Object detection using model based prediction and motion parallax

Stefan Carlsson
Telecommunication Theory
and
Jan-Olof Eklundh
Dep. of Numerical analysis and Computing Science
Royal Institute of Technology , S-100 44 Stockholm, Sweden

1. Motion parallax and object background separation

When a visual observer moves forward the projections of the objects in the scene will move over the visual image. If an object extends vertically from the ground its image will move differently from the immediate background. This difference is called motion parallax [1,2]. Much work in automatic visual navigation and obstacle detection has been concerned with computing motion fields, or more or less complete 3-D information about the scene [3-5]. These approaches in general assume very unconstrained environments and motion. If the environment is constrained, e.g. motion occurs on a planar road, then this information can be exploited to give more direct solutions to e.g. obstacle detection.[6]

Fig. 1.1 shows superposed the images from two successive times for an observer translating relative a planar road. The arrows show the displacement field, i.e. the transformation of the image points between the successive time points.

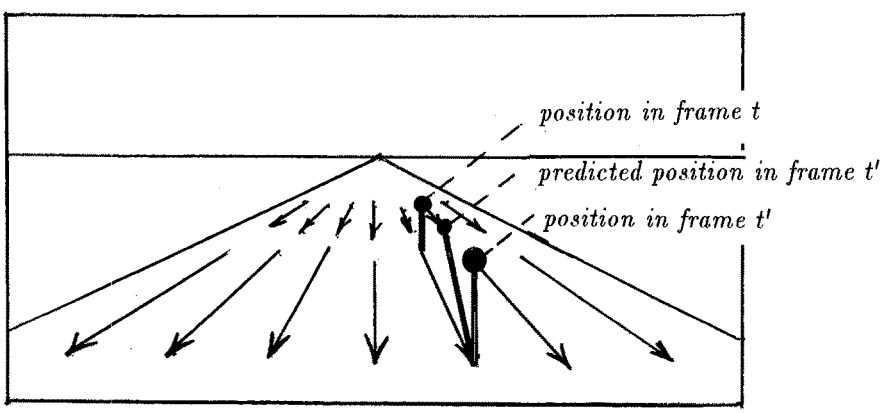

Fig. 1.1 Displacement field from road with predicted and actual position of vertically extended object

Fig. 1.1 also shows a vertically extended object at time t and t'. Note that the top of the object is displaced quite differently from the immediate road background. This effect is illustrated by using the displacement field of the road to displace the object. A clear difference between the actual image and the predicted image is observable for the object. This fact forms the basis of our approach to object detection. (Fig. 1.2) For a camera moving relative a planar surface the image transformation of the surface is computed and used to predict the whole image. All points in the image that are not on the planar surface will then be erroneously predicted. If there is intensity contrast at those parts we will get an error in the predicted image intensity. This error then indicates locations of vertically extended objects.

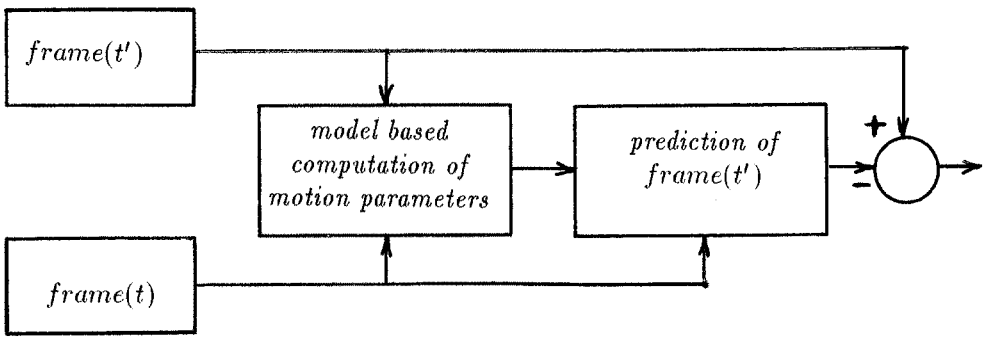

Fig. 1.2 Block diagram of processing for vertical object detection

2. Image transformation for motion relative a planar surface

With a moving camera each point in the scene will map to a different point in the image at different times. The transformation of the mapped image point over time is determined by the motion of the camera and the position of the point in the 3-dimensional scene. If the point is on a planar surface the transformation can be computed using the camera motion and position of the surface in space. Fig. 2-1 shows the coordinate system of the camera and the image plane. A rigid displacement of the camera can be decomposed into a translation with components D_X, D_Y, D_Z along the coordinates and a rotation around an axis passing through the point of projection, which can be decomposed into rotations around the axis of the coordinate system ϕ_X, ϕ_Y, ϕ_Z. Assuming small rotations, a point in the scene with coordinates X, Y, Z is then transformed to the point X', Y', Z' where:

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \begin{pmatrix} 1 & -\phi_Z & \phi_Y \\ \phi_Z & 1 & -\phi_X \\ -\phi_Y & \phi_X & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} D_X \\ D_Y \\ D_Z \end{pmatrix} \qquad [2.1]$$

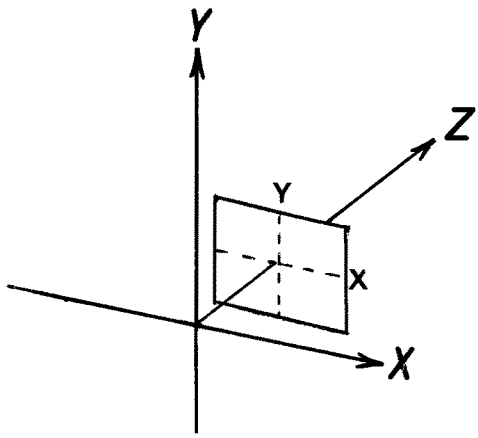

Fig. 2.1 Coordinate system of camera and image plane

If the image plane is located at unit distance from the point of projection the image coordinates (x, y) of a point X, Y, Z under perspective projection are

$$x = \frac{X}{Z} \qquad y = \frac{Y}{Z} \tag{2.2}$$

The transformation of the projected image point of a point in the scene with depth Z is therefore [7]:

$$x' = \frac{x - \phi_Z y + \phi_Y + D_X/Z}{1 - \phi_Y x + \phi_X y + D_Z/Z} \qquad y' = \frac{y + \phi_Z x - \phi_X + D_Y/Z}{1 - \phi_Y x + \phi_X y + D_Z/Z} \tag{2.3}$$

If the point X, Y, Z is located on a planar surface with equation $K_X X + K_Y Y + K_Z Z = 1$, the transformation in the image plane the becomes:

$$x' = \frac{(1 + K_X D_X)x + (K_Y D_X - \phi_Z)y + \phi_Y + D_X K_Z}{(1 + K_X D_Z - \phi_Y)x + (K_Y D_Z + \phi_X)y + K_Z D_Z} \tag{2.4 - a}$$

$$y' = \frac{(1 + K_Y D_Y)y + (K_X D_Y - \phi_Z)x - \phi_X + D_Y K_Z}{(1 + K_X D_Z - \phi_Y)x + (K_Y D_Z + \phi_X)y + K_Z D_Z} \tag{2.4 - b}$$

This is a nonlinear transformation of the image coordinates, determined by 9 parameters. The actual number of degrees of freedom of the transformation is however just 8, since parameters $\bar{K}$ and $\bar{D}$ always occur as products. which means that their absolute values are irrelvant.

3. Estimation of parameters by minimisation of prediction error

The transformation of the projected image points due to the motion of the camera will manifest itself as a transformation of the image intensity $I(x, y)$. If t and t' are the time instants before and after the transformation, we shall assume that:

$$I(x',y',t') = I(x,y,t) \qquad [3.1]$$

where x, y and x', y' are related according to eq. 2.4.

I.e. we assume that the transformation of the image intensity is completely determined by the geometric transformation of the image points. This is not strictly true in general since we neglect factors as changing illumination etc.

For points on a planar surface our assumption implies that the transformation of the intensity is determined by the 3 vectors $\bar{\phi}, \bar{D}$ and $\bar{K}$ charcterising the camera motion and surface orientation respectively. The determination of these parameters can therefore be formulated as the problem of minimising the prediction error:

$$P(\bar{\phi},\bar{D},\bar{K}) = \sum_{x,y}[\ I(x'(x,y,\bar{\phi},\bar{D},\bar{K}),y'(x,y,\bar{\phi},\bar{D},\bar{K}),t') - I(x,y,t)\ ]^2 \qquad [3.2]$$

where the summation is over image coordinates containing the planar surface.

For the minimisation we use gradient descent, i.e. the values of the parameters are adjusted iteratively according to:

$$\bar{\phi}^{(i+1)} = \bar{\phi}^{(i)} - \mu_1 \frac{\partial P^{(i+1)}}{\partial \bar{\phi}}$$

$$\bar{D}^{(i+1)} = \bar{D}^{(i)} - \mu_2 \frac{\partial P^{(i+1)}}{\partial \bar{D}} \qquad [3.3]$$

$$\bar{K}^{(i+1)} = \bar{K}^{(i)} - \mu_3 \frac{\partial P^{(i+1)}}{\partial \bar{K}}$$

where i denotes iteration index. The derivatives with respect to the parameter vectors are taken componentwise.

The convergence properties of the gradient descent minimisation depend heavily on the structure of the image intensity function $I(x,y)$. Assume that we have computed approximate values $\hat{\phi}, \hat{D}, \hat{K}$ for the parameters. These values transform the point (x,y) to the point $(\hat{x'},\hat{y'}) = (x' + \delta x', y' + \delta y')$ If the approximate parameter values are close enough to the true values, $\delta x'$ and $\delta y'$ will be small. For the prediction error we then have:

$$e \ = \ I(\hat{x'},\hat{y'},t') - I(x,y,t) \ = \ I(x' + \delta x', y' + \delta y', t') - I(x,y,t) \ \approx$$

$$\approx \frac{\partial I}{\partial x'}\delta x' + \frac{\partial I}{\partial y'}\delta y' \qquad [3.5]$$

In order to compute well defined values of motion and surface orientation parameters the prediction error e should be sensitive to small changes in these. From eq.3.5 we

see that the size of the image intensity gradient is very important in this respect since it directly amplifies any variations in $\delta x', \delta y'$. Preferably image points x, y with high intensity gradient should be used in the computation of the prediction error in eq. . The selective use of points with high gradient also reduces the volume of the computations involved. For our application we therefore first applied a version of the Canny-Deriche edge detector to the image [8,9], and used only the edge points. The choice of edge detector for this problem is probably not critical. What is needed is just a selection of points with high intensity gradient.

The derivatives of the prediction error P with respect to the parameters were computed by systematically varying the parameters. If the difference between the transformed coordinates x', y' and the original x, y is small, e.g. by choosing the time interval $t' - t$ to be small, these derivatives could be computed more efficiently using the relation 3.5. In principle direct methods of determination of the parameters can be used. [10].

4. Sequential estimation using recorded sequence

Since the algorithm assumes motion relative a planar surface we first have to select points in the image, projected from the road , to be used for parameter estimation. Under normal driving conditions, the part of the image immediately in front of the car can be assumed to project from the planar road surface. The position of the car realtive to the road boundaries can also be considered as relatively stable. The points to be used in the algorithm are therefore selected from a rectangular window in the image choosen so that points from the road immediately in front of the car are contained in the window as shown in fig. 4-1. This window is fixed in time relative to the image. As obstacles are detected the window could be made adaptive so that these objects are not included in the pixels used for parameter estimation.

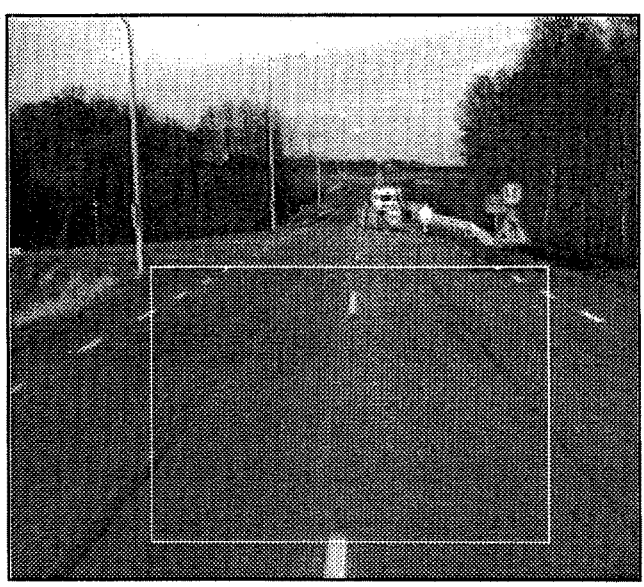

Fig. 4.1 Window for selection of points on planar road surface

The gradient descent algorithm for estimation of parameters can now be applied to succesive image frames in the sequence. If the time between the succesive frames is short enough, we can expect a high correlation in time between computed parameter values. This can be exploited in the gradient descent algorithm by using parameter values computed in the previous frame pair as start values for the next pair. For an ideal planar road surface there will in fact be a coupling between motion parameters, $\bar{\phi}, \bar{D}$ and surface orientation $\bar{K}$, since the change in surface orientation over time is determined by the motion. This can be introduced as an extra constraint in the algorithm in order to build a true spatio-temporal model of the position and orientation of the camera relative to the road. At this stage however we did not consider this coupling between parameters.

5. Object detection using prediction error

If the estimated parameters $\hat{\phi}, \hat{D}$ and $\hat{K}$ are correct and the road conforms to the planar surface model, the error in the prediction of the image intensity acc. to eq. 3.5 will be 0. Any errors in the estimated parameter values or errors in the model will however give rise to a non zero prediction error. Any vertically extended object in the scene will obviously violate the planar surface model and therebye cause a prediction error. The prediction error is thus an important variable to be used for deciding whether any objects are present in front of the car.

The effect of a vertically extended object on the prediction error is however highly dependent on its position in the scene. If we consider the ideal case of a camera aligned with optical axis parallell to the planar surface, translating in the direction of the optical axis only with no rotation, we have for a point in the scene at depth Z the following transformation in the image:

$$x' = \frac{x}{1 + D_Z/Z} \qquad y' = \frac{y}{1 + D_Z/Z} \qquad [5.1]$$

If we choose units so that the height of the camera over the ground, $-1/K_Y = 1$ we have for points on the planar road surface:

$$x^m = \frac{x}{1 - D_Z y} \qquad y^m = \frac{y}{1 - D_Z y} \qquad [5.2]$$

A point in the image with coordinates x, y at height h above the road will be at depth $Z = -1 + h/y$. For this point the difference between the actual coordinates and thoose predicted by the model is:

$$x' - x^m = \frac{hD_Z xy}{(1 - D_Z y - h)(1 - D_Z y)} \qquad y' - y^m = \frac{hD_Z y^2}{(1 - D_Z y - h)(1 - D_Z y)} \qquad [5.3]$$

Note that this applies only to points x, y in the image that project to the road surface. This means that they are below the horison, i.e. $y < 0$ and $h < 1 - D_Z y$.

We see that the error in the prediction of the coordinates grows monotonically with height h above the ground. However it also grows with distance from focus of expansion $x = 0, y = 0$. Objects close to the F.O.E. will therefore give rise to very small errors in the coordinates. This is natural since motion at the F.O.E. is zero, independent of

the objects coordinates in space. If we use scene coordinates X and Z instead of image coordinates we get for the coordinate error:

$$x' - x^m = \frac{hD_Z X}{(Z + D_Z)(Z + D_Z - D_Z h)} \qquad y' - y^m = \frac{h(h-1)D_Z}{(Z + D_Z)(Z + D_Z - D_Z h)} \quad [5.4]$$

From this we see that the error scales inversely with depth, i.e. distance in front of the camera. It also grows with the distance X to the side of the optical axis, and with D_Z the translatory displacement along the optical axis.

For small errors in the coordinates we can estimate the error in the predicted image intensity by projecting the coordinate errors on the intensity gradient acc. to eq. 3.5 . This means that errors in the predicted image intensity will only show up at the edges of the vertically extended objects, unlesss the errror in the predicted coordinates are large enough. Important to note is also the fact that the orientation of the edges of the vertically extended objects influence the size of the error in the predicted image intensity. From eq.5.3 we see that $\delta x', \delta y'$ will be oriented radially out from the focus of expansion. For maximum prediction error acc to eq. 3.5 the gradient shouls be parallell to this orientation, i.e. the edges should be orhogonal to the lines radiating out from the focus of expansion. However we must emphasise that this only applies when coordinate errors $\delta x', \delta y'$ are small

Another cause of prediction error is errors in the estimated parameters. For these errors we also have the dependence on the distance from the F.O.E. This is important to consider in the evaluation of false alarm detections e.g.

6. Experimental results and conclusions

The algorithm for sequential model based motion estimation and object detection was simulated using a digitised video tape recorded from a camera placed on top of a moving car. 100 frames with a frame rate of 25/sec was selected from the video sequence. In this sequence the car approaches a roadwork where the right lane of the road is blocked by warning signs and fences. In the first frame the car is about 150 m from the roadwork.

In order to get sufficiently large prediction errors for objects close to the F.O.E. the prediction was made 3 frames ahead. Parameters were computed for every frame however, i.e. the sequence of frame pairs used were 1-4, 2-5, 3-6, etc. A special problem was the initiation of the algorithm. Since only the pixels at the edges from the Canny-Deriche edge detector were used in the prediction error computation the convergence of the algorithm depended on the initial parameters not beeing too far away from the correct values. For the first frame pair therefore, all the pixels in the window were used for computation of the prediction error.

For each frame pair the gradient descent algorithm was iterated 30 times. No complexity considerations were considered in choosing this number. For more iterations the improvement of the prediction was found to be negligible.

Fig.6.1.a-b shows the unpredicted difference and the prediction error respectively for frames 97-100. We see that a clear reduction of the difference image is obtained in

the prediction error image for image points projecting from the ground. For vertically extended objects the reduction is significantly less, depending on verical extent and distance from F.O.E. The diagrams in fig 6.1.c-f illustrates this in more detail. The curves show the image intensity from the difference and prediction error image respectively. The intensity along two different horisontal lines, indicated in fig.6.1.a-b are plotted. The first line containing white marking from the road is clearly reduced in the prediction error image, while the second containing the vertically extended objects in the roadwork shows comparatively less reduction from difference to prediction error image.

In fig.6.2 is shown with white markings thresholded prediction error images from several different times. The images were thresholded at the edge points only and the threshold was increased systematically with distance from F.O.E. The threshold was choosen so that at most one marking was obtained from the part of the road containing the parameter estimation window. This means that the threshold was adapted to any errors in the estimated parameters.

From fig.6.2 we see that as expected the sensitivity of the algorithm increases with distance from the F.O.E and height over ground. A very important factor is also the intensity contrast of the objects relative background. Some vertically extended objects are not over the threshold in every frame. By combining detections from several frames the performance should be increased however. In order to do this the detections from the different objects have to be grouped and anlysed separately. The main purpose of this prediction error based detection can therefore be seen as a preprocessing mechanism which directs resources of the system for further processing.

Acknowledgements

The video recording of the road sequence was made by Lars J. Olsson, Mercel AB and digitised data was provided by the Computer Vision Laboratory, Linkoping University. The work was performed under contract within the Prometheus-Sweden project. It was also supported by STU, the Swedish national board for technical developement under contract 88-01749.

References

[1] H. von Helmholtz, Physiological optics, vol.3, 1866

[2] J.J. Gibson, The perception of the visual world, *Houghton, Muffin, Boston 1950*

[3] H.C. Lounguet-Higgins and K.Prazdny, "The interpretation of a moving retinal image ", *Proc. Roy. Soc. London B-208, pp. 385 - 397. 1980*

[4] R.C. Nelson and J. Aloimonos, "Using flow field divergence for obstacle avoidance: Towards qualitative vision ", *Proc. of 2:nd Int. Conf. on ComputerVision, Tampa Florida, pp.188 - 196 1988*

[5] W.B.Thompsom, K.M.Much and V.A.Berzins, "Dynamic occlusion analysis in optical flow fields", IEEE Trans. Pattern Anal. Machine Intelligence, Vol. PAMI-7, No.4, July 1985

[6] H.A. Mallot, E Schulze and K. Storjohann, "Neural network strategies for robot navigation " *Proc.of nEuro'88, Paris, June 6-9, pp. 560 - 569, 1988*

[7] R.Y. Tsai and T.S. Huang, "Estimating three-dimensional motion parameters of a rigid planar patch ", IEEE-ASSP, Vol ASSP-29, No. 6, December 1981.

[8] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Machine Intelligence, Vol. PAMI-8, 679-698, 1986,

[9] R. Deriche, Using Canny's criteria to derive a recursively implemented optimal edge detector,Int. J. of Comp. Vision , 1, 167-187, 1987.

[10] B.K.P. Horn and E.J. Weldon, " Direct methods for recovering motion ", Int. J. of Comp. Vision 2, 51-76, 1988.

Fig. 6.1.a Difference frames 97-100

Fig. 6.1.b Prediction error frames 97-100

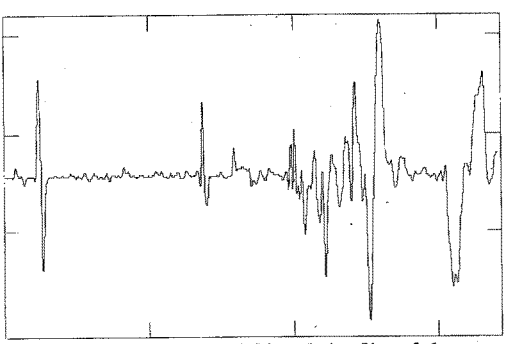

Fig. 6.1.c Intensity of line 1 in fig. 6.1.a

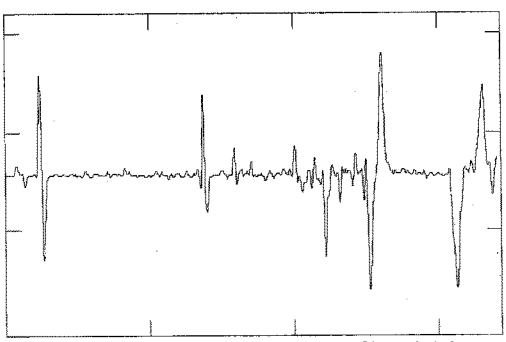

Fig. 6.1.d Intensity of line 1 in fig. 6.1.b

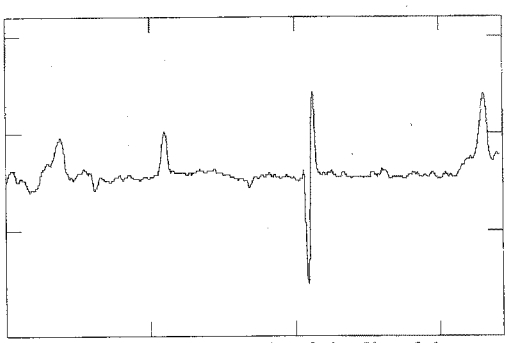

Fig. 6.1.e Intensity of line 2 in fig. 6.1.a

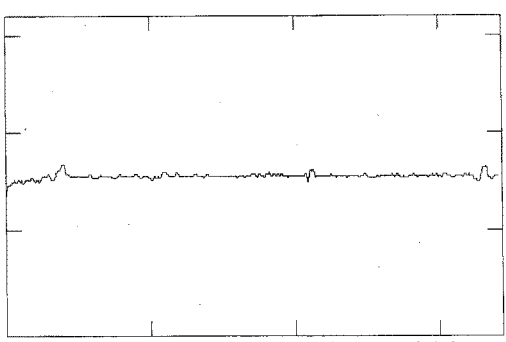

Fig. 6.1.f Intensity of line 2 in fig. 6.1.b

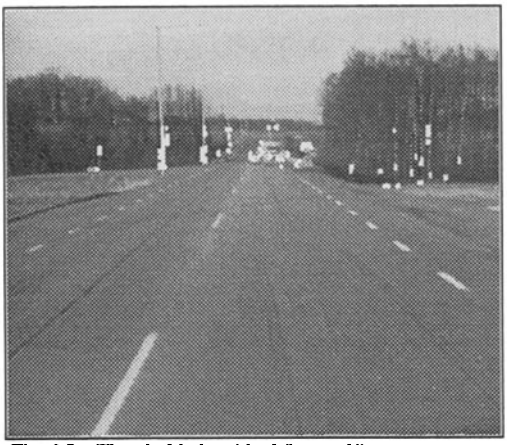

Fig. 6.2.a Thresholded residual frame 10

Fig. 6.2.b Thresholded residual frame 30

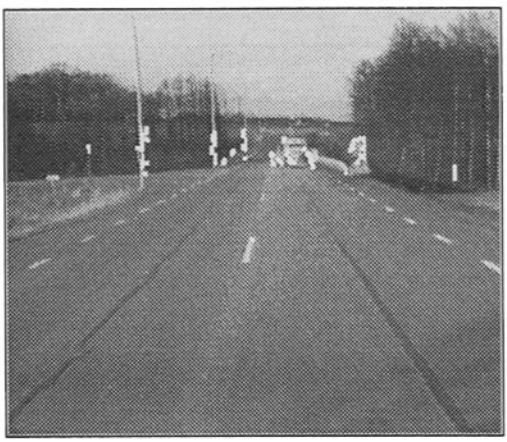

Fig. 6.2.c Thresholded residual frame 50

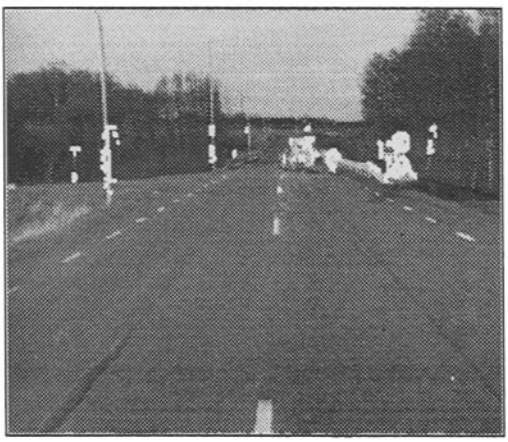

Fig. 6.2.d Thresholded residual frame 70

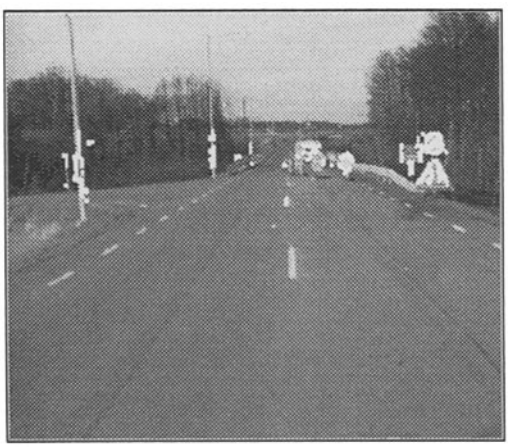

Fig. 6.2.e Thresholded residual frame 80

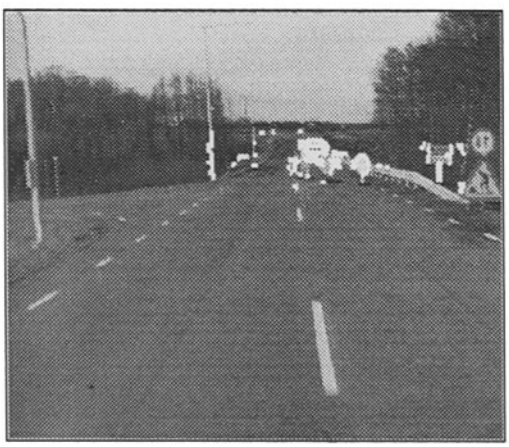

Fig. 6.2.f Thresholded residual frame 90