detection between the first and the warped second view can label image points as changing in the two views or not. Points that remain unchanged belong to the plane under consideration. To account for the fact that typical change detection algorithms fail in uniform, textureless areas, a pixel is assumed to belong to a plane when it is labeled as not changing by the change detection algorithm and the magnitude of its gradient is above some threshold. The plane having the largest number of points is declared to be the dominant one. As it will be clear from the following sections, the result of change detection does not have to be very accurate, since the part of the proposed method for IMD that makes use of the location of the dominant plane is tolerant to errors. In our implementation, the change detection algorithm described in [28] is employed. This algorithm is based on a test regarding the variance of the intensity ratios in small neighborhoods of two images.

4 Robust Parametric Estimation of Optical Flow

The problem of estimating 2D image velocity, or optical flow, from image sequences is generally very difficult. This difficulty mainly stems from the fact that transparencies, specular reflections, shadows, occlusions, depth boundaries and independent motions give rise to discontinuities in the optical flow field [18]. This in turn implies that an optical flow field is typically only piecewise smooth [5]. Since the estimation of optical flow involves the combination of constraints arising from an image region, no guarantee is given that the selected region will contain only a single motion. In other words, the primary difficulty of most optical flow estimation techniques is that they lack any information regarding the region of support of a particular motion. This problem is referred to in [5] as the *generalized aperture problem.*

In the case that an image region is known to correspond to a plane in the scene, the optical flow within the region can be accurately modeled as a parametric function of the image coordinates [1]. More specifically, assuming that the equation of the imaged plane in image coordinates is $\frac{1}{Z} = px + qy + r$, substitution into Eq. (1) yields an eight parameter model for optical flow. This model is known as the *quadratic* model since it contains terms that are of degree two in the image coordinates:

$$\begin{aligned} u &= a + bx + cy + gx^2 + hxy \\ v &= d + ex + fy + gxy + hy^2 \end{aligned} \tag{8}$$

At this point, it should be noted that, if the camera is not calibrated, the unknown intrincic parameters (i.e. focal lengths and location of principal point) are absorbed in the eight parameters $a, \ldots, h$. By employing the quadratic model, the estimation of optical flow amounts to the estimation of the eight parameters involved. Substitution of Eq. (8) into Eq. (2), permits the derivation of a model relating the planar flow parameters to the spatiotemporal intensity derivatives. This model is linear in the parameters to be estimated and is overdetermined, since each point of the plane contributes one constraint regarding the eight unknown parameters. To account for errors in the computation of derivatives, violations of the intensity conservation assumption, errors in the determination of the region corresponding to the image of the plane, etc, the LMedS estimator is again employed to give a robust estimate of the parameters satisfying the majority of the constraints. This ‘‘robustification’’ of the optical flow estimation problem has already been suggested by Black and Anandan [5], with the major difference being that they employed M-estimators which are less robust compared to LMedS that is employed in this work.

5 Planar Parallax

Most motion analysis methods express rigid image motion as the sum of two displacement fields, namely a translational and a rotational one. Recently, however, it has been shown that if image motion is expressed in terms of the motion of a parametric surface and a *residual parallax field*, important problems in motion analysis become considerably simpler [15, 23, 25, 14]. In this section, the equations describing the residual field are derived, assuming that the employed parametric surface is a plane.

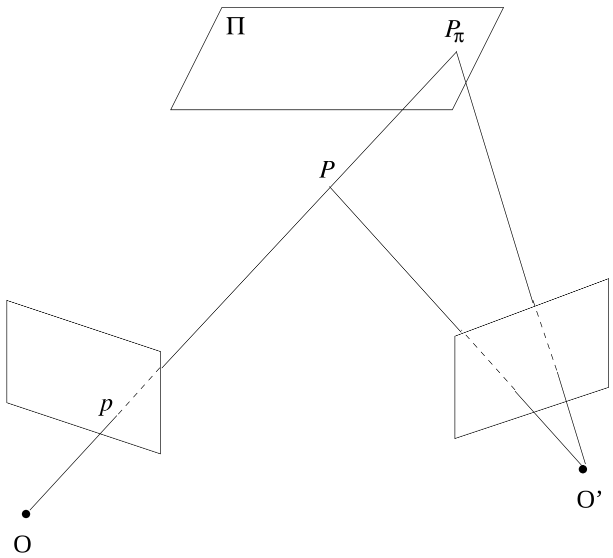


Figure 2: Planar Parallax.

Let (u, v) be the displacement field between two images $\mathcal{I}_t$, $\mathcal{I}_{t+dt}$ acquired at time instants t and $t + dt$ respectively. Let also Π be a 3D plane in the viewed scene and let (u_π, v_π) be the 2D motion vector of a single point belonging to Π. As shown in Section 4, (u_π, v_π) is defined by a linear model with eight parameters. Warping $\mathcal{I}_t$ towards $\mathcal{I}_{t+dt}$ according to (u_π, v_π) will register $\mathcal{I}_t$ and $\mathcal{I}_{t+dt}$ over regions of Π, while regions not belonging to Π will be unregistered. According to Eq. (1), the residual flow (u^r, v^r) between the warped $\mathcal{I}_t$ and $\mathcal{I}_{t+dt}$ will be [14, 6]

$$\begin{aligned} u^r = \quad & u - u_\pi \quad = (xW - Uf)\left(\frac{1}{Z} - \frac{1}{Z_\pi}\right) \\ v^r = \quad & v - v_\pi \quad = (yW - Vf)\left(\frac{1}{Z} - \frac{1}{Z_\pi}\right) \end{aligned} \tag{9}$$

where $\frac{1}{Z_\pi}$ is the depth of the 3D plane at pixel (x, y). As can been seen from Eq. (9), the residual flow field is purely translational. This is because the rotational component of the motion does not depend on depth and is thus canceled by the warping step. Consequently, all optical flow vectors of the residual flow point towards the FOE[2]. In a similar manner, the *residual normal flow field* between the warped $\mathcal{I}_t$ and $\mathcal{I}_{t+dt}$ is given by:

$$u_{nr} = \quad u^r n_x + v^r n_y \quad = \{(xW - Uf)n_x + (yW - Vf)n_y\}\left(\frac{1}{Z} - \frac{1}{Z_\pi}\right) \tag{10}$$

where (n_x, n_y) is the unit vector in the direction of the intensity gradient.

Figure 2 depicts geometrically the notion of planar parallax. O and O' are the camera focal points in the two views, Π a plane in the viewed scene and P_π, P are two 3D points

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

with P_π belonging to Π. Since P_π and P project to the same image point p in the first view, their corresponding optical flow vectors share the same rotational component.

6 Independent Motion Detection

Consider a rigid observer that is moving with unrestricted egomotion in 3D space. Due to this motion, a reliable normal flow vector can be computed at each point where the image intensity gradient is sufficiently large. Let (n_x, n_y) be the unit vector in the gradient direction. The magnitude u_n of the normal flow vector is given by $u_n = un_x + vn_y$, which, by substitution from Eq. (1), yields:

$$\begin{aligned} u_n &= -n_x f\frac{U}{Z} - n_y f\frac{V}{Z} + (xn_x + yn_y)\frac{W}{Z} \\ &+ \left\{\frac{xy}{f}n_x + \left(\frac{y^2}{f} + f\right)n_y\right\}\alpha \\ &- \left\{\left(\frac{x^2}{f} + f\right)n_x + \frac{xy}{f}n_y\right\}\beta + (yn_x - xn_y)\gamma \end{aligned} \tag{11}$$

Equation (11) 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 constraints regarding $n+6m$ unknowns. Evidently, the problem cannot be solved without any additional information on depth.

Let us now suppose that at least one of the surfaces in the scene is planar or can be well approximated by a plane. This assumption is often satisfied in practice, especially in scenes containing man-made objects [32]. Using the technique described in Section 3, the dominant plane in the scene can be extracted. Following this, the parametric model describing the motion of this plane can be estimated as described in Section 4. The residual planar parallax flow can then be computed from Eq. (9). Irani and Anandan [13] have recently described a method for IMD that computes the *relative projective 3D structure* from this residual parallax flow. Their method, however, requires the computation of a dense optical flow field, a difficult problem in its own right. Noting that the residual flow field is translational, another approach to detect independent motion is to locate the FOE and then, similar to [24], label points that violate the epipolar constraint as independently moving. The major drawback of this approach is that it depends critically on the correctness of the estimated FOE. To avoid this problem, the proposed method for IMD does not attempt to estimate the FOE. Instead, it combines the information from two residual normal flow fields computed at consecutive time instants.

Assume that three consecutive images $\mathcal{I}_{t-dt}$, $\mathcal{I}_t$ and $\mathcal{I}_{t+dt}$ are captured at time instants $t-dt$, t and $t+dt$ respectively. Let $\mathcal{I}_0$ be a fourth image that along with $\mathcal{I}_t$ permits the extraction of the dominant plane. Also, let u_{nr} be the residual normal flow computed by warping $\mathcal{I}_t$ towards $\mathcal{I}_{t+dt}$ using the motion of the dominant plane. Similarly, let u'_{nr} be the residual normal flow computed by warping $\mathcal{I}_t$ towards $\mathcal{I}_{t-dt}$ using the dominant plane. According to Eq. (9), u_{nr} and u'_{nr} are given by

$$\begin{aligned} u_{nr} &= \{(xW - Uf)n_x + (yW - Vf)n_y\}\,(\frac{1}{Z} - \frac{1}{Z_\pi}) \\ u'_{nr} &= \{(xW' - U'f)n_x + (yW' - V'f)n_y\}\,(\frac{1}{Z} - \frac{1}{Z_\pi}) \end{aligned} \tag{12}$$

where (U, V, W) and (U', V', W') are the translational velocity vectors for the displacement between t and $t+dt$ and t and $t-dt$ respectively.

Both residual normal flow fields given by Eqs. (12) are defined in the the same reference frame, namely $\mathcal{I}_t$. This implies that at each point (x, y) of $\mathcal{I}_t$ having considerable gradient magnitude, two normal flow vectors along the same direction (n_x, n_y) can be computed. Solving the first of Eqs. (12) for $\frac{1}{Z} - \frac{1}{Z_\pi}$ and substituting into the second results into the following equation

$$\begin{aligned} W(xn_x + yn_y)u'_{nr} \;-\; & Ufn_xu'_{nr} - Vfn_yu'_{nr} - \\ & W'(xn_x + yn_y)u_{nr} + U'fn_xu_{nr} + V'fn_yu_{nr} = 0, \end{aligned} \tag{13}$$

in which the terms related to depth have been eliminated. The above equation is linear in the variables $\phi_1 = W$, $\phi_2 = Uf$, $\phi_3 = Vf$, $\phi_4 = W'$, $\phi_5 = U'f$, $\phi_6 = V'f$. These variables involve the 3D motion parameters and the camera focal length. Assuming that the dominant plane is not independently moving, violations of Eq. (13) signal the presence of independently moving objects. LMedS estimation can be applied to a set of observations of the model of Eq. (13) as a means to estimate the parameters ϕ_i, $1, \ldots, 6$. To avoid the trivial solution $\phi_i = 0$, the solutions tried by LMedS are computed with an eigenvector technique that imposes the constraint $||(\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6)||^2 = 1$. 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_{nr} and/or u'_{nr} 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 vectors 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. (13) 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. The proposed algorithm for IMD is summarized in the block diagram of Fig. 3. The postprocessing step is described in the following subsection.

When implementing the method presented in the preceding paragraphs, the residual normal flow can be computed without actually warping the first image towards the second according to the estimated planar flow. Knowledge of the eight parameters in Eq. (8) enables the prediction of the normal flow that would result if the dominant plane covered the whole visual field. The residual normal flow can then simply be estimated as the difference between the normal flow computed directly from the pair of input images and the predicted planar normal flow. Normal flow between a pair of input images is computed from the spatiotemporal derivatives I_x, I_y and I_t of the image intensity function. To reduce the effects of noise, images are smoothed by convolving them with a 3×3 Gaussian prior to the computation of derivatives.

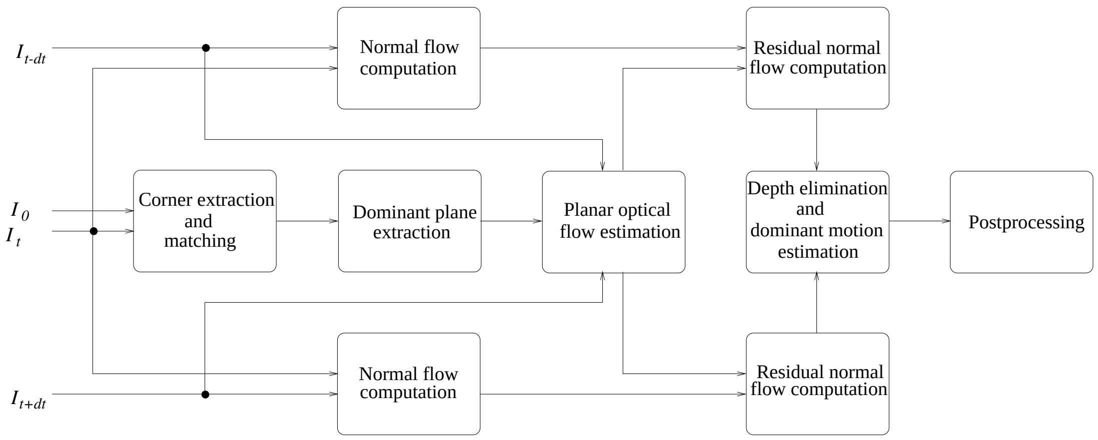


Figure 3: Block diagram of the proposed method (see text for explanation).

6.1 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 reasons why the points of a motion segment may 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. at 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 onto 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.

6.2 Egomotion Estimation

Besides the inlier/outlier characterization, LMedS provides estimates $\hat{\phi}_i$ of the parameters ϕ_i in the linear model of Eq. (13). Since these parameters are directly related to the translational motion of the observer, the FOE can be estimated as $(\frac{\hat{\phi}_2}{\hat{\phi}_1}, \frac{\hat{\phi}_3}{\hat{\phi}_1})$. The rotational motion parameters can then be recovered from the eight parameters defining the optical flow corresponding to the dominant plane [14].

7 Experimental Results

The proposed method has been evaluated experimentally with the aid of several real-world image sequences. During the course of all experiments, quantitative information regarding camera motion and calibration parameters was not available. Due to space limitations, only two of the conducted experiments are reported here.

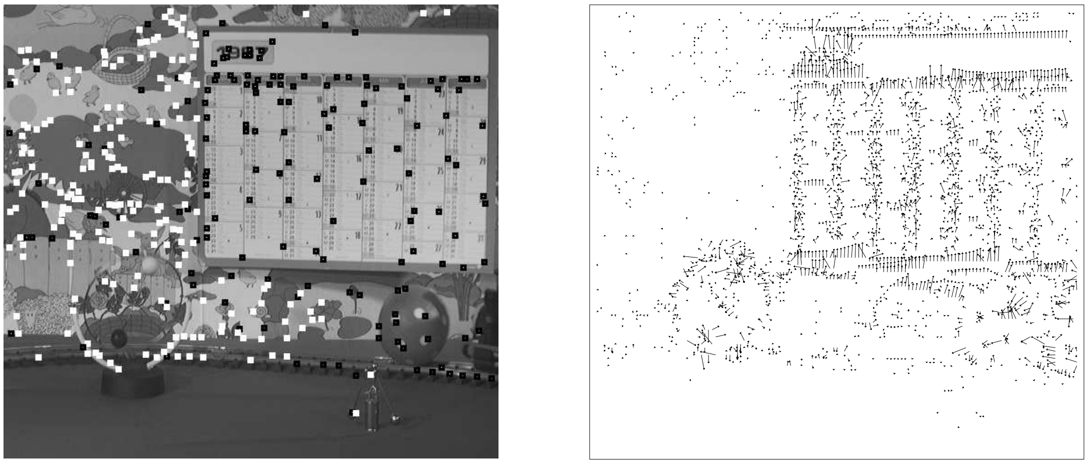
The first experiment is based on the well known ‘‘calendar’’ image sequence. Frames 2 and 30 of this sequence are shown in Fig. 4.

(a)

(b)

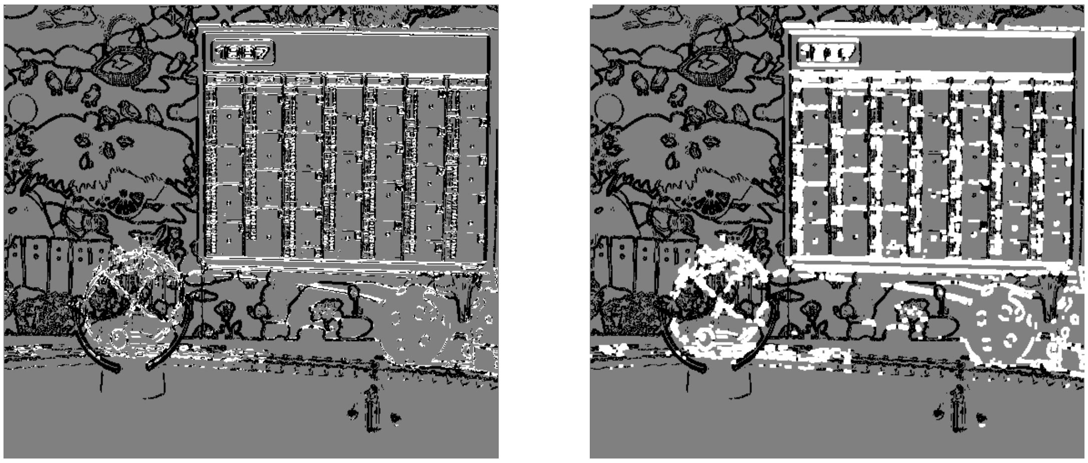
Figure 4: Frames (a) 2, and (b) 30 of the ‘‘calendar’’ sequence.

In this sequence, the camera is panning with a right to left direction and the viewed scene consists of a planar background and a nonplanar foreground. The background contains a stationary wall and a calendar that is independently moving upwards. The foreground contains three independently moving objects. A pair of spheres is rotating in the left side of the scene, while a ball followed by a toy train are moving in a right to left direction. The dominant plane was extracted using frames 2 and 30. Corners belonging to the dominant plane are marked with white rectangles in Fig. 5(a), while all other corners are black.



(a) (b)

Figure 5: (a) Corners belonging to the dominant plane for the ‘‘calendar’’ sequence, (b) residual normal flow field for frames 2-3.



(a) (b)

Figure 6: Motion segmentation for the ‘‘calendar’’ sequence (a) before and, (b) after postprocessing.

The pair of residual parallax normal flow fields is computed between frames 2 - 3 and 2 - 1. The residual parallax normal flow for frames 2 - 3 is shown in Figure 5(b). As can be seen from this figure, the residual flow field is zero over the area corresponding to the dominant plane, indicating that the dominant plane has been successfully registered. Figure 6 illustrates the results of motion segmentation on the ‘‘calendar’’ sequence.

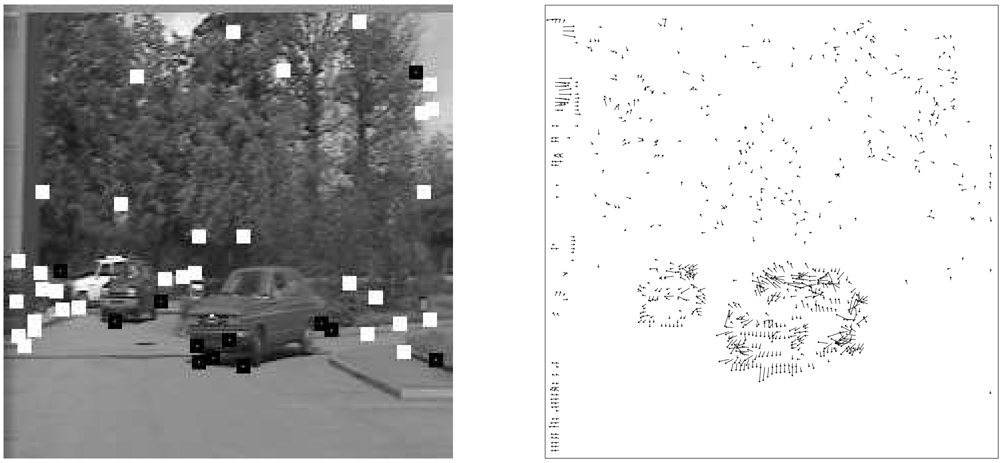
Figure 6(a) shows the intermediate segmentation results. Black color corresponds to egomotion and white color corresponds to independent motion. Gray color corresponds to points where no decision can be made, due to low image gradient and, therefore, lack of normal flow vectors. It can be verified that the largest concentration of white (i.e. independently moving) points is indeed over the regions of the independently moving objects. Note that independent motion was not detected along the vertical edges of the calendar. This is because the intensity gradient is perpendicular to the direction of motion on these edges, which results in the corresponding normal flow vectors being equal to zero. The elongated areas below the calendar that are marked as independently moving are actually shadows, cast by the calendar and the rotating spheres, that are moving during time. Figure 6(b) presents the same result 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. It is clear that after this step, the bodies of the four independently moving objects have been successfully identified as such. An MPEG video demonstrating the results of applying the proposed method on the first 10 frames of the ''calendar'' sequence can be found at http://www.ics.forth.gr/proj/cvrl/demos/lourakis/IMD/calendar.mpg

The second experiment concerns the ''cars'' image sequence. Frames 5 and 20 of this sequence are shown in Fig. 7.

(a) (b)

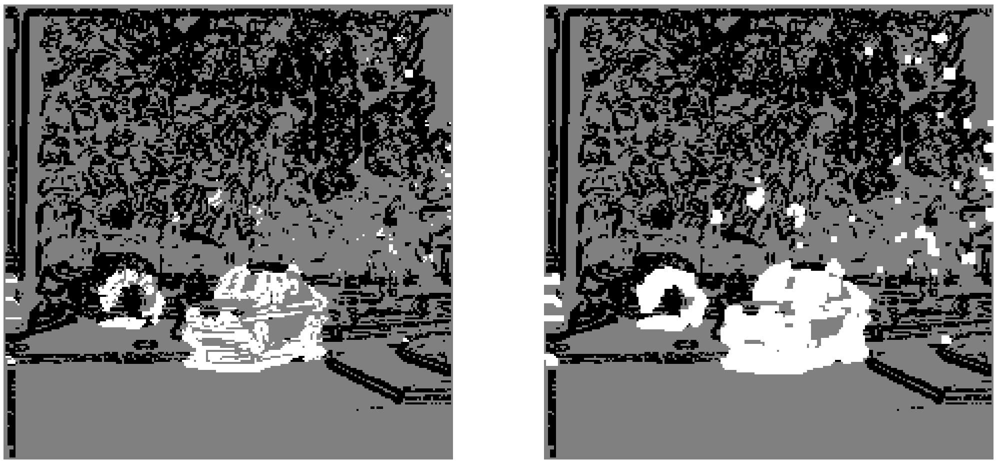
Figure 7: Frames (a) 5, and (b) 20 of the ''cars'' sequence.

In this sequence, the camera is again panning with a right to left direction. The two dark gray cars in the foreground move independently while the white car on the far left is stationary. A few trees in the background form an approximately planar surface. Frames 5 and 20 were used to extract the dominant plane. Figure 8(a) shows corners belonging to the dominant plane marked with white rectangles, while all other corners are black.



(a) (b)

Figure 8: (a) Corners belonging to the dominant plane for the ‘‘cars’’ sequence, (b) residual normal flow field for frames 5-6.



(a) (b)

Figure 9: Motion segmentation for the ‘‘cars’’ sequence (a) before and, (b) after postprocessing.

Frames 5 - 6 and 5 - 4 are used to compute the pair of residual parallax normal flow fields. Figure 8(b) shows the residual parallax normal flow computed from frames 5 - 6. The results of motion segmentation on the ‘‘cars’’ sequence before and after postprocessing are illustrated in Figures 6(a) and 9(b) respectively. Black color corresponds to egomotion and white color corresponds to independent motion. Gray color corresponds to points with low intensity gradient, and thus without normal flow vectors. As it can be seen from Fig. 9, the two cars are correctly identified as independently moving. Moreover, the independent motions of small parts of the tree foliage are also detected.

8 Conclusions

Artificial seeing systems should operate in dynamic environments that consist of both stationary as well as moving objects. The perception of independent 3D motion is crucial because it provides useful information on where attention should be focused and, possibly, maintained. In this paper, independent 3D motion detection was based on a pair of residual parallax normal flow fields that are computed by an observer that moves freely in the 3D space. The proposed method employs 3D motion models and is able to perform satisfactorily even in scenes with considerable depth variations. Both rigid and non-rigid independent motion can be detected. Moreover, apart from the requirement for the existence of a planar surface in the viewed scene, no further assumptions regarding the structure of the external world are made. The method avoids a complete solution to the ill-posed correspondence problem by matching only carefully selected sets of image points. To guard against errors caused by false matches, robust estimation techniques are employed. Experimental results from the application of the proposed method on real image sequences were also presented. Ongoing research aims at integrating 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.A. Argyros, M.I.A Lourakis, P.E. Trahanias, and S.C. Orphanoudakis. Independent 3D Motion Detection Through Robust Regression in Depth Layers. In *Proceedings of BMVC '96, Edinburgh, UK*, Sep. 9-12 1996.

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

[4] A.A. Argyros and S.C. Orphanoudakis. Independent 3D Motion Detection Based on Depth Elimination in Normal Flow Fields. In *Proceedings of CVPR '97, San Juan, Puerto Rico*, pages 672--677, Jun. 17-19 1997.

[5] M.J. Black and P. Anandan. The Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow Fields. *Computer Vision and Image Understanding*, 63(1):75--104, 1996.

[6] R. Cipola, Y. Okamoto, and Y. Kuno. Robust Structure from Motion Using Motion Parallax. In *Proceedings of ICCV'93*, pages 374--382, 1993.

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

[8] R. Hartley and R. Gupta. Computing Matched-epipolar Projections. In *Proceedings of CVPR'93*, pages 549--555, 1993.

[9] R.I. Hartley. In Defense of the 8-Point Algorithm. In *Proceedings of ICCV'95*, pages 1064--1070, 1995.

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

[11] B.K.P. Horn and B. Schunck. Determining Optical Flow. *Artificial Intelligence*, 17:185--203, 1981.

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

[13] M. Irani and P. Anandan. A Unified Approach to Moving Object Detection in 2D and 3D Scenes. In *Proceedings of ICPR'96*, pages 712--717, Vienna, Austria, 1996.

[14] M. Irani, B. Rousso, and S. Peleg. Recovery of Ego-Motion Using Region Alignment. *IEEE Trans. on PAMI*, 19(3):268--272, Mar. 1997.

[15] R. Kumar, P. Anandan, and K. Hanna. Direct Recovery of Shape from Multiple Views: A Parallax Based Approach. In *Proceedings of ICPR'94*, pages 685--688, Jerusalem, Israel, 1994.

[16] N. V. Lobo and J. K. Tsotsos. Computing Egomotion and Detecting Independent Motion from Image Motion Using Collinear Points. *Computer Vision and Image Understanding*, 64(1):21--52, July 1996.

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

[18] A. Mitiche and P. Bouthemy. Computation and Analysis of Image Motion: A Synopsis of Current Problems and Methods. *IJCV*, 19(1):29--55, Jul. 1996.

[19] J.L. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision.* MIT Press, Cambridge, MA, 1992.

[20] 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.

[21] L. Robert and O.D. Faugeras. Relative 3D Positioning and 3D Convex Hull Computation from a Weakly Calibrated Stereo Pair. *Image and Vision Computing*, 13(3):189--196, April 1995.

[22] P.J. Rousseeuw. Least Median of Squares Regression. *Journal of American Statistics Association*, 79:871--880, 1984.

[23] H. Sawhney. Simplifying Motion and Structure Analysis Using Planar Parallax and Image Warping. In *Proceedings of ICPR'94*, pages 403--408, Jerusalem, Israel, 1994.

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

[25] A. Shashua and N. Navab. Relative Affine Structure - Canonical Model for 3D From 2D Geometry and Applications. *IEEE Trans. on PAMI*, PAMI-18(9):873--883, Sep. 1996.

[26] D. Sinclair. Motion Segmentation and Local Structure. In *Proceedings of ICCV'93*, pages 366--373, 1993.

[27] D. Sinclair and A. Blake. Quantitative Planar Region Detection. *IJCV*, 18(1):77--91, Apr. 1996.

[28] K. Skifstad and R. Jain. Illumination Independent Change Detection for Real World Image Sequences. *Computer Vision, Graphics and Image Processing*, 46:387--399, 1989.

[29] S. M. Smith and J. M. Brady. SUSAN - A New Approach to Low Level Image Processing. *IJCV*, 23(1):45--78, May 1997.

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

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

[32] T. Viéville, C. Zeller, and L. Robert. Using Collineations to Compute Motion and Structure in an Uncalibrated Image Sequence. *IJCV*, 20(3):213--242, 1996.

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

[34] 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.

[35] Z. Zhang. A New and Efficient Iterative Approach to Image Matching. In *Proceedings of ICPR'94*, pages 563--565, Jerusalem, Israel, 1994.

[36] Z. Zhang. Determining the Epipolar Geometry and its Uncertainty: A Review. Technical Report 2927, INRIA, July 1996.