IEEE Workshop on 'Visual Motion'
Princeton, NJ., Oct. 7-10,1991

Simultaneous Estimation of 3D Shape and Motion of Objects by Computer Vision

Jens Schick, Ernst D. Dickmanns

Universität der Bundeswehr Munich W.
Heisenberg Weg 39, D-8014 Neubiberg, Germany

Abstract: A recursive estimation method based on the 4D-approach to real-time computer vision for simultaneously determining both 3D shape parameters and motion state of objects is discussed. The recognition processes exploit structurally given shape models and motion models given by difference-equations. This allows to confine the image analysis to feature evaluation of the last frame of the sequence only; no differencing between images has to be done, yet the spatial motion components (satisfying planar motion constraints) are recovered directly without inverting the perspective projection equations of the imaging process explicitly.

Object recognition has been confined to a well-structured, but otherwise general dynamic scene for the beginning: road traffic with a limited class of vehicles.

1 Introduction

When image sequence interpretation is conceived as a measurement process in which only partial output variable measurements (as opposed to direct state variable measurements) are possible and in which the full state vector of the dynamical system has to be recovered exploiting a generic dynamical model of the real world process, a completely different approach to machine vision than the one favored by AI-groups results. In [2] this approach had been pioneered. It took five years until in 1987, based on experience in four different application areas, the general scheme for dynamic machine vision became validated by real world experiments [3].

Initially, only the egomotion of the system relative to a static environment could be handled in real time with the processing power available. Then, one additional moving object of simple appearance could be dealt with [4j; both its relative 3D state including spatial velocity components and its size (height and width) could be determined. In the present article, this approach is expanded to the recognition of 3D bodies and their trajectories satisfying a generic shape description each; both the actual shape parameters of the object and of the trajectory as well as the state vector of the system are estimated recursively by parallel processes in real time. Central to all these activities is the integral spatio-temporal approach based on edge element feature extraction and object oriented feature aggregation ('Gestalt' idea) which will be discussed in the next section.

2 The 4D approach to dynamic vision

The main goal of this approach from the beginning has been to take advantage of the full spatio-temporal framework for internal representation and to do as few reasoning as possible in the image plane and hi between frames. Instead, temporal continuity in physical space according to some model for the motion of objects is being exploited in conjunction with spatial shape rigidity in this **'analysis-by-synthesis'** approach.

Dynamical models link time to spatial motion, in general. The shape models exhibit the spatial distribution of visual features on the surface which allow objects to be recognized and tracked. In order to exploit both types of models at the same time, the prediction error feedback scheme for recursive state estimation developed by Kalman and successors [1] has been extended to image sequence processing by our group. There are many publications on this approach so that only Fig. 1 will be given here as a short summary (see e.g. the survey article [3] and the references given there).

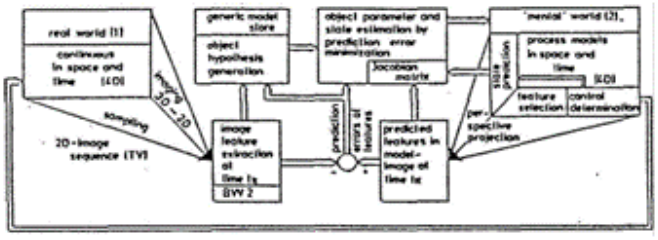


Figure 1: Survey block diagram of the 4D approach to dynamic machine vision

In the initialization phase there is hi general context dependent generic knowledge about the task domain available. This should be exploited for hypothesis generation as far as possible. In the highway traffic environment of interest here, generic models of the road and of different types of vehicles are readily available. The general 3D road recognition problem is treated hi [5]; the vehicle part is treated in the sequel.

3 A priori knowledge about objects

Confining the discussion to those objects which are allowed to participate hi the task domain under standard

conditions allows to reduce the variety of objects and situations considerably. For a well-defined task area like highway driving there is an appreciable amount of background knowledge available as to what can be expected under normal conditions:

1. Traffic participants are massive vehicles pulled to the ground by Earth gravity; direct lateral translation is not possible due to frictional constraints between tires and ground. Control degrees of freedom are: longitudinally by engine propulsion or by braking, and laterally by steering control of front wheels which is characteristically linked to yawing rotation around the vertical axis. This knowledge is best represented by dynamical models for vehicle motion capabilities.
2. The size and shape of highway traffic participants is geared to their transportation task of carrying people and goods. Most frequent are passenger cars of dimensions [width, height and length hi m] around (1.5 to 2) * (1.3 to 1.7) * (3 to 6). Then- shape usually is very close to mirror symmetry relative to a vertical longitudinal plane.
3. The roads are smoothly curved both horizontally and vertically; on highways the radius of curvature is of the order of magnitude of hundreds of meters. Usually, roads have more than one lane, each of which is about 2.5 to 4 m wide.
4. Independently moving objects usually are well apart by a safety distance depending on the speed travelled (a '2 seconds'-distance is recommended, though very often disregarded by drivers).

Knowing what to look for alleviates the task considerably and allows to make processing of the huge amounts of data from TV-cameras more goal oriented and more intelligent. Intelligence hi the common sense means the capability that physical or mental processes can be recognized by observing only small but somehow characteristic parts of that process (and of properly adjusting the own behavior to this).

4 Generic shape models

A generic 3D shape model defines the spatial structure of the shape without specifying the values of the parameters occurring in that model. In fig. 2 an idealized generic polyhedral 3D passenger car model is given in four orthonormal views: Top down view (center), front and back view (left and right), left side view (bottom). The different edge lengths are the model parameters yet to be determined in order to best match groups of perspectively mapped image features. The types of features for measurements still have to be determined too. In the polyhedral model corner point positions at first glance seem to be reasonable features defining the shape; however, on real cars the corners are more or less rounded yielding no chance for precise measurements of corner positions since well visible silhouette points are rather sensitive to aspect condition changes on parts with medium curvature. Parts with either very strong or less pronounced curvature yield more stable contour feature positions under varying aspect angles.

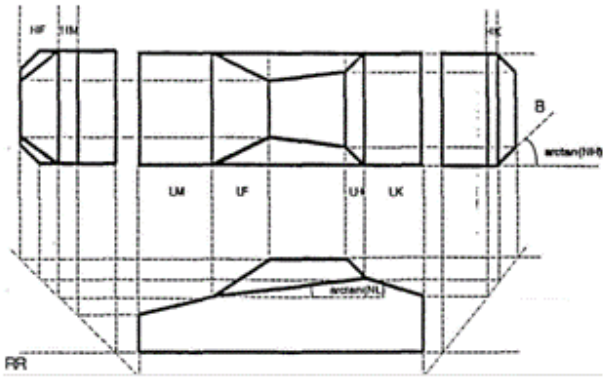


Figure 2: Idealized generic shape model of a passenger car

For these reasons, linearly extended edge- or contour elements and their relative positions with respect to each other may be best suited as features to be tracked and measured for spatial shape recognition hi the general case. Knowing what to look for, the choice of meaningful feature combinations is much reduced. This is what psychologists refer to as the 'Gestalt'-idea [8]: One is able to see an object definitely only after the overall frame has been set by some means in an ambiguous situation. Features which are less pronounced from a mere image processing point of view may become the major ones supporting the object hypothesis despite the presence of well visible other ones which just happen to be there by pure chance; a typical example for this situation is the existence of shadows hi the scene with strong variations hi brightness over the shadow boundary while the object boundary within or outside the shade is much less pronounced. This also points to the fact that hi more complex scenes there may be nothing like an optimal feature extractor, but that it may be background knowledge which makes the system work reasonably well.

Since what the system tries to interpret are sequences of 2D images of 3D scenes, hypotheses for the object and for the aspect conditions have to be generated hi conjunction. Knowing about the limited mobility of the objects hi the real world, this interpretation process can be stabilized over time by incorporating this knowledge about center of gravity (c.g.) motion into the recursive least squares estimation process. Shape rigidity links feature motion to that of and around the c.g. by taking perspective mapping into account.

The polyhedral generic 3D shape representation model for a passenger car as given in fig.2 consists of 12 polyhedrons with 26 edges and 16 corners. The body-fixed x-axis lies hi the plane of symmetry parallel to the base plane polyhedron, pointing into the forward direction; the y-axis points to the right, the z-axis positive downward hi the plane of symmetry. Each edge between two corner points may be given as the space diagonal of a rectangular box of length/width/height fx/fy/fz oriented

like the entire vehicle. Relative to the overall vehicle, these three length components in vehicle x,y,z-direction are the shape parameters to be estimated in parallel to the motion state.

For a model having 26 edges, 78 shape parameters are obtained. The number of independent shape parameters, however, is much less due to the following geometric constraints: The vehicle is symmetric to the plane $y=0$; at each corner point at least three edges meet (resulting in the closedness of the shape). Assuming that the bottom and the top plane of the vehicle are parallel to the wheel base, and that the lower front-, back- and left and right side planes are normal to it, the number of independent shape parameters is reduced to 12. These are represented by ten distances and two distance ratios (see fig.2): Overall length L, width B and height H-RR; height of front windshield HF, height and length of motorhood HM and LM, length of front windshield LF, length of rear windshield LH, length of trunk LK, height of trunk hood HK; two slopes NH and NL. The bottom line of the body is at height RR above the ground.

By proper choice of these shape parameters different types of vehicles may be represented: Selecting trunk length LK as zero results in a liftback; fixing also the motorhood length to zero results in a bus-like vehicle. The closedness constraints are straightforward and not detailed here.

5 Generic motion models

Motion understanding is achieved in two steps: 1. the x and y translational position and the rotational orientation of the object around the vertical axis have been estimated using an approximate dynamical model for planar motion to be discussed in the next subsection. 2. starting from these results, in a second interpretation step the vehicle motion is interpreted in an object oriented frame of reference as longitudinal and lateral egomotion of the object observed yielding insight into the control actuation applied. Space does not allow going into details. The basic approach is discussed in [4,9].

6 The estimation scheme

The shape parameters according to section 4 and the motion state variables according to the previous section are being iterated in parallel in order to minimize prediction errors for the features selected. The iteration scheme is an extension of the modified Kalman filter [1] including UDU factorization in a sequential formulation. Of course, all the nonlinear relations in the mathematical models for motion and measurements have to be linearised around the actual nominal values in order to achieve linear relationships required by the iteration scheme. The resulting Jacobian matrices carry all the information for performing a least squares fit in the spatio-temporal sense taking all the model constraints into account. Since convergence occurs over space and time simultaneously, the method has been dubbed '4D approach'. In order to avoid error prone analytical derivations for the elements of the Jacobian matrices, these may be computed numerically by proper differencing.

The shape parameters (fx, fy, fz) of the edges are iterated independently from each other; for this estimation step they are derived from the independent shape parameters by a set of algebraic relations. These relations contain the geometrical side constraints discussed above; shape components fixed by geometrical side constraints (like $fz=0$ for bottom or top plane edge elements) are not iterated.

The independently estimated new shape parameters in general will no more satisfy all geometrical constraints; these will have to be enforced by iteration including proper balancing and adjustment through averaging. This procedure deliberately disregards some measurement results for the sake of satisfying apriori known facts about physical objects to be recognized, like symmetry; it is hoped (and indeed achieved in the simulation experiments) that for frequent measurements the estimation process converges rapidly to a stable shape and motion interpretation close to the real one. The geometrical side constraints make the decoupled recursive estimation processes much more stable.

Fig.3 shows the partitioning of the entire recursive estimation task into three pipeline stages, each of which may be parallelized depending on the complexity of the object shape and the processing power of the nodes. The pipeline stages are: 1. 4D-estimation and prediction of object shape and state; 2. image prediction including visibility checking, feature selection and Jacobian matrix computation; 3. feature extraction from measured image.

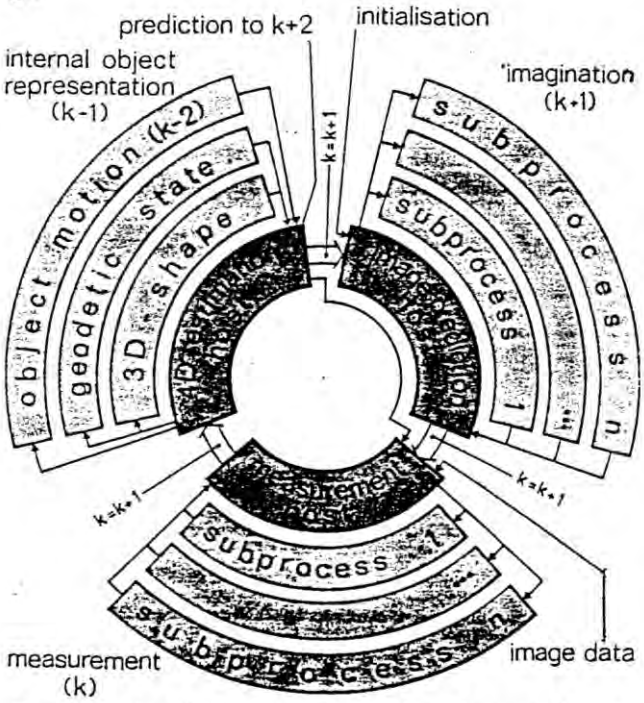


Figure 3: Parallelisation of recursive shape and state estimation into a three-stage pipeline

At time kT, measured feature data supplied to the estimation process are from frame (k-1)T, while this

process after finishing the estimation part makes a prediction over three cycles (3T) to provide the image prediction stage with data expected for (k+2)T; after time T has elapsed, the index k is incremented by one and data are transmitted to the next stage.

7 Experimental results

The general arrangement is given in fig.4. The vehicle initially is at rest at x = 20m, y = -3.125m heading straight to the right ($C_0 = 0$)under an angle of $\psi = 90^o$ (marked 'start'). The vehicle is controlled by the following control sequence given in table 1. With 50 ms simulated cycle time and 300 cycles maneuver length, the total duration is 15 seconds. At the end of phase 1 the vehicle speed is 5 m/s remaining constant during the clothoid arcs including a straight line section in the opposite direction (2 to 6); in phase 7 the vehicle is brought back to standstill again. The clothoid arcs are designed in such a way that at a heading change of 90^o the steering wheel turn rate is reversed leading to the opposite driving direction in phases 4 and 7. At the beginning of phase 7 the original straight line is reached again (loop closure).

phase No.	cycle No.	duration sec	long. acc. m/s^2	curv.rate rad/(s*m)	remarks
1	1-50	2.5	2	0	const. acc.
2	51-90	2	0	$\pi/20$	start right curve
3	91-130	2	0	$-\pi/20$	end right curve
4	131-157	1.35	0	0	straight, 180 ^{o}
5	158-197	2	0	$\pi/20$	start right curve
6	198-237	2	0	$-\pi/20$	end right curve
7	238-287	2.5	-2	0	const. decel.
8	288-300	.65	0	0	at rest

Table 1: Driving maneuver for oval path of fig.4

In order to be able to better visualize what type of images the system has to interpret, eight different perspective views along the trajectory are given for the positions marked.

Note that the directly measured quantities are not corner points which more easily are corrupted by noise (and may not even be there in images of real cars with rounded corners), but instead, several edge segments are combined by collinearity and continuity conditions into a best fitting line. This again helps smoothing noisy data, and this approach may be easily extended to the more general case of curved edges abundant on real cars.

Corner points in the image corresponding to the measured lines are determined in a second step by computing intersections of measured lines in correspondence to the 3D 'Gestalt'-model underlying the predicted model image. Through this step, line segments are obtained when two neighboring lines have been determined. These segments have a certain length and position (given by the center point on it); the corner points thus obtained are interpreted as a 3D structure approximating the generic shape, position and orientation model in a least squares sense exploiting the Jacobian matrix neighborhood relationship. Since each straight line segment is treated independently in the recursive estimation process, the closedness condition of the 3D shape will be lost during this step, in general.

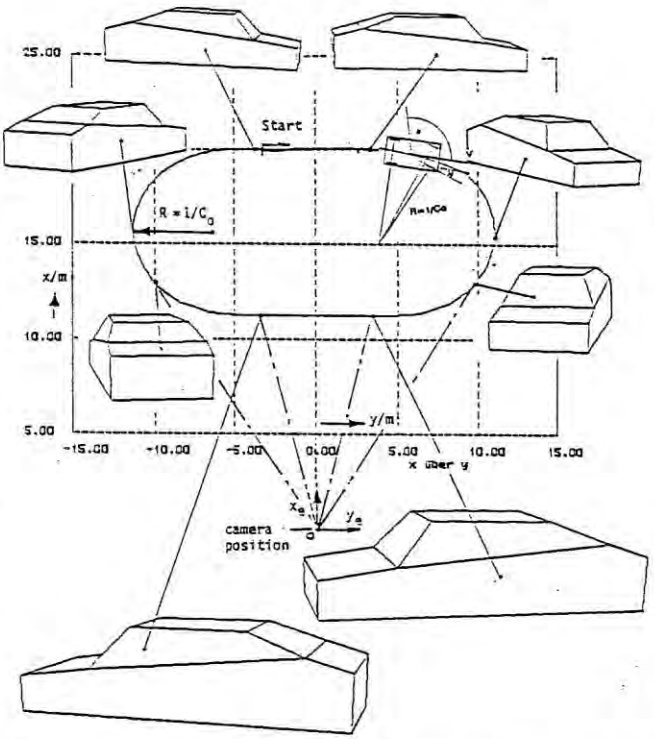
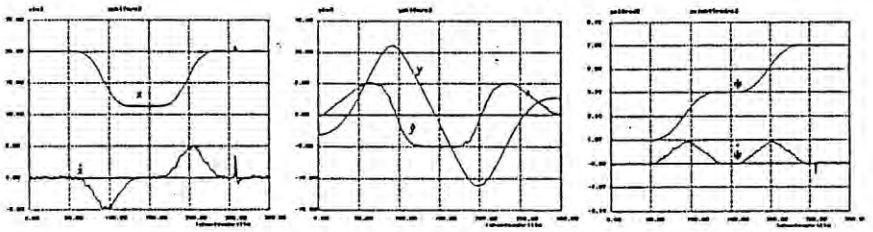
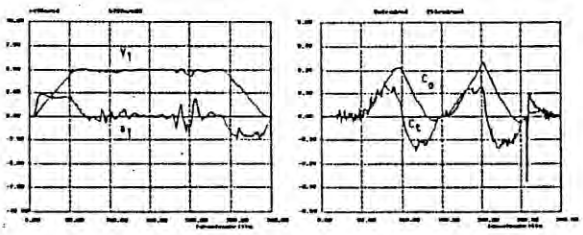


Figure 4: Oval track driven and corresponding perspective view of the idealized passenger car



a) x and $\dot{x}$ time history b) y and $\dot{y}$ time history c) ψ and $\dot{\psi}$ time history



d) Long. speed and acc. e) Curv. and temp. curv. rate

Figure 5: Estimation results of Earth oriented track and motion estimation results for the oval course

This is remedied by computing a best fit to the 3D model obtained by transition to the minimal parameter set as given in fig. 2. This step enforces symmetry and all the other characteristics coded as a priori knowledge about passenger cars into the system. Numerical experiments have validated this approach, which tries to simplify each step as far as possible by decoupling it from other ones; this feature allows to distribute computing loads onto parallel processors thereby reducing cycle times.

Trajectory estimation results

Fig. 5 shows the estimated time histories of the state variables which correspond well to the actual path driven (known from the simulation process). Speed components are more noise corrupted than positions, especially $\dot{x}$ (fig.5a); at around cycle 260 an outlier occurred which was due to some error in the visibility calculations, as it turned out later. It is seen to disturb all estimated values considerably. By modelling accelerations also and by limiting its maximal values, outliers like this one could be eliminated. Experiments with real scenes will have to show whether this is necessary. The y-components in fig. 5b nicely show the acceleration phase at the beginning (cycles 0 to 50) and the deceleration phase at the end (240 to 290). Fig. 5c shows the good recovery of the linear yaw rates over time corresponding to the constant positive and negative steer rates during the clothoid arcs.

In figure 5 the results of the second, object-oriented trajectory estimation step fed by the speed components of the previous one are shown. Speed is estimated rather well, while longitudinal acceleration is much more noisy especially around cycle numbers 100 and 200 (see fig.5d) when the $\dot{x}$ and $\dot{\psi}$ components have their maximal magnitude (fig.5a and 5c). The acceleration step at 50 and the deceleration step input at 240 are not recovered well; a rather shallow transition results.

Fig. 5e shows that except for a small arc with negative curvature when entering the straight line segments (around cycle numbers 140 and 250), the path curvature estimation C_0 looks very reasonable, while the curvature rate C_t is rather noisy and not steplike as indicated in table 1. The outlier at cycle 260 again causes a heavy disturbance; however, such jumps could be eliminated by physical reasoning in space/time should they occur in real experiments.

Shape estimation results

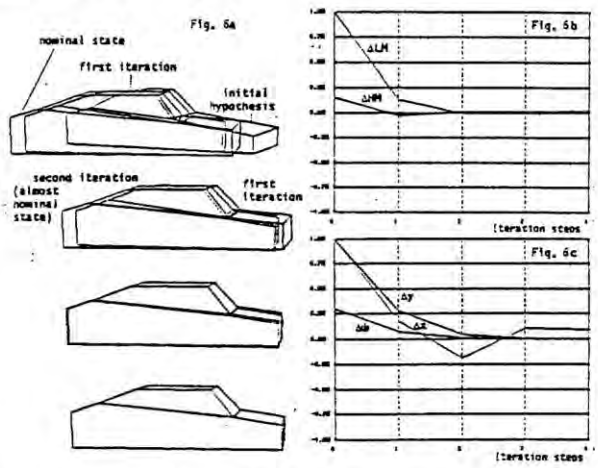
The shape parameters according to fig. 2 for the simulation results shown here are given in table 2. Convergence behavior has been studied for several cases; for the parameters indicated in table 2 the results of figs. 6 and 7 have been obtained. The initial position and orientation and the hypothesis selected was the following:

coordinate	actual state	hypothesised state
position x	20.000 m	21.000 m (+1.0 m)
position y	-3.125 m	-2.125 m (+1.0 m)
orientation ψ	1.57 rad	1.87 rad (+0.3 rad)

Shape parameter	actual value / [m]		
length of vehicle	L	4.60	
length of motorhood	LM	1.00	(hyp.: 2.00)
length of front wind shield	LF	0.50	
length of rear wind shield	LH	0.60	
length of trunk	LK	0.80	
height of vehicle	H-RR	1.30	
height of motorhood	HM	0.15	(hyp.: 0.30)
height of front window	HF	0.60	
height of trunk	HK	0.10	
width of vehicle	B	1.80	
lateral slope	NH	0.50	
longitudinal slope	NL	0.10	

Table 2: Vehicle parameters underlying the simulation experiment; LM and HM have been disturbed by factor 2 to study convergence

The camera was positioned at 0;0 and 3m above the ground. Fig. 6a shows three perspective views of the vehicle during the first step of the estimation process in an overlay: The actual state (leftmost image), the starting hypothesis with the motorhood twice as long as the actual one and the offsets corresponding to the table above (1 m in position to the right and backwards, and 17 degrees in orientation, rightmost image) and the view as 'imagined' after the first iteration (in between). The following frames are the results for the next three iteration steps; it is seen that after four iterations the estimated state and shape are practically identical to the real (simulated) ones.



a) Perspective views b) Shape parameter deviations
c) State deviations

Figure 6: Simultaneous vehicle state and shape parameter estimation

Fig. 6b shows the discrepancy reduction for the two shape parameters as function of the iteration number; in fig. 6c the same is done for the state variables. Except for the depth-component x which is still 0.1 m off, convergence practically occurs within three steps.

In fig. 7 the iteration results for the shape parameters are given while the vehicle drives along the oval course.

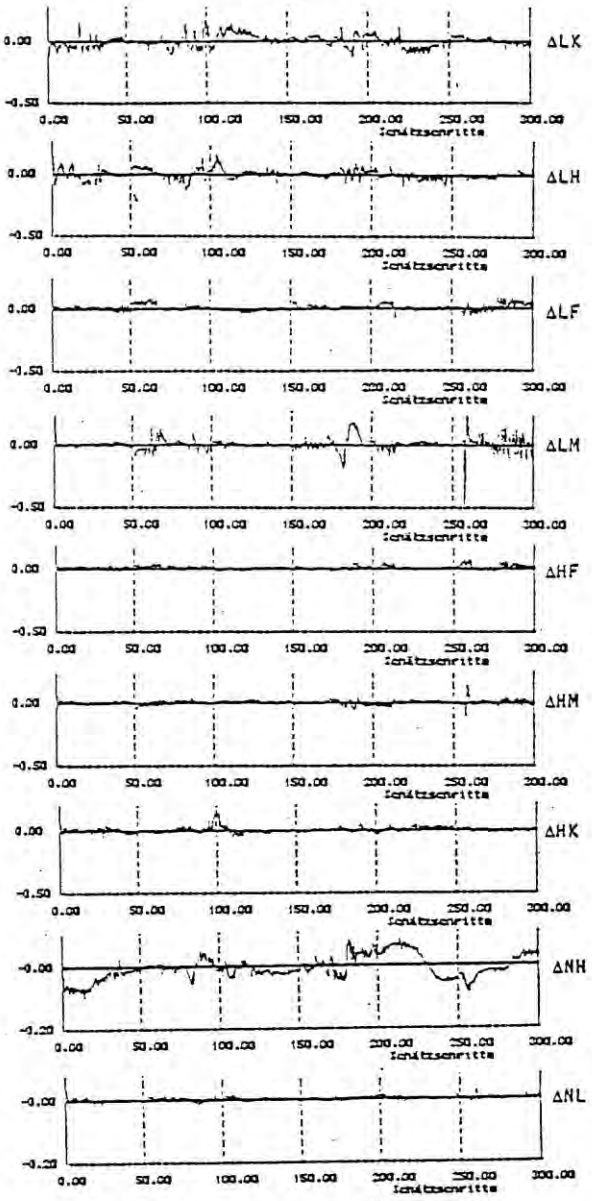


Figure 7: Shape parameter fluctuations along the oval course of fig. 4

It is seen how the measurement noise (assumed to be two pel) affects the shape parameter estimation differently according to the aspect conditions. The parameters, however, remain close to their true values in the average. This means, that low pass filtered results after some time may be frozen as the best shape estimate available; from this time on, a much simpler pure state estimation process may be sufficient for tracking the vehicle with this set of shape parameters (which may even be reduced to a still simpler one).

8 Conclusions

It has been shown that recursive estimation in the 4D approach to dynamic machine vision - using both generic 3D shape and motion models - can be put to work to estimate shape and motion of road vehicles concurrently; processing times with today's microprocessors (e.g. Intel 80386 with floating point numerical coprocessor) are less than 1 second for one iteration cycle. Distributing the processes onto parallel processors will allow to achieve update rates of several Hz which are desirable in real world traffic situations.

The approach still has to be tested under more general conditions in order to determine its robustness. However, from first results it seems to be very promising. For more accurate shape representations for real road vehicles, the lines of the polygons will have to be replaced by curves of smoothly changing curvature; this work is under progress.

Working with integral spatio-temporal models right from the outset allows to perform image sequence processing very efficiently, confining the analysis to the last image of the sequence only.

*This research has been supported by the German Science Foundation DFG.

9 Literature

[1] G.J. Bierman: Factorization Methods for Descrete Sequential Estimation. AP, New York, 1977.

[2] H.G. Meissner, E.D. Dickmanns: Control of an Unstable Plant by Computer Vision. In: Huang T.S. (ed) Image Sequence Processing and Dynamic Scene Analysis. Springer-Verlag, Berlin, pp 532-548.

[3] E.D. Dickmanns, V. Graefe: a) Dynamic monocular machine vision. Machine Vision and Applications, Springer International, Vol. 1, pp 223-240, 1988. b) Applications of dynamic monocular machine vision (ibid), pp 241-261, 1988.

[4] E.D. Dickmanns, Th. Christians: Relative 3D-state estimation for autonomous visual guidance of road vehicles. In: Kanada, T.e.a. (ed): 'Intelligent Autonomous Systems 2', Amsterdam, Vol. 2, Dec. 1989, pp 683-693.

[5] E.D. Dickmanns, B. Mysliwetz: Recursive 3D Road and Relative Ego-State Recognition. Special Issue of IEEE-Trans. PAMI 'Interpretation of 3D Scenes' (to appear Sept. 1991).

[6] R.E. Kalman: A new Approach to Linear Filtering and Prediction Problems. Trans. ASME, Series D, Journal of Basic Engineering, 1960, pp 35-45.

[7] P.S. Maybeck: Stochastic models, estimation and control. Vol. 1, Academic Press, 1979.

[8] M. Wertheimer: Untersuchungen zur Lehre von der Gestalt. Psychologische Forschung, Vol. 4, 1923, pp 301-350; Engl. translation: Investigations on the Gestalt Theory, in Ellis, W.D. (ed): A Source Book of Gestalt Psychology. Harcourt, New York, 1938, pp 71-88.

[9] H.J. Wuensche: Detection and Control of Mobile Robot Motion by Real-Time Computer Vision. In: Marquino N. (ed) Advances in Intelligent Robotics Systems. Proceedings of the SPIE, Vol. 727, Cambridge, Mass., pp 100-109, 1986.