Fast Obstacle Detection using Flow/Depth Constraint

S. Heinrich

DaimlerChrylser AG
P.O.Box 2360, D-89013 Ulm, Germany
Stefan.Heinrich@DaimlerChrysler.com

Abstract

The early recognition of potentially harmful traffic situations is an important goal of vision based driver assistance systems. Pedestrians, in particular children, are highly endangered in inner city traffic. Within the DaimlerChrysler UTA (Urban Traffic Assistance) project, we are using stereo vision and motion analysis in order to manage those situations. The flow/depth constraint combines both methods in an elegant way and leads to a robust and powerful detection scheme.

1 Introduction

Within the DaimlerChrysler UTA (Urban Traffic Assistance) project, different vision modules for inner city traffic have been developed [1,2]. This includes fast stereo vision for Stop&Go, traffic sign and light recognition as well as pedestrian recognition and tracking. It is the goal of our current investigations to add collision avoidance capabilities to the existing system. In particular, we intend to recognize situations that implicate a high risk of accidents with children running across the road. A city roller coming from the side, a child looming between parking cars as shown in Fig. 1.1 indicate such dangerous situations. A warning as well as an emergency reaction have to take place instantaneously in order to prevent accidents with serious injuries.

Fig. 1.1 A child behind a car

Relevant objects must be detected and classified in real-time from the moving car. For obstacle recognition, we generally use stereo analysis followed by a classification stage.

Stereo vision delivers three-dimensional measurements. A height threshold is applied in order to distinguish between ground and obstacle features. Points above ground are grouped to objects. Detected objects are tracked over time to estimate their motion.

Although very powerful, stereo analysis has three drawbacks with respect to the application that we have in mind. First, the grouping process tends to merge objects which are close to each other, e.g. a pedestrian in front of a vehicle or a child behind a car. Secondly, the height threshold implies the risk to miss small obstacles which are close to the ground. Thirdly, motion information included in the sequence is exploited for the detected objects only.

Motion analysis, on the other hand, allows to estimate the motion of any pixel based on the analysis over time and thus detection of any moving object.

In vehicles, a precise recovery of the ego-motion is necessary in order to distinguish between static and moving objects. Unfortunately, the ego-motion estimation is a difficult problem which requires considerable computational power and usually lacks from robustness. Neither the presence of optical flow automatically indicates a moving object, nor a flow equals zero does mean a zero risk. Depending on depth, a collision could take place in any case.

A proper combination of both techniques promises the optimal exploitation of the available information in space and time. In this paper, we present an elegant method which uses the fact that stereo disparity and optical flow are connected via real-world depth. The so called "flow/depth constraint" allows to test each motion vector directly against the stereo disparity to detect moving objects. The detection works within a few image frames with very low computational cost.

In chapter 2. we describe the used systems for stereo and motion analysis. The fusion of stereo and motion data by means of the flow/depth constraint is presented in chapter 3.

2 Stereo and Motion

2.1 Stereo Vision

Our stereo analysis [3] is based on a correlation-based approach. In order to reach real-time performance on a standard PC, design decisions need to be drawn carefully.

First of all, we use the sum-of-squared (SSD) or sum-of-absolute (SAD) differences criterion instead of expensive cross correlation to find the optimal fit along the epipolar line. Wrong results due to different mean and variance of the image pairs can be avoided if gain and shutter of the cameras are appropriately controlled.

Secondly, in order to speed up the computation, we use a multi-resolution approach in combination with an interest operator. The idea is to find correspondences on a coarse level that can be recursively refined. First, a gaussian pyramid is constructed for the left and right stereo image. Areas with sufficient contrast are extracted by means of a fast vertical Prewitt edge detector.

Pixels with sufficient gradient are marked, from which a binary pyramid is constructed. A pixel (i,j) at level n is marked if one of its 4 corresponding pixels at level n-1 is set. A non-maximum suppression is applied to the gradient image in order to further speed up the processing. In this case, we find about 1100 attractive points at pyramid level zero (original image level), 700 at level one and 400 at level 2 on typical image sequences. Only those correlation windows with the central pixel marked in these interest images are considered during the disparity estimation procedure.

Depending on the application, the correlation process starts at level one or two of the pyramid. If D is the maximum searched disparity at level zero, it reduces to $(D/2)^n$ at level n. At level 2 this corresponds to a saving of computational burden of about 90% compared to a direct computation at level zero. Furthermore, smaller correlation windows can be used at higher levels which again accelerates the computation.

The result of this correlation is then transferred to the next lower level. Here, only a fine adjustment

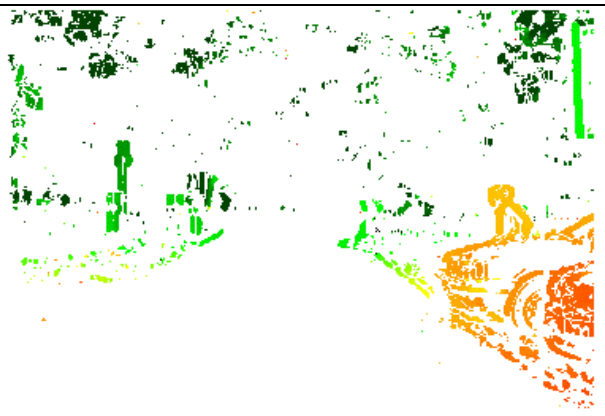

Fig. 2.1: Color encoded disparity image generated by the correlation approach. Red signals close, green means far.

has to be performed within a small horizontal search area of +/- 1 pixel. This process is repeated until the final level is reached. At this level, subpixel accuracy is achieved by fitting a parabolic curve through the computed correlation coefficients.

The price we have to pay for this fast algorithm is that mismatches in the first computed level propagate down through the pyramid and lead to serious errors. Since the quality of a found match cannot be judged by the measured SSD or SAD, we compute the normalized cross correlation coefficient for the best matches at the highest correlation level and eliminate bad matches from further investigations. In addition, a left-right check can be applied to the disparity images on the different pyramid levels. In case of ambiguities, the best match or the match with the smaller disparity is selected. The latter strategy avoids the erroneous detection of close obstacles caused by periodic structures.

Usually, we start at level 2 (resolution 91x64 pixels) and allow a maximum disparity of 60 pixels corresponding to a minimum distance of 4 meters. In this case, the total analysis including pyramid construction runs at about 30 milliseconds on a 700 MHz Pentium III on an average. Starting at higher levels causes problems in our field of applications, since relevant structures may be lost.

Fig. 2.1 shows the disparity image that we get by this scheme for the situation of Fig. 1.1.

2.2 Motion Analysis

Stereo object detection usually is done by clustering disparity features to gather 3D objects. As mentioned in the introduction, this method is not sufficient if the distance between two objects is lower than a predefined threshold. Objects with a close distance will merge to a single object even if velocities vary. For a fast detection of moving objects, regardless size and distance, it is necessary to measure motion within the images directly.

Based on performance comparison of a number of optical flow techniques, emphasizing the accuracy and density of measurements on realistic image sequences [6], we are using a basic differential (gradient based) optical flow method after Lukas and Kanade [13].

The gradient based method assumes that gray values of moving objects do not change over time which is usually the case in a wide range of our environmental scenes. The computation of the optical flow is illustrated by Fig. 2.2. which is leading to the one dimensional continuity equation

Fig. 2.3: computed gradient flow field

(2.1) where the gray value shift Δu is given as the ratio between the temporal and spatial derivatives g_t and g_u.

$$g_t + \Delta u \cdot g_u = 0 \qquad (2.1)$$

$$g_t + \Delta u \cdot g_u + \Delta v \cdot g_v = 0 \qquad (2.2)$$

Accordingly, equation (2.2) can be derived for the two dimensional case. The two dimensional optical flow $(\Delta u, \Delta v)$ is given by the least squares solution of (2.2) within a small image region. As an example, the resulting optical flow field is shown in Fig. 2.3.

Of course, many different methods for optical flow computation like region-based matching [8], energy-based [9] and phase based [10] methods are available. The basic gradient method can also be

improved by using either second order derivatives or smoothness constraints for the flow field [11].

However, none of the above methods is capable to compute dense optical flow fields under real-time conditions. Usually, special hardware and parallel processing is needed in order to reach acceptable frame rates whereas the basic gradient flow can be computed in real-time on a standard PC. Furthermore, we will show that in combination with stereo the basic method is more than sufficient for our detection problem.

3 Fusion of Stereo and Motion

Both methods, stereo and motion, have certain disadvantages for object detection. As described above, stereo extracts depth information without correlation over time. The optical flow on the other hand is able to detect even small gray value changes providing the possibility for early detection of moving objects. But with a moving camera, it lacks from suppression of background-flow without depth information.

In order to use the information of both systems in an optimal way, we suggest a sensor fusion method. We will show that with the proposed

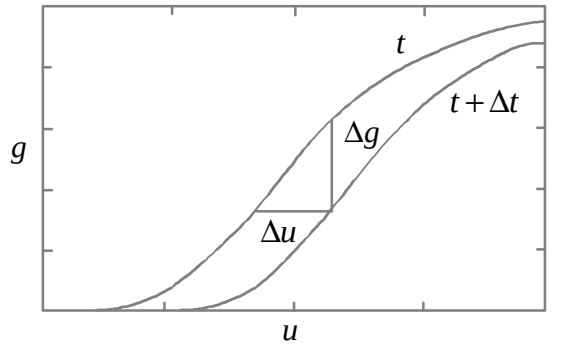

Fig. 2.2: gradient based optical flow

fusion of stereo and motion both methods supplement their shortcomings leading to a robust detection of arbitrary moving objects.

3.1 Flow/Depth constraint

Let us assume a purely longitudinal moving camera and a stationary environment for the moment. For the transformations between the 3D world coordinate system (x, y, z) and the corresponding 2D image coordinate system (u, v), we are using a pinhole camera model with the focal length f and s_u as the size of a sensor element of the camera chip. With the pinhole camera model and the stereo base line b, we can derive the disparity D and the optical flow $(\dot{u}, \dot{v})$ from triangulation leading to the following equations:

$$D = \frac{f \cdot b}{s_u \cdot z} \quad , \quad \frac{\dot{u}}{u} = \frac{\dot{z}}{z} \quad , \quad \frac{\dot{v}}{v} = \frac{\dot{z}}{z} \qquad (3.1)$$

Both, disparity and optical flow, depend on the real-world depth z. Therefore, the optical flow field can be computed from depth information and vice versa for stationary objects.

However, computation of the real-world depth is not necessary in our case. Switching variables for vehicle speed $\dot{z} = \Delta s$ and the horizontal and vertical components of the optical flow $\dot{u} = F_u$, $\dot{v} = F_v$, the depth factor is eliminated by building the quotient between the optical flow and the disparity. Separately applied to the horizontal and vertical components of the optical flow, this leads to the following constraints:

$$\frac{F_u}{D} = \frac{s_u \cdot \Delta s}{b \cdot f} \cdot u \qquad \frac{F_v}{D} = \frac{s_u \cdot \Delta s}{b \cdot f} \cdot v \qquad (3.2)$$

Equations (3.2) can be illustrated by inclined planes over the image region (u, v). The gradient of the planes is determined by the stereo base line b, the size of a sensor element of the camera chip s_u, the focal length f and the vehicle speed Δs [m/frame]. Fig. 3.1 shows this plane for the horizontal component of equation 3.2.

Using our in-vehicle stereo camera system, the camera parameters f, s_u and b usually remain constant while only the speed varies over time. Therefore, the inclination of the plane changes as a function of the vehicle's speed only.

The quotient values for pixels belonging to stationary objects will match the plane. If the quotient does not match the value of the plane, we have to consider a moving object at this image position. Fig. 3.2 shows four consecutive images of a test sequence. All objects within the scene are stationary except one vehicle which backs into the street from the right while the camera is moving forward. The flow/depth quotient is computed for one line in the image center only. The corresponding values are displayed in blue. The value of the quotient plane is displayed in red. If stationary objects are present, the quotient measurements follow the predefined value of the plane. Quotient values corresponding to the moving object vary distinctively from the plane.

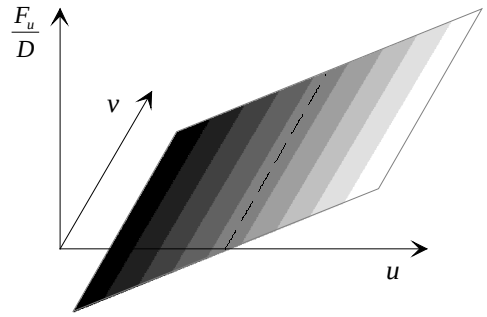

Fig. 3.1: flow/depth quotient plane

3.2 Quotient Noise

As we see from Fig. 3.2, there is some measurement noise from the underlying stereo and optical flow within the flow/depth quotient which complicates segmentation of moving objects. But since the measurement noise for the disparity and optical flow preprocessing is well known, we can derive the maximum error of the quotient and use it as a threshold function for the segmentation.


Fig. 3.2: deviation of flow/depth value (blue) from the quotient plane (red) if a moving object

From

$$Q + \Delta Q = \frac{F + \Delta F}{D + \Delta D} = \frac{F^*}{D^*} \qquad (3.3)$$

we get the following function for the maximum quotient error for the horizontal and vertical flow, respectively:

$$\Delta Q = \frac{1}{D^*}(\Delta F - \Delta D Q) \qquad (3.4)$$

where Q is the value of the flow/depth plane and D^* is the current measurement value for the disparity. $\Delta D, \Delta F$ are the known maximum errors for the disparity and optical flow preprocessing. Together with equation (3.2) the maximum error of the horizontal quotient value is given by:

$$\Delta Q = \frac{1}{D^*}(\Delta F - \Delta D \cdot \frac{s_u}{b \cdot f} \cdot u \cdot \Delta s) \qquad (3.5)$$

Except u, equation (3.5) is the same for the vertical quotient. Fig. 3.3 illustrates the maximum allowed deviation from the plane, which basically is the absolute value of equation (3.5). We will use this as the threshold function.

Segmentation of moving objects is a three step process:

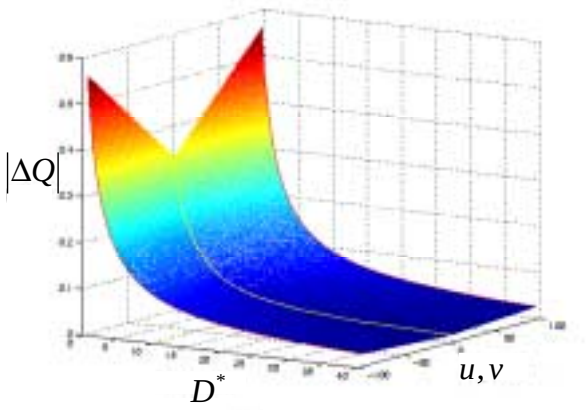

Fig. 3.3: maximum flow/depth quotient noise

1. Compute the horizontal and vertical quotients from the optical flow and the disparity for every pixel (u,v) for which depth and motion information is present.
2. Compare the computed quotient values with the reference values from the flow/depth plane at position (u,v) multiplied with the known vehicle speed Δs.
3. Tag image position (u,v) as "moving object" if the difference between reference value and quotient is more than $|\Delta Q|$ in at least one direction.

3.3 Stabilization

So far, pure longitudinal camera motion was assumed. We use the above method within a demonstrator vehicle where the camera is mainly moving in longitudinal direction. Additionally, there are rotational components about all three axes.

There is a distinct flow pattern corresponding to rotation and translation along every camera axes. As the camera movement is a combination of camera translation and rotation, the optical flow is a linear combination of independent components.

In order to use the flow/depth constraint as described above, we have to stabilize the image so that all rotational components are zero and only the translational flow remains.

Our stabilization is estimating self-motion using a matched filter method [12]. Each filter is tuned to one flow pattern generated by either camera pitch, yaw or roll according rotation for the three camera axes. We assume that the flow preprocessing stage provides the optical flow as an input to the matched filters. The elimination of the rotational flow components is done in three steps:

1. Compute the filter output from the weighted sum of the scalar product between the optical flow and the matched filter pattern at each image position. This results in a single scalar which is the rotational speed for this axis.
2. An estimate for the rotational flow field is given by the product of the matched filter pattern and the rotational speed from the first step.
3. The compensated flow is given by the difference between the measured optical flow and the estimated rotational flow field from step 2.

The method is very well adapted to our stabilization task. Based on the optical flow which we take from the preprocessing stage there is only few extra computational power needed for the stabilization within every image cycle. The matched filter patterns for all three axes do not change over time, so they can be computed only once when the system is initialized. If we assume, that the optical flow is present for *n* pixels within the image, we only need *2n* MUL, 2*n-1* SUM and *1* DIV operation to compute the rotational speed from step 1. The flow prediction from step 2 needs *2n* MUL and the compensation from step 3 needs 2*n* SUB operations.

3.4 Results

The system has been tested on several inner city image sequences with pedestrians involved. As an example, one of these scenes is shown in Fig. 3.4.

The sequence has been taken from our in-vehicle stereo camera system. The vehicle speed is 18 km/h and the slight pitch and yaw movement of the camera has been compensated by the described matched filter method.

The result of the flow/depth constraint is overlaid onto the image. As can be seen, the algorithm is very sensitive to movements that don't match the motion of a static environment with respect to the moving camera, while background noise is very low.

The robust and fast detection can only be achieved because our fusion method is using the information from the stereo and the optical flow subsystems in an optimal way. The head of the child is detected within only three image frames after its first appearance behind the front window of the car. From stereo or optical flow alone this wouldn't be possible.

The detection is independent of size or shape of the object. Since everything is done on a small pixel based neighborhood, the detection even works for non-rigid motion from pedestrians where motion varies for different parts of the body. However, in Fig. 3.4, the motion of legs and arms with respect to their size is fairly high and therefore out of the measuring range of our current motion analysis.

The flow/depth constraint also works on areas where the flow is zero. Due to the fact that our camera is moving, a flow equals zero does not automatically mean a zero risk. The subimages in Fig. 3.4 have been cropped from the original video at a fixed position. Even though the child is moving with respect to world coordinates, there is almost zero optical flow for the child's head since its position within the image stays nearly constant. But as one can see there is no difference in detection even under this extreme conditions.

The current system works for low vehicle speed. Due to our optical flow algorithm, the range for valid image motion is restricted to ± 2 pixel/frame. With the current camera and a video rate of 25 frames/s, this restricts the maximum vehicle speed to 25km/h. As flow range is limited, the used stabilization is optimal for small rotational velocities only.

In order to overcome this restrictions, we are working on a multi-scale approach for optical flow which will extend the current measurement range.

4 Summary

The early detection of dangerous situations in urban traffic is a serious challenge for image understanding systems. Up to now, we had stereo vision to detect obstacles in front of the car only.

The presented fusion of stereo and motion analysis is a new powerful scheme that allows early detection of moving obstacles even if they are

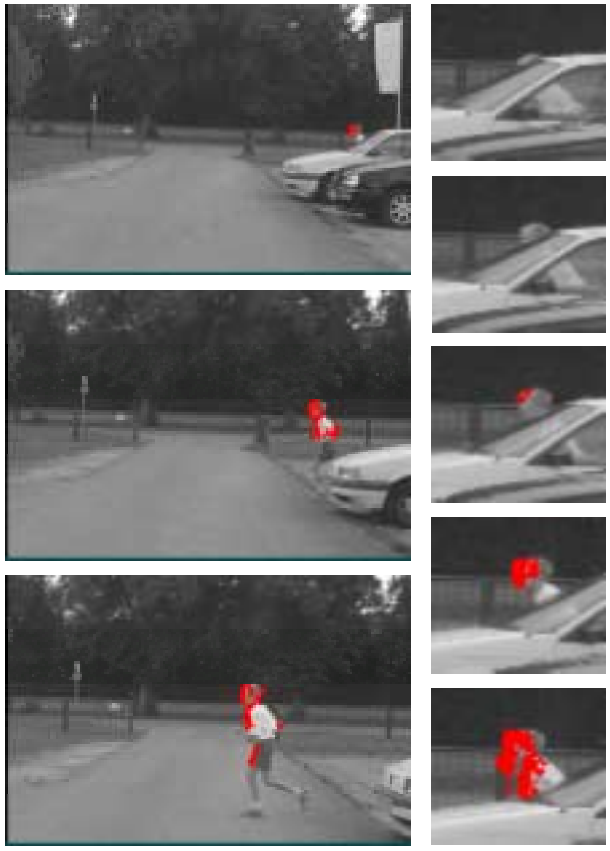

Fig. 3.4: detection of a child within an image sequence taken from a moving camera

partially occluded and non-rigid. The disparity information is already available in our vehicle and the simple motion analysis runs in real-time, too. Since the fusion algorithm has to compare the flow/depth quotient against a threshold function at distinct points only, it is computationally highly efficient. Its current limitation to low vehicle speed due to the used optical flow measurement shall be overcome by means of a multi-scale approach.

References

[1] U.Franke, D.Gavrila, S.Görzig, F.Lindner, F.Paetzold, C.Wöhler: "Autonomous Driving Goes Downtown", *IEEE Intelligent Systems*, Vol.13, No.6, Nov./Dec.1998, pp.40-48

[2] U.Franke, D.Gavrila, A.Gern, S.Goerzig, R.Janssen, F.Paetzold and C.Wöhler: "From door to door – principles and applications of computer Vision for driver assistant systems", in *Intelligent Vehicle Technologies: Theory and Applications*, Arnold, 2001

[3] U.Franke: "Real-time Stereo Vision for Urban Traffic Scene Understanding", *IEEE Conference on Intelligent Vehicles 2000*, October, Detroit

[4] C. Wöhler, J. K. Anlauf. An Adaptable Time Delay Neural Network Algorithm for Image Sequence Analysis. *IEEE Transactions on Neural Network*, vol. 10, no. 6, pp. 1531-1536, 1999.

[5] U.Franke, A.Joos, B.Aguirre: "Early Detection of potentially harmful situations with children", *Intelligent Vehicles 2001*, Tokyo, Mai 2001

[6] J.L. Barron, D.J.Fleet, S.S. Beauchemein: "Performance of Optical Flow Techniques", *International Journal of Computer Vision 1*, 1994

[7] W.B. Thompson and Ting-Chuen Pong: "Detecting Moving Objects", *Int. Journal of Comp. Vision 4*, 1990

[8] P. Anandan; "A computational framework and an algorithm for the measurement of visual motion", *Int. Journal of Comp. Vision 2*, 1989

[9] D.J. Heeger: "Optical flow using spatiotemporal filters", *Int. Journal of Comp. Vision 1*, 1988

[10] D.J. Fleet and A.D. Jepson: "Computation of component image velocity from local phase information", *Int. Journal of Comp. Vision 5*, 1990.

[11] H.H. Nagel: "Displacement vectors derived from second-order intensity variations in image sequences", *Comp. Graph. Image Processing 21*, 1983

[12] M.O. Franz: "Minimalistic Visual Navigation", *VDI Reihe 8 Nr. 739*, 1999

[13] B. Lucas and T. Kanade: "An iterative image registration technique with an application to stereo vision", *Proc. 7th Int. Conf. On Artificial Intelligence*, 1981