6D-Vision: Fusion of Stereo and Motion for Robust Environment Perception

Uwe Franke, Clemens Rabe, Hernán Badino, and Stefan Gehrig

DaimlerChrysler AG, 70546 Stuttgart, Germany
{uwe.franke,clemens.rabe,hernan.badino,stefan.gehrig}@daimlerchrysler.com

Abstract. Obstacle avoidance is one of the most important challenges for mobile robots as well as future vision based driver assistance systems. This task requires a precise extraction of depth and the robust and fast detection of moving objects. In order to reach these goals, this paper considers vision as a process in space and time. It presents a powerful fusion of depth and motion information for image sequences taken from a moving observer. 3D-position and 3D-motion for a large number of image points are estimated simultaneously by means of Kalman-Filters. There is no need of prior error-prone segmentation. Thus, one gets a rich 6D representation that allows the detection of moving obstacles even in the presence of partial occlusion of foreground or background.

1 Introduction

Moving objects are the most dangerous objects in many applications. The fast and reliable estimation of their motion is a major challenge for the environment perception of mobile systems and of driver assistance systems in particular. The three-dimensional information delivered by stereo vision is commonly accumulated in an evidence-grid-like structure [9]. Since stereo does not reveal any motion information, usually the depth map is segmented and detected objects are tracked over time in order to obtain their motion. The major disadvantage of this standard approach is that the performance of the detection highly depends on the correctness of the segmentation. Especially moving objects in front of stationary ones – eg. the bicycle in front of the parking vehicles shown in figure 1 – are often merged and therefore not detected. This can cause dangerous misinterpretations and requires more powerful solutions.

Our first attempt to overcome this problem was the so called flow-depth constraint [6]. Heinrich compared the measured optical flow with the expectation stemming from the known ego-motion and the 3D stereo information. Independently moving objects do not fulfil the constraint and can easily be detected. Unfortunately, this approach turned out to be very sensitive to small errors in the ego-motion estimation, since only two consecutive frames are considered.

Humans do not have the above mentioned problems since we simultaneously evaluate depth and motion in the retinal images and integrate the observations over time [10]. The approach presented in this paper follows this principle. The

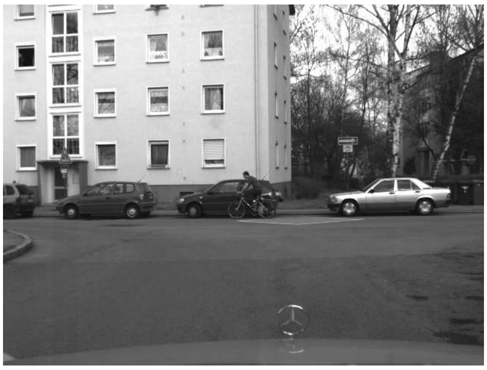


Fig. 1. Typical scene causing segmentation problems to standard stereo systems.

basic idea is to track points with depth known from stereo vision over two and more consecutive frames and to fuse the spatial and temporal information using Kalman Filters. The result is an improved accuracy of the 3D-position and an estimation of the 3D-motion of the considered point at the same time. Since we get a rich 6D-state vector for each point we like to refer to this method as 6D-Vision. Taking into account the motion information, the above mentioned segmentation problem can be solved much more easily and robust.

The fusion implies the knowledge of the ego-motion. Fortunately, practical tests reveal that for the detection of moving obstacles in front of a moving car the standard inertial sensors are sufficient. If neccessary, the ego-motion can be computed from image points found to be stationary (e.g. see [8] or [1]).

The mentioned accuracy improvement is already exploited by a satellite docking system described in [7]. After an application-specific initialization, predefined markers are tracked in the images of a pair of stereo cameras yielding a very precise estimation of the relative position. In [3] Dang combines stereo and motion to decide whether a group of points underlies the rigid motion.

In our real-time application we track about 2000 image points. So far, the best results are obtained using a version of the well-known Kanade-Lucas-Tomasi (KLT) tracker [11] that was optimized with respect to speed. The depth estimation is based on a hierarchical correlation based scheme [4]. However, any comparable optical flow estimation and any other stereo system can be used.

The paper is organized as follows: section 2 describes the system model and the measurement equation for the proposed Kalman Filter. Section 3 studies the rate of convergence of the considered system and presents a multi-filter system for improved convergence. Section 4 gives practical results including crossing objects and oncoming traffic.

2 System Description

In the following we use a right handed coordinate system with the origin on the road. The lateral x-axis points to the left, the height axis y points upwards and the z-axis represents the distance of a point straight ahead. This coordinate system is fixed to the car, so that all estimated positions are given in the coordinate system of the moving observer. The camera is at $(x, y, z)^T = (0, height, 0)^T$ looking in positive z-direction.

2.1 System Model

The movement of a vehicle with constant velocity v_c and yaw rate $\dot{\psi}$ over the time interval Δt can be described in this car coordinate system as

$$\Delta \underline{x}_c = \int_0^{\Delta t} \underline{v}_c(\tau)\, d\tau = \frac{v_c}{\dot{\psi}} \begin{pmatrix} 1 - \cos \dot{\psi} \Delta t \\ 0 \\ \sin \dot{\psi} \Delta t \end{pmatrix} .$$

The position of a world point $\underline{x} = (X, Y, Z)^T$ after the time Δt can be described in the car coordinate system at time step k as

$$\underline{x}_k = R_y(\psi)\left(\underline{x}_{k-1} + \underline{v}_{k-1}\Delta t - \Delta \underline{x}_c\right)$$

with its associated velocity vector $\underline{v}$ and the rotational matrix around the y-axis $R_y(\psi)$. Combining position and velocity in the 6D-state vector $\underline{\tilde{x}} = (X, Y, Z, \dot{X}, \dot{Y}, \dot{Z})^T$ leads to the discrete system model equation

$$\underline{\tilde{x}}_k = A_k \underline{\tilde{x}}_{k-1} + B_k v_c + \underline{w}_{k-1}$$

with the state transition matrix

$$A_k = \begin{pmatrix} R_y(\psi) & | & \Delta t R_y(\psi) \\ — & & — \\ 0 & | & R_y(\psi) \end{pmatrix}$$

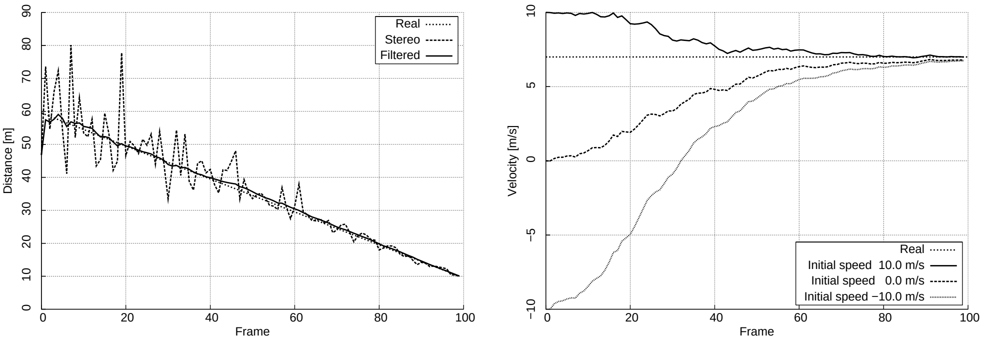
and the control matrix

$$B_k = \frac{1}{\dot{\psi}} \begin{pmatrix} 1 - \cos\left(\dot{\psi}\Delta t\right) \\ 0 \\ -\sin\left(\dot{\psi}\Delta t\right) \\ 0 \\ 0 \\ 0 \end{pmatrix} .$$

The noise term $\underline{w}$ is assumed to be Gaussian white noise with covariance matrix Q.

2.2 Measurement Model

The measurement consists of two pieces of information: the image coordinates u and v of a tracked feature and the disparity d delivered by stereo vision working



(a) Distance estimation of static world point.

(b) Velocity estimation of moving world point at a speed of $v_z = 7.0\,\frac{\mathrm{m}}{\mathrm{s}}$ using three different initializations.

Fig. 2. Estimation results of the presented Kalman Filter. The considered world point is at the initial position $(10.0\,\mathrm{m}, 1.0\,\mathrm{m}, 60.0\,\mathrm{m})^T$. The observer moves at a constant speed of $v_z = 10\,\frac{\mathrm{m}}{\mathrm{s}}$ in positive z-direction (20 fps).

on rectified images. Assuming a pin-hole camera the non-linear measurement equation for a point given in the camera coordinate system is

$$\underline{z} = \begin{pmatrix} u \\ v \\ d \end{pmatrix} = \frac{1}{Z} \begin{pmatrix} X f_u \\ Y f_v \\ b f_u \end{pmatrix} + \underline{\nu}$$

with the focal lengths f_u and f_v and the baseline b of the stereo camera system. The noise term $\underline{\nu}$ is assumed to be Gaussian white noise with covariance matrix R.

2.3 Simulation Results

The benefit of filtering the three-dimensional measurement is illustrated by figure 2(a). It shows the estimated relative distance of a simulated static world point measured from an observer moving at a speed of $10\,\frac{\mathrm{m}}{\mathrm{s}}$. The initial position of the point is $(10.0\,\mathrm{m}, 1.0\,\mathrm{m}, 60.0\,\mathrm{m})^T$. White gaussian noise was added to the image position and the disparity with a variance of $1.0\,\mathrm{px}^2$. The dashed curve shows the unfiltered 3D position calculation which suffers from the additive noise. The continuous curve gives the excellent result of the filter.

In the above example the speed of the point was correctly initialized to zero. How does the filter perform if the point is in motion? Let us assume the point moves at a speed of $v_z = 7.0\,\frac{\mathrm{m}}{\mathrm{s}}$ in positive z-direction. Figure 2(b) shows the estimation results of three differently initialized filters. Although very large initial values of the P-Matrix are used, the speed of convergence is only fair. Better results can be obtained by a multi-filter approach described in the following.

3 Multiple Filters for Improved Rate of Convergence

As shown above in figure 2(b), the closer the first guess is to the correct value, the less time it takes until the estimate is below a given error threshold. This strong dependency on the initial value can be overcome by running multiple Kalman Filters initialized at different speeds in parallel, estimating the world position and velocity using the same input data.

How can we decide which state is the best? One way is to calculate the distance between the real measurements and the predicted measurements using the Mahalanobis distance, also known as the normalized innovation squared (NIS) [2]:

$$D_M\left(\underline{z}, \underline{x}\right) = \left(\underline{z} - \underline{x}\right) \Sigma^{-1} \left(\underline{z} - \underline{x}\right)^{\top}$$

with the measurement $\underline{z}$, the predicted measurement $\underline{x}$ and the innovation covariance matrix Σ.

Alternatively, the probability density function, also called likelihood, can be used as an indicator to decide whether a given measurement $\underline{z}$ matches a certain Kalman Filter model. This is used for example in the interacting multiple model estimator (IMM) [2]. However, the likelihood calculation tends to suffer from too small floating point data types.

In order to avoid these numerical problems, we base our decisions on the NIS criterion. Figure 3(a) shows the low pass filtered NIS values for the three differently initialized filters of figure 2(b). It is obvious that the initialization quality corresponds to the discrepancy in measurement space between the measured and predicted position.

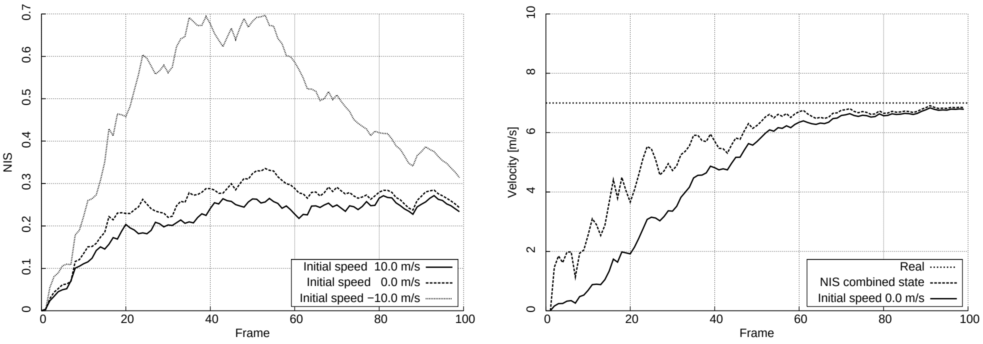
Selecting one of the three (in general n) filter states as the correct one would ignore valuable information contained in the other filters. Assuming a limited initial state space, i.e. the tracked point has a limited absolute velocity, we initialize the filters on different velocities including the boundaries. Consequently, the real state must lie in between these boundaries and can be expressed as a weighted sum

$$\widetilde{\underline{x}} = \frac{1}{\sum \beta_i} \sum_{i=0}^{n} \beta_i \underline{x}_i \qquad \text{with} \qquad \beta_i = \frac{1}{NIS_i}$$

where the weights β_i represent the matching quality of each Kalman Filter.

It is beneficial not to base the decision or weighting on the current measurement quality only, since this would lead to undesired effects due to measurement noise. Therefore, we apply a low pass filtering to the weights thus accumulating the errors over a certain time.

Figure 3(b) shows the result obtained by the above approach, if the three filters are initialized at speeds -10.0, 0.0 and 10.0 $\frac{\mathrm{m}}{\mathrm{s}}$. For comparison, the same filter initialized at -10.0 $\frac{\mathrm{m}}{\mathrm{s}}$ shown in figure 2(b) is considered. It can be seen that the multi-filter approach converges two times faster than the simple one. A comparison with figure 2(b) reveals that the combined system shows a better performance than each of the three single filters.



(a) Low pass filtered normalized innovation squared of the filters from figure 2(b).

(b) Improved convergence obtained by the Multi Kalman Filter system.

Fig. 3. Multi Kalman Filter estimation result (20 fps).

4 Real World Results

The five most probable practical situations are: stationary objects, vehicles driving in the same direction as our own vehicle with small relative speed, oncoming traffic, traffic from left, and traffic from right. The sketched multi-filter approach offers the chance to run independent filters tuned to any of these five situations in order to reach the desired fast convergence.

Let us first concentrate on the crossing situation already shown in figure 1. The result of the velocity estimation is given in figure 4. The cyclist drives in front of parking vehicles while the observer moves towards him at a nearly constant speed of 4 $\frac{\mathrm{m}}{\mathrm{s}}$. The arrows show the predicted position of the corresponding world point in 0.5 s projected into the image. The colors encode the estimated lateral speed; the warmer the colour the higher the velocity. In order to prove the results, the right image in figure 4 shows the same situation 0.5 s later. As can be seen, the prediction shown in the left image was very accurate.

Figure 5 shows the estimation results for a typical oncoming traffic situation in which the observer moves at a constant speed of 50 $\frac{\mathrm{km}}{\mathrm{h}}$. Here the color encodes the Euclidean velocity of the tracked points. The prediction matches the real position shown in the right image.

Practical tests at intersections and on highways reveal that the algorithm is highly robust with respect to measurement noise, induced by pitch and yaw of the vehicle. Simply spoken, it doesn't matter how a point in the world precisely moves from A to B, because those details are filtered out by the Kalman Filter. On the other hand, it turns out that measurement outliers sometimes cause serious misinterpretations. This problem is overcome by using a standard 3σ-test to detect and reject those outliers.

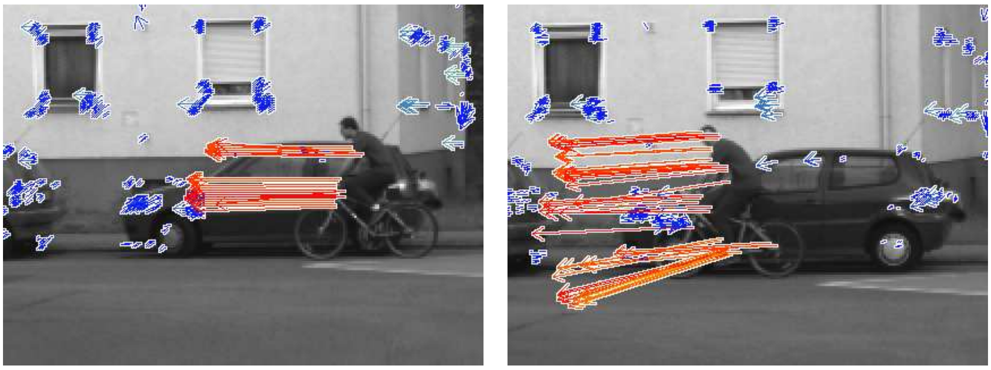


Fig. 4. Velocity estimation results for cyclist moving in front of parking cars. The arrows show the predicted position of the world point in 0.5 s. The right image was taken 0.5 s later allowing a comparison of the estimation from the left image. Blue encodes stationary points.

5 Summary

The proposed fusion of stereo and optical flow simultaneously improves the depth accuracy and allows estimating position and motion of each considered point. Segmentation based on this 6D-information is much more reliable and a fast recognition of moving objects becomes possible. In particular, objects with certain direction and speed of motion can directly be detected on the image level without further non-linear processing or classification steps that may fail if unpredicted objects occur.

Since the fusion is based on Kalman Filters, the information contained in a number of frames is integrated. This leads to much more robust estimations than differential approaches like pure evaluation of the optical flow. The proposed multi-filter approach adopted from our depth-from-motion work [5] speeds up the rate of convergence of the estimation, which is important for fast reactions. For example, practical tests confirm that a crossing cyclist at an intersection is detected within 4-5 frames. The implementation on a 3.2 GHz Pentium 4 proves that the described approach runs in real-time. Currently, we select and track about 2000 image points at 12-16 Hz, depending on the used optical flow algorithm (the images have VGA resolution).

The ego-motion is assumed to be known throughout the paper. For many in-door robotic applications on flat surfaces the usage of inertial sensors will be sufficient. At the moment, the ego-motion of our demonstrator vehicle (UTA, a Mercedes Benz E-Class vehicle) is determined based on the inertial sensors only. Thanks to the Kalman Filter, the results are sufficient for obstacle avoidance. The most dominant pitching motion results in an apparent vertical motion that is ignored for this application. Nevertheless, in order to reach maximum accuracy, the next step will be to estimate the six degree of freedom ego-motion precisely using those image points that have been classified as static.

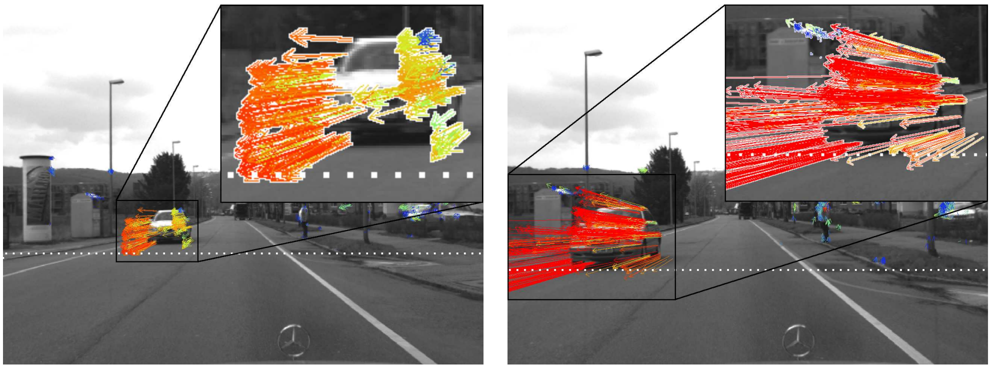


Fig. 5. Velocity estimation results for oncoming car. The observer moves at a constant speed of 50 $\frac{\mathrm{km}}{\mathrm{h}}$.

References

1. Badino, H.: A Robust Approach for Ego-Motion Estimation Using a Mobile Stereo Platform. 1^{st} International Workshop on Complex Motion (IWCM04), Günzburg, Germany, October 2004
2. Bar-Shalom, Y., Kirubarajan, T., Li, X.: Estimation with Applications to Tracking and Navigation. John Wiley & Sons, Inc., 2002
3. Dang, T., Hoffmann, C., Stiller, C.: Fusing Optical Flow and Stereo Disparity for Object Tracking. Proc. of the IEEE V. International Conference on Intelligent Transportation Systems, Singapore, 3-6 September, pp. 112–117, 2002.
4. Franke, U.: Real-time Stereo Vision for Urban Traffic Scene Understanding. IEEE Conference on Intelligent Vehicles 2000, October 2000, Dearborn
5. Franke, U., Rabe, C.: Kalman Filter based Depth from Motion Estimation with Fast Convergence. to appear in IEEE Conference on Intelligent Vehicles 2005, June 2005, Las Vegas
6. Heinrich, S.: Real Time Fusion of Motion and Stereo using Flow/Depth constraint for Fast Obstacle Detection. DAGM 2002, September 2002
7. Lin, C.-F.: Three-dimensional relative positioning and tracking using LDRI. US patent no 6,677,941, Jan. 13, 2004
8. Mallet, A., Lacroix, S., Gallo, L.: Position estimation in outdoor environments using pixel tracking and stereovision. Proc. of the 2000 IEEE International Conference on Robotics and Automation ICRA 2000, pp. 3519–3524, San Francisco, April 2000
9. Martin, M., Moravec, H.: Robot Evidence Grids. The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, March 1996 (CMU-RI-TR-96-06)
10. Sekuler, R.,Watamaniuk, S.,Blake R.: Perception of Visual Motion. In Stevens' Handbook of Experimental Psychology, Volume 1, Sensation and Perception, Jan. 2004
11. Tomasi, C., Kanade, T.: Detection and Tracking of Point Features. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, April 1991 (CMU-CS-91-132)