Statistical Detection of Independent Movement from a Moving Camera

P H S Torr and D W Murray
Robotics Research Group
Department of Engineering Science
Oxford University, Parks Road, Oxford, OX1 3PJ, UK

Abstract

This paper describes the use of a low level, computationally inexpensive closed form motion detector to define regions of interest within an image, based upon statistical measures. The algorithm requires only the first order properties of the image intensities and does not require known camera motion. It has been tested on a variety of real imagery. A b-spline snake is initialised on the occluding contours of this region of interest.

1 Introduction

Amongst the important tasks which rely on motion data is that of motion segmentation. This paper addresses the problem of how to detect a set of moving objects in the two dimensional projection of an otherwise rigid scene, given that the camera is moving in an arbitrary and unpredetermined manner. A closed form analytic solution is supplied for the detection of non-rigid motions given the spatio-temporal gradients.

The problem of motion segmentation has received considerable attention over the years. For example, Nelson [10] has described an algorithm designed to solve the object segmentation problem for the case of known camera translation and rotation. Burt and his co-workers [3] have developed a multi-scale pyramidal motion segmentation algorithm designed for use in conjunction with control of the sensor and the parameters of the algorithm. François and Bouthemy [7] have designed an algorithm that uses qualitative information about the camera motion to aid motion segmentation, using a Markow Random Field (MRF) approach to segment the scene into regions with common affine flows. Adiv [1] and Waxman and Duncan [15] make second order approximations to the flow field over small regions. They then use various methods to test the compatibility of these regions, in order to decide whether or not to merge them. Thompson, Mutch and Berzins [13] and Schunck [12] present algorithms for motion boundary detection based upon early edge detection algorithms using the Laplacian of the Gaussian.

In this paper we propose a global solution to the problem of motion segmentation, in order to overcome the problems posed by regions contributing a paucity of data (e.g. aperture problem), by considering the whole image. Many methods assume dense and accurate velocity fields as input and impose a continuity constraint upon the projected motion field identifying motion boundaries along lines where this continuity is violated. Without highly accurate estimates

BMVC 1992 doi:10.5244/C.6.9

of the projected motion, small parts of the image are unlikely to contain sufficient information to reconstruct full flow and its deformation parameters, thus numerical differentiation of the projected vector field or the merging of small areas will be very ill-conditioned. As in [4] we assume an affine background flow, and successive quantitative estimates are made about the affine deformation parameters of the background motion over the image. Parts of the image that do not accord with this estimate are identified as regions of interest. We use the first order intensity properties of the image as input to our algorithm, so that we do not throw any information away.

Our aim then is to partition the five dimensional space of discrete image points (pixels) and the spatio-temporal intensity gradients calculated at those pixels $\{(E_x, E_y, E_t, x, y)\}$ into disjoint sets corresponding to either the designated background or to foreground objects undergoing independent motions. To achieve this we

1. Fit a hyper-plane through the points in a given region using least-squares, assuming that the points undergo an affine transform (Sections 2.1 and 2.2);
2. Check for collinearity and appropriateness of the assumption (Section 2.3);
3. Identify the outliers to the fit (Section 2.1); and
4. Cluster the outliers to form regions of interest (Section 3).

2 Least squares fit to intensity gradients

Within this section we shall outline how to discover outliers from the flow predicted by the affine scene model. These outliers will usually arise from either noise or occlusions. The interested reader is referred to [2, 6] for a more thorough coverage of the theory and methods of diagnostic techniques.

2.1 Testing for Outliers within Least Squares

Given a set of equations

$$y_\alpha = \vec{\mathbf{d}}_\alpha \vec{\mathbf{b}}^T \qquad \alpha = 1 \ldots n \tag{1}$$

where y_α is a known variable, $\vec{\mathbf{d}}_\alpha$ is a known p dimensional vector and $\vec{\mathbf{b}}$ is an unknown p dimensional vector, termed the vector of coefficients We shall term y_α the *dependent* variable and $\vec{\mathbf{d}}_\alpha$ the vector of *independent* variables. Let $[\mathbf{D}]$ be a matrix whose rows are $\vec{\mathbf{d}}_\alpha$ then from equation 1 we can see that:

$$\vec{\mathbf{y}}^T = [\mathbf{D}]\vec{\mathbf{b}}^T \tag{2}$$

we can then use the pseudo inverse to solve for $\vec{\mathbf{b}}$:

$$\vec{\beta}^T = [[\mathbf{D}]^T[\mathbf{D}]]^{-1}[\mathbf{D}]^T\vec{\mathbf{b}}^T \tag{3}$$

where $\vec{\beta}$ is our estimate of the coefficients $\vec{\mathbf{b}}$ which is unknown.

The general procedure for assessing the influence of a given point in a regression analysis is to determine the changes that occur when the point is omitted. Several measures of influence exist in the literature. They differ in the particular regression result on which the effect of the deletion is measured, and the standardization used to make them comparable over observations. All the influence measures discussed can be computed from the results of a single regression. Below we discuss three influence measures, each of which takes account of the deletion of the ith observation or equation (e.g. what would be the solution to the set of equations 1 if we delete the equation $y_i = \vec{\mathbf{d}}_i \vec{\mathbf{b}}^T$) on some regression variable. Cook's D measures the effect on $\vec{\beta}$ our estimator of $\vec{\mathbf{b}}$. **DFFITS**–F measures the effects on our prediction of the dependent variable $\vec{\mathbf{y}}_i'$ given $\vec{\beta}$. **COVRATIO**–C which measures the effect on the variance-covariance matrix of the parameter of estimates.

Potentially influential points are data points that are far from the centre of the [**D**]-space. A measure of the distance of the ith data point from the centroid of all the points in [**D**]-space $\overline{\mathbf{d}}$ is provided by h_{ii}, the ith diagonal element of the hat matrix [**H**]. The hat matrix is derived as follows, from equation 3 we can see that:

$$\vec{\mathbf{y}}'^T = [\mathbf{D}] \left([\mathbf{D}]^T [\mathbf{D}] \right)^{-1} [\mathbf{D}]^T \vec{\mathbf{y}}^T \tag{4}$$

where $\vec{\mathbf{y}}'$ is our prediction of $\vec{\mathbf{y}}$ given $\vec{\beta}$. $[\mathbf{H}] = [\mathbf{D}] \left([\mathbf{D}]^T [\mathbf{D}] \right)^{-1} [\mathbf{D}]^T$ is termed the hat or the orthogonal projection matrix on the column space of [**D**]. h_{ii} is termed the *leverage* or *potential* of the ith case in that is gives an indication of the effect of y_i on $\hat{y}_i$, the closer h_{ii} is to 1 the smaller the residual e_i. The estimate of the variance of the dependent variable is

$$\hat{\sigma_t}^2 = \frac{\sum e_i^2}{n - p} \tag{5}$$

Where e_i is the ith element of the residual vector $\vec{\mathbf{e}}$. Note that $\vec{\mathbf{e}} = ([\mathbf{I}] - [\mathbf{H}])\vec{\mathbf{y}}$, $\mathbf{Var}(e) = \mathbf{Var}(([\mathbf{I}] - [\mathbf{H}])\vec{\mathbf{y}})$. Thus the residuals do not have common variance. The heterogeneous variances in the residuals are corrected by dividing each residual by an estimate of its standard deviation given by the square root of the diagonal elements of $([\mathbf{I}] - [\mathbf{H}])\hat{\sigma_t}^2$. Standardized (or internally Studentized) residuals s are given by [11]:

$$s_i \stackrel{\text{def}}{=} \frac{\hat{e_i}}{\hat{\sigma_t}\sqrt{1 - h_{ii}}} \tag{6}$$

Belsley, Kuh and Welsch [2] suggest standardizing the residuals with an estimate of its standard deviation independent of the residual. This is accomplished using $\sigma_{(i)}$, the estimate of the standard deviation without the ith observation which can be obtained by:

$$(n - p - 1)\sigma_{(i)}^2 = (n - p)\sigma_t^2 - \frac{e_i^2}{1 - h_{ii}} \tag{7}$$

Let t_i be the ith Studentized residual such that [2]

$$t_i \stackrel{\text{def}}{=} \frac{e_i}{\sigma_{(i)}(1-h_{ii})^{\frac{1}{2}}} = s_i \left(\frac{n-p-1}{n-p-s_i^2} \right)^{\frac{1}{2}} \tag{8}$$

t_i will follow a t-distribution with $n-p-1$ degrees of freedom if the errors in y_i are normally distributed. Cook's D [5] test is designed to measure the shift in $\vec{\beta}$ when a particular observation is omitted. Cook's D_i is defined as

$$D_i \stackrel{\text{def}}{=} \frac{(\vec{\beta}_{(i)} - \vec{\beta})^T([\mathbf{D}]^T[\mathbf{D}])(\vec{\beta}_{(i)} - \vec{\beta})}{p\sigma_t^2} = \frac{s_i^2}{p} \left(\frac{h_{ii}}{1-h_{ii}} \right) \tag{9}$$

Where $\vec{\beta}_{(i)}$ is the set of parameters fitted to the data without the ith observation included. D_i has approximately an F-distribution thus $D_i \approx F_{(\alpha,p,n-p)}$. Cook [5] suggests that if $\alpha = .50$ from omitting a single data point then this is significant. The 50th percentile for F is 1.0 when the numerator and denominator are large thus a value of D_i near 1.0 is significant. This is extreme and the literature suggest a more modest threshold of $\frac{4}{n}$, where we recall n is the number of observations. The **DFFITS** [2] statistic F can be computed from the Studentized residual. F_i gives a measure in the change of $\vec{\mathbf{y}}'$ when the ith observation is not included in the estimation of $\vec{\beta}$.

$$F_i \stackrel{\text{def}}{=} \frac{y_i' - y'_{(i)i}}{\sigma_{(i)}\sqrt{h_{ii}}} = \left(\frac{h_{ii}}{1-h_{ii}} \right)^{\frac{1}{2}} t_i \tag{10}$$

where $y'_{(i)i} = x_i\vec{\beta}_{(i)}$ i.e. the estimated y_i' for the ith observation where the ith observation was not used to estimate $\vec{\beta}$. After tests it was found that the significant areas identified by Cook's D and F_i are very nearly identical. An approximation to the impact of the ith observation on the variance of the estimated coefficients is measured by the ratio of the determinants of the two variance-covariance matrices **COVRATIO** $= C$.

$$C \stackrel{\text{def}}{=} \frac{\left| \sigma_{(i)}^2 \left[[\mathbf{D}]_{(i)}{}^T [\mathbf{D}]_{(i)} \right]^{-1} \right|}{\left| \sigma_t^2 \left[[\mathbf{D}]^T [\mathbf{D}] \right]^{-1} \right|} = \left[\left(\frac{n-p-1+t_i^2}{n-p} \right)^p (1-h_{ii}) \right]^{-1} \tag{11}$$

The determinant of the variance-covariance matrix is a generalised measure of variance. Thus C reflects the impact of the ith observation on the precision of the estimates of the regression coefficients. Values near 1 indicate that the ith observation has little effect, greater than 1 indicates that the presence of the ith observation increases the precision of the estimates, the converse is true. A range of $1 \pm 3p/n$ is suggested to be considered the extremes for identifying influential points. Thus in this section we have presented a set of computationally simple methods for determining outliers to a set of linear equations.

2.2 Affine Flow

In this paper we assume that the spatial structure of the projected flow, with the exception of independently moving foreground objects, is coherent and may be approximated by a linear vector field. According to the proposed method these foreground objects may be detected as inconsistent with the affine background motion. The affine assumption is approximately correct when the distance to background objects is large when compared to the variations in these distances, this occurs in many outdoor scenes. It is also approximately correct for rotations seen over a small field of view.

Rather than first computing the flow, then fitting, we fit directly to the spatiotemporal image surface. Let the image intensity at inhomogeneous pixel coordinate $\vec{\mathbf{x}} = (x, y, -f)$ be $E(x, y)$. The motion constraint equation [9] is:

$$\frac{\partial E}{\partial x}\frac{dx}{dt} + \frac{\partial E}{\partial y}\frac{dy}{dt} + \frac{\partial E}{\partial t} = 0 \tag{12}$$

Verri and Poggio [14] have shown that equation 12 does not hold in general, but increases in accuracy as the spatial gradient increases. Thus in general points are only used if $|\nabla E|$ exceeds a certain threshold. Let $\vec{\mathbf{u}} = (u, v)$ be the projected velocity at $\vec{\mathbf{x}}$ then $\nabla E \cdot \vec{\mathbf{u}} + E_t = 0$ given the flow varies linearly:

$$\begin{pmatrix} E_x & E_y \end{pmatrix} \begin{bmatrix} u & \frac{\partial u}{\partial x}\Delta x & \frac{\partial u}{\partial y}\Delta y \\ v & \frac{\partial v}{\partial x}\Delta x & \frac{\partial v}{\partial y}\Delta y \end{bmatrix} = -E_t \tag{13}$$

We can rewrite equation 13 separating the observables and unobservables into two vectors of the form given by equation 1 where

$$\begin{aligned} \vec{\mathbf{d}} &= \begin{pmatrix} E_x & E_y & \Delta x E_x & \Delta y E_x & \Delta x E_y & \Delta y E_y \end{pmatrix}^T \\ \vec{\mathbf{b}} &= \begin{pmatrix} u & v & \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} & \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}^T \\ \vec{\mathbf{y}} &= \begin{pmatrix} E_{t_1} & E_{t_2} & \dots & E_{t_N} \end{pmatrix}^T \end{aligned} \tag{14}$$

We may solve the set of equations presented in 14 by a least squares method, given N image points for which we know the spatio-temporal derivatives. Outliers to this system of equations are then be deemed to be independently moving objects.

2.3 Collinearity

If there are near-singularities among the columns of [**D**] then we have insufficient data within this region to reconstruct the fit. This could have arisen from a highly structured image e.g. a series of vertical bars. Alternatively sparse or unbalanced data could give rise to collinearity. When the data is collinear then we must rely on past estimates for information. Waxman [16] referred to this as the *aperture problem in the large* in which insufficient contour structure leaves the set of deformation parameters undetermined, even over large regions of the image. For instance, if the data had arisen from a single conic section then there would be at least one affine dependency in [**D**] and the Taylor coefficients β would be undetermined.

Geometrically, this means there is poor dispersion in one of the dimensions of [**D**]-space. The presence of collinearity can be detected by an eigen analysis of $[\mathbf{D}]^T[\mathbf{D}]$. The six eigenvalues λ_i^2 provide measures of the amount of dispersion for each of the principal component axes in [**D**]-space [11]. The condition number is defined as the ratio of the largest to the smallest singular value λ_i. This gives a measure of sensitivity of β to small changes in [**D**]. The condition number concept is extended to the condition index for each (principal component) dimension of the [**D**]-space. The condition index K_i for the ith principal component dimension in [**D**]-space is the ratio of the maximum singular value to the ith singular value. We shall take values of the condition index around 10 to indicate moderate dependencies, values from $30 - 100$ to indicate strong dependencies and values in excess of 100 to indicate severe collinearity problems [11]. The number of condition numbers in the critical range indicating the number of near-dependencies. Given a large number of dependencies warning must be given that the result of the regression is suspect.

The size of the eigenvalues depends on the scale of the columns of [**D**], thus we shall scale [**D**] so that the length of each column vector is one (i.e the sum of squares of the elements is unity) to prevent the eigen-analysis being dominated by one or two independent variables e.g the term xE_x will be always be larger than the E_x column but we do not wish to give it any greater weight when testing for collinearity.

Thus we have presented a statistically well founded technique for determining whether the image is indeed sufficiently structured enough to allow us to recover β.

3 Clustering

Once we have identified a set of outliers we then need to form a hypothesis about whether they are consistent with one or more rigid three dimensional objects moving independently of the background. We utilise a method of spatial clustering by merging nearby outliers into groups and defining the region of interest as the convex hull of the group, following an algorithm presented in [8].

A problem with this is that a lone outlier (a result of noise for instance) might seriously distort the convex hull. Thus we utilise further robust statistical methods to differentiate between outliers caused by noise and outliers caused by objects moving differently to the background. The image is tessellated into equal sized overlapping regions. Given an estimate of the noise (by observation of some static scenes) we calculate a 99% confidence interval that the number of outliers within the region must exceed to determine that the outliers within that region are not due to noise. From the set of points delineating the convex hull we initialise a b-spline snake onto the occluding contour of the object.

4 Results

We have implemented a movement detector based on the above principals using the COVRATIO test–C and Cook's D measures for outliers on a linear fit. Empirically it was found that the results of DFFITS was indistinguishable from

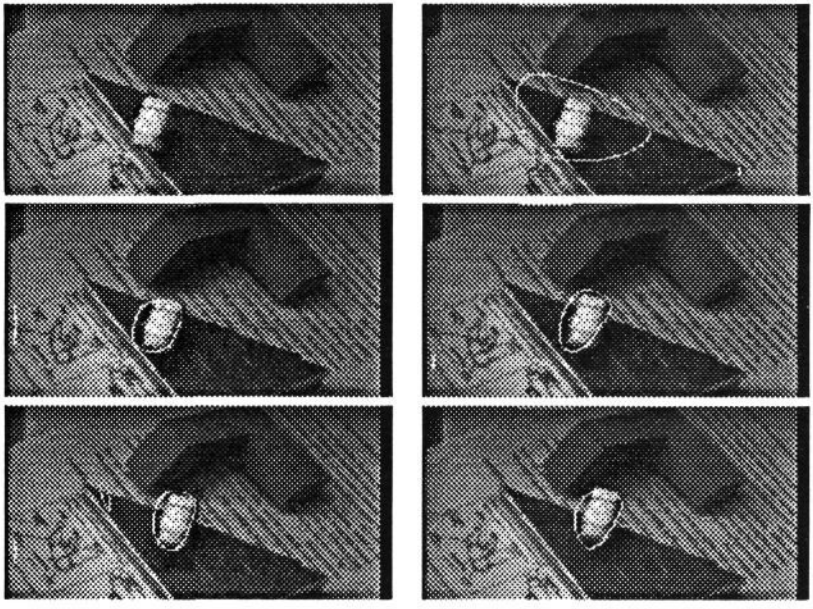

Figure 1: *Showing a sequence of images of a white object translating* 4 *pixels to the right as the background translates* $0-2$ *pixels down and* $0-1$ *pixels right depending on the depth.*

Cook's D. The first is of a moving object taken from a camera in motion in an indoor scene. The second is of several moving objects as the camera translates. No heuristics or "magic numbers" were used to derive the thresholds. Instead they were derived from the statistical theory underpinning the work. Furthermore, these thresholds are self scaling to the type of image concerned.

The first sequence of images shown in figure 1 is of a white object translating 4 pixels to the right as the background translates 2 pixels down and $0-1$ pixels right depending on the depth. The image is 256×128. To reduce the amount of redundant information in the least squares points with low gradient were excluded, this reduced the number of points under consideration from 32768 to 29039. The reason for the exclusion of these points is that they are clustered about the origin in observation space through which any fit must pass and thus provide redundant information, their exclusion reduces the amount of calculation. The thresholds for points to be considered outlying were $C = 1-\frac{3p}{n}$ and $D = \frac{4}{n}$ taken from [11], e.g. $C < 0.997864$, $D > 0.000285$. Figure 2 shows the result of the variance test superimposed on the motion shown in in figure1, areas in black are outliers, white areas are background and the grey areas where points excluded from the fit and are shown in their original intensities. Note that some of the edge of the lower triangle has been indicated as outlying. Care must be taken when handling the output of the outlier tests. As we are making a linear flow assumption depth and velocity discontinuities are both shown. It is hoped that the inclusion of a matching strategy over time might reduce the number of false outliers. The second pair of images, in figure 3, depicts several

Figure 2: *Showing the result of the C-test for outliers superimposed on the images shown in figure 1, white areas are the background and black areas are outliers. Grey areas are the intensities of the original regions with low* (E_x, E_y, E_t) *that are excluded from the regression.*

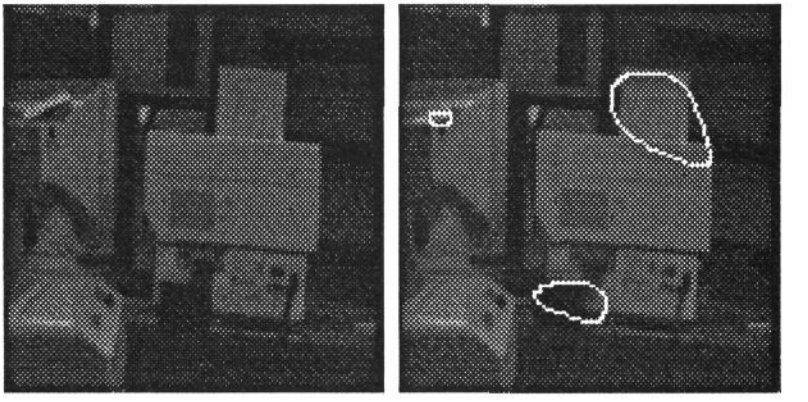

Figure 3: *Showing several lab objects moving in different directions as the background moves* $0-3$ *pixels left. In the bottom centre of the image a black box moves quickly* 10 *pixels to the left and* 3 *pixels up, on the top left a piece of paper contracts as it slips down the back of the monitor* $3-5$ *pixels and moves* 3 *pixels to the left (i.e* 0 *left pixels relative to the background). In the centre a large white box moves to the right by* 6 *pixels and down by* 1 *pixel*

lab objects moving in different directions as the background moves $0-3$ pixels left. In the bottom centre of the image a black box moves quickly 10 pixels to the left and 3 pixels up, on the top left a piece of paper contracts as it slips down the back of the monitor $3-5$ pixels and moves 3 pixels to the left (i.e 0 left pixels relative to the background). In the centre a large white box moves to the right by 6 pixels and down by 1 pixel. The image is 256×256, points with low gradient were excluded, this reduced the number of points under consideration from 65536 to 21216. The thresholds for points to be considered outlying were $C < 0.998586$, $D > 0.000189$

Overall given accurate estimates of the first order properties of the image the algorithm successfully localises the independently moving objects, providing that there motion is sufficiently different to the background motion.

5 Conclusions

In this paper we have presented a method for detection of non-rigid motion given information derived from time varying imagery. The method is founded upon the examination of the differences between the observed temporal difference and a predicted form given an affine transformation. Thus we make a global estimate of the background motion using a scene constraint, as a heuristic. The algorithm does not attempt to establish point correspondences, estimate the optic flow, or make a three dimensional reconstruction. It does not require knowledge of camera motion or calibration. Instead successive quantitative estimates are made about the deformation parameters of background motion and parts of the image that do not accord with this estimate are identified as regions of interest. This identification is done using recent statistical work on the analysis of regressions. The thresholds are determined in a principled manner and are self scaling to the variances of the image intensities.

6 Future Work

There are two avenues of current research. The first is to improve the method of grouping the outliers into cohesive groups. The second is generalisation to a more realistic set of motions. An inherent problem with an affine approximation is that it is only valid in a limited number of situations. Current work addresses the problem of how to detect a set of moving objects in the two dimensional projection of an otherwise rigid scene, given that the camera is moving in an arbitrary and unpredetermined manner. We utilise the fact that point correspondences having arisen from a projective $3D$ transformation can be described by a 3×3 *Essential Matrix* [**E**] linking the coordinates of the points before and after the transformation. The Essential Matrix is derived by an analytic $O(N^3)$ least squares method, assuming at least half of the image is undergoing a coherent projective transformation. Thus points with non-rigid motion (modulo a projectivity) are deemed to be those statically inconsistent from the calculated value of [**E**].

Acknowledgements

This work was supported by SERC grant GR/G30003. Thanks are due to Andrew Zisserman and Paul Bearsdley for helpful suggestions and to Charlie Rothwell for the convex hull software.

References

[1] G. Adiv. Inherent ambiguities in recovering 3-d motion and structure from a noisy flow field. In *Proceedings, CVPR '85 (IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, June 10–13, 1985)*, IEEE Publ. 85CH2145-1., pages 70–77. IEEE, IEEE, 1985.

[2] E. Belsley, D.A. Kuh and Welsch R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley, 1980.

[3] P. J. Burt. Image motion analysis made simple and fast, one component at a time. In *Proc. BMVC*, pages 1–8, 1991.

[4] M. Campani and A. Verri. Computing optical flow from an overconstrained system of linear algebraic equations. In *Proceedings of the Third International Conference on Computer Vision*, pages 22–25, 1990.

[5] R.D. Cook and S. Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22:337–344, 1980.

[6] R.D. Cook and S. Weisberg. *Residuals and Influence in Regression.* Chapman Hall; London, 1982.

[7] E. François and Bouthemy P. Multiframe based identification of mobile components of a scene with a moving camera. In *Proc. CVPR.*, 1991.

[8] Green and Silverman. Constructing the convex hull of a set of points in the plane. *Computer Journal*, vol.22:262–266, 1979.

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

[10] R.C. Nelson. Qualitative detection of motion by a moving observer. *IJCV*, pages 33–46, 1991.

[11] J.O. Rawlings. *Applied Regression Analysis.* Wadsworth and Brooks, California, 1988.

[12] B.G. Schunck. Image flow segmentation and estimation by constraint line clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:1010–1027, 1989.

[13] W.B. Thompson, K.M. Mutch, and V.A. Berzins. Dynamic occlusion analysis in optical flow fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7:374–383, 1985.

[14] A. Verri and T. Poggio. Against quantitative optical flow. In *First International Conference on Computer Vision, (London, England, June 8–11, 1987)*, pages 171–180, Washington, DC., 1987. IEEE Computer Society Press.

[15] A.M. Waxman and J.H. Duncan. Binocular image flows: Steps toward stereo-motion fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:715–729, 1986.

[16] A.M. Waxman and K. Wohn. Contour evolution, neighborhood deformation, and global image flow: Planar surfaces in motion. *International Journal of Robotics Research*, 4(3):95–108, 1985.