4-th International Symposium on Robotics Research (ISRR)
University of California, Santa Cruz, Aug 9-14, 1987

4D-dynamic scene analysis with integral spatio-temporal models

E. D. Dickmanns[1]
Aerospace Technology Department
Universität der Bundeswehr München
W. Heisenberg Weg 39, 8014 Neubiberg
Federal Republic of Germany

Abstract

A method for interpreting high frequency image sequences is presented that confines image data processing to the last image of the sequence and yet - by using smoothing integration-operations - allows to determine velocity components in space explicitly. This is achieved by simultaneously exploiting 3D-object and - motion models together with the laws of perspective projection. These integral models of the process observed in the world and of the (very flexible) measurement process using vision are utilized in the sense of modelbased feedback control theory (Kalman filter, observer), to estimate the state variables in space and time directly. The concept has been tested in three application areas: a planar docking maneuver between two 3D-objects, autonomous road vehicle guidance and autonomous aircraft landing.

Introduction

The usual way to process image sequences today is characterized by pictorial interpretation of single images and an ensuing comparison of the position of objects; from this the motion of objects in space is reconstructed. This procedure may be based in the historical development of digital image processing which began in the area of remote sensing taking temporally well separated single images.

It is, however, well known from biology and physiology, that pictorial vision and motion vision are two separate developments, motion vision being the phylogenetically elder one. The psychologist Yonas has shown, that also in human children motion vision is developed first <Yonas 83>.

If a slide show, say on the last vacation adventures, is copied onto a movie film and shown at normal image frequency (18-24 frames per second) the observer will turn away or close the eyes since a continuous development of action is missing. From this one can conclude that for meaningful vision, temporal continuity is an essential prerequisite. High image frequency is not detrimental since it does not alter the dynamics of the process being observed; it is, on the contrary, beneficial since it reduces the so-called correspondence problem: Within the small sampling period T, which at the usual TV-frame rate is 16 2/3 ms, features being tracked will move only by a small amount. When the process being watched is "recognized", reasonably good extrapolations to the next frame can be achieved by linear prediction with a temporal model.

If one then succeeds, by exploiting the difference between the predicted and the actually measured feature position, in determining the parameters and the state variables of the model, which served as the basis for "recognition", and in servoing these variables fast and precisely enough so that the measured values are well approximated over time, then a symbolic representation of the process in the world has been generated in the computer. This stable condition is called "recognition" of the process by computer vision or "understanding" of the dynamic scene.

It is immediately clear that because of the temporal extrapolation required, time has to be an essential component of the model. In order to achieve this, the dynamical models of modern control theory are utilized which have been developed around 1960 in the form of linearized systems of differential equations or difference equations in the discrete linear state space model for sampled control systems <Kalman 60; Kailath 80>.

This model based approach eliminates the necessity to have access to data of previous images (in order e.g. to compute differences or optical flow) and it thereby relieves computational loads considerably. This has to be paid for by having to deal with a somewhat more involved initial orientation phase when the vision process starts and when reasonable model hypotheses have to be found. However, besides confining image processing to the last image of the sequence, this approach has several additional advantages hardly to be overestimated:

+ With respect to actions required, gaps in measurement data may be bridged over a certain period in time by just using the extrapolated state variables of the model.

+ The quality of new measurement data may be judged relative to the values predicted by the model; depending on the quality of the model, the dynamics of the process and preknowledge about general environmental conditions (e.g. perturbations) a situation-specific reaction is possible: from throwing away the new data up to the initiation of a new subprocess in order to obtain a more precise analysis of the situation. In addition, the dynamical model allows the application of adapted filter algorithms for data smoothing.

+ The interpretation of the image sequence proceeds simultaneously in space and time. Meaningfull continuity conditions for features are formulated easier in 3D-space than in the 2D-image, e.g. the disappearance or appearance of features when aspect angles change or when occlusion occurs due to (spatial) object- or ego-motion.

+ Well proven spatiotemporal models allow the prediction of events or the appearance of objects, features of which can be actively looked for; in this case good hypotheses for interpretation and parameter adaptation are readily available. (It is easier to orient oneself in a "known environment" than in an unknown one. Whole objects may be recognized by detecting only a few characteristic features.)

Coming from systems dynamics this approach to image sequence interpretation is readily proposed; it does not seem to have been investigated more closely up to now, however. Except for our group at UniBw M, where dynamical models for image sequence processing have been used since 1979 <Meissner 82>, only a few hints are found in the literature to similar approaches <Gennery 81>, <Rives 86>, <Broida, Chellappa 86>. We have tested this approach at four motion control applications:

1. balance of an inverted pendulum on an electro cart <Meissner, Dickmanns 83>,<Dickmanns, Wünsche 86a>

2. autonomous road vehicle guidance <Dickmanns, Zapp 85, 86, 87>

3. planar docking maneuver between 3D-objects with a model control plant <Wünsche 86, 87>, <Dickmanns, Wünsche 86b>

4. autonomous aircraft landing (simulation) <Eberl 87>, <Dickmanns, Eberl 87>

In all four applications real-time motion control has been achieved with a real (CCD-) TV-camera and a MIMD-multi-microprocessor system for image sequence processing <Graefe 84> in the loop.

In the sequel, the approach is first described in general terms; then the applications 2 to 4 are discussed in somewhat more detail as an introduction to the references cited. Finally, an outlook is given on the growth potential up to the recognition and visual tracking of other moving objects under egomotion.

2. Integral spatio-temporal models

Figure 1 shows a juxtaposition of the usual procedure in image sequence processing (upper half) and the "cybernetic" approach based on "difference feedback" (lower half). Spatiotemporal processes in the real world (1) are imaged by a TV-camera, usually via a sequential analog video signal into an image sequence (left side). In conventional image sequence interpretation for the detection of motion two or more frames out of the sequence have to be accessed simultaneously in order to find corresponding features (corners, contours or lines) and to obtain displacement vectors in the image plane by differencing. Knowing the sampling period, the velocity components in image coordinates can theoretically be determined from this (optical flow). Due to the inevitable measurement noise, these computed velocities become the more corrupted by noise the shorter the sampling period between the two frames evaluated is (the well known roughening property of differentiation for noisy data). Based on these position and velocity data in image coordinates, one then tries to infer the imaged spatiotemporal processes (3D-motion); in this step the nonunique inversion of the perspective projection has to be performed. There are many publications to this ill-conditioned problem.

In the model-based approach, on the contrary, through a successfull recognition process over time there results a symbolic spatiotemporal model-instantiation in the computer of the dynamical scene being observed (in fig. 1 termed "world 2" (right) in accordance with <Popper 77>). The spatial symbolic representation may be done using known methods in computer vision or computer graphics; the temporal symbolic representation is realized by differential equations for the spatial position, orientation and velocity-components as state variables of the object, e.g.

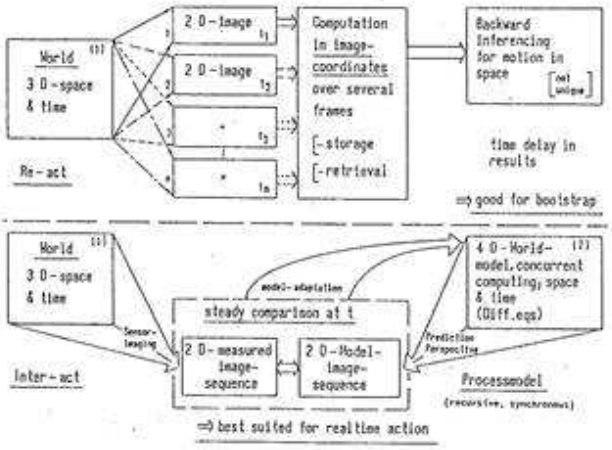

Figure 1: Two basically different ways of image sequence processing in computer vision: difference computation between images (top), model based integration (bottom).

center of gravity (c.g.) coordinates and angular orientation around the c.g..

Very often it may be sufficient to take linear approximations to the differential equations with time- or state-dependent coefficients; in a discrete formulation with respect to time and the basic sampling period, a system of difference equations results

$$x[(k+1)T] = A(k,T)\ x\ (kT) + B(k,T)\ u(kT) + v(kT) \quad (1)$$

where x is an n-component state vector, k is the running index of the discrete time, T the sampling period, A the n*n transition matrix, B the n*m control effect matrix, u an m-component control vector and v a superimposed disturbance.

Equation (1) without the disturbance term allows to determine the state vector at time (k+1)T as extrapolation based on this model by simple matrix-vector multiplication (nominal prediction). With this 3D-state the perspective projection (3D to 2D forward) is applied to the 3D-shape features of the object model, in order to arrive at predicted positions of 2D- features to be measured in the image. This leads to an impoverished "model image", the components of which are compared to the measured actual image. The computer, therefore, needs to have access to the last image in the sequence only. In the sense of modern feedback control theory (see e.g. <Kalman 60; Kailath 80>) an additional prediction error term may now be fed back through a (yet to be determined) gain matrix in such a way that the discrepancies vanish over time. Depending on the noise statistics of the process and the measurements, either filter- or observer-techniques may be selected.

Note, that this formulation contains the state variables in 3D-space as primary variables and that all components are reconstructed or estimated, also when a smaller number of output variables is being measured (observability assumed as given). The numerical operations required are integrations (summations in the discrete case), which tend to suppress noise effects; if the selectable error decay dynamics are chosen properly (observer-eigenvalues slightly larger than those of the plant), the resulting gain factors lead to an acceptable behaviour of this cybernetic vision process, provided the sampling period is small, compared to the characteristic time scale of the process being watched, and the model is sufficiently good.

The sequence of image comparisons (lower middle in fig. 1) thus leads to an adaptation of the model and of the state variables, converging over time towards the process running in reality. Within the computing process, thus, the dynamical scene analysed is duplicated in a symbolic form by servocontrolled instantiations of elements out of a store of components for a world model. The state variables of the dynamical model are obtained explicitly as complete time histories; they are taken instead of the state variables in the real world 1 as the basis for decisions with respect to actions, e.g. control activities. - There are interesting parallels to old philosophical ideas.[2]

A detailed treatment of special applications is not possible in the framework of this survey paper; for this, the reader is referred to the original publications cited. The following treatment of the applications investigated is intended to help clarify the general principle.

3. Automous guidance of road vehicles

Forced by gravity and the supportance of the ground, road vehicles move parallel to the local Earth surface essentially. In order to improve riding comfort, man has shaped the areas for vehicle movement with a smooth surface (roads); i.e. only radii of curvature are allowed that are large as compared to the wheel- or axle-distance. The curvatures of these "surface-strips", both in the vertical plane defined by the gravity vector and the road tangent, and in the plane tangential to the surface determine the driving behaviour of vehicles. To recognize both of those is one of the essential tasks for vehicle guidance by both man and computer using vision. This has to be achieved in a certain look ahead range concurrently while driving.

In a coordinate frame fixed to the vehicle when moving along the road, the geographically fixed road curvature appears in the car as a temporally variable state of the environment. Since the law, according to which highways are built (clotoids, i.e. linearly varying curvature over arc length), and the ranges for the parameters yielding reasonable results are known, effective filtering methods may be based on this road model for smoothing noisy image processing data <Dickmanns, Zapp 86; 87>. These methods evaluate recursively, by exploiting the dynamical model for the road being driven at speed V, the two relevant state variables in each of the two planes mentioned above: 1. the actual curvature and 2. the rate of curvature change with arc length (differential geometry road skeleton model). The image of the road in a look ahead distance, however, also depends on the position and the orientation of the camera relative to the road. If the camera position and its orientation in the vehicle are fixed, then the state variables of the vehicle relative to the road (lateral position y and heading angleψ) determine its perspective image. This holds true spatially. Temporal continuity conditions result from the vehicle having only limited mobility: Its wheels revolve in a plane normal to their axis; the sliding angle β due to soft tires and slipping are small, but not negligeable. From this, side constraints in the form of differential equations for vehicle motion result: If the vehicle does have an angle relative to the road not equal to zero, there will be resulting a lateral offset y in the future; this, in addition, depends on the centrifugal force ($\sim V^2$) and the steering control. Introducing the knowledge of these interactions into the process of image sequence interpretation, again a very effective recursive approach for estimating the entire state vector of the vehicle results. Though only some feature positions are being measured, all position and speed components are determined exploiting always the last image of the sequence only <Dickmanns, Zapp 85, 86, 87>. Fig. 2 shows the cooperation of the two dynamical models for curvature determination (upper part) and for vehicle state estimation (lower part) in the feedback loop.

Road curvature c determined in the look ahead range is not only used for driving the anticipatory part of the lateral control $u_{a\ lat}$, but also for automatically adapting longitudinal speed V. This is adjusted in such a way that the lateral acceleration $a_y = cV^2$ stays below a preset limit value (e.g. 0.1 of Earth gravity g). Both in simulations with real sensors in the loop and in

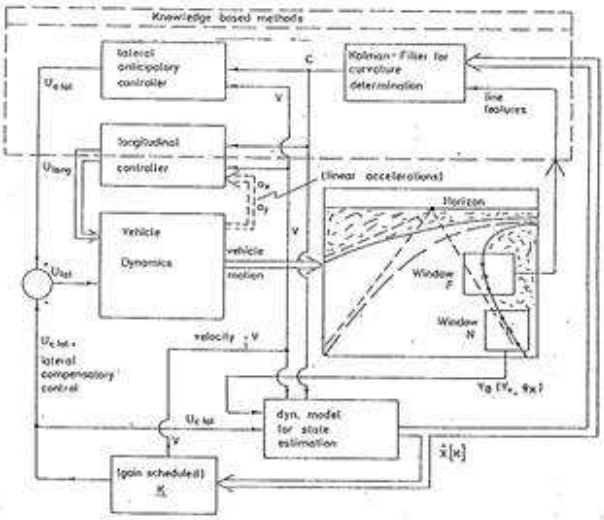

Figure 2: Block diagram for high speed road vehicle guidance by computer vision: model based estimation of curvature (upper right) and vehicle state relative to the road (lower right).

real experiments with our 5-ton test vehicle for autonomous mobility and computer vision, VaMoRs (fig. 3), this method has proven to be very efficient.

Figure 3: The UniBw M test vehicle for autonomous mobility and computer vision VaMoRs.

In fully automatic test runs speeds up to 60 km/h have been achieved, where the vehicle adapted speed to the curvature of the track automatically. One of these test tasks is shown in figure 4. Image sequence evaluation- and control cycle time has been 0.08 to 0.1 seconds.

Other vehicles as partners in road traffic may be observed and tracked using very similar methods and the same camera (see below).

4. Planar relative positioning

A frequent task in robotics is to position a controllable three-dimensional vehicle relative to another 3D-object. Using the dynamic approach to computer

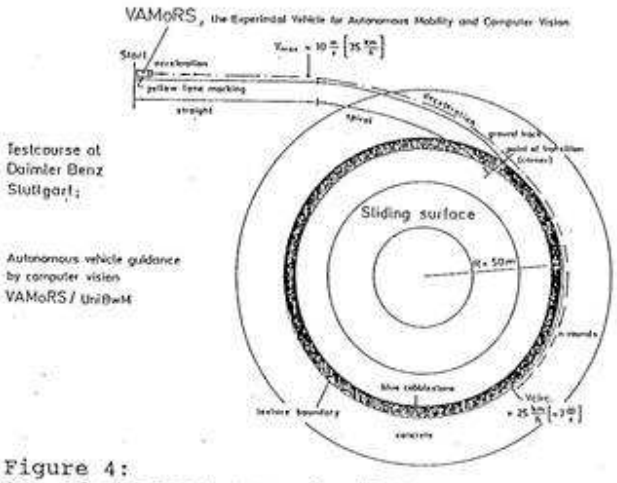

Figure 4:
One of the test runs for VaMors.

vision described above, H.J. Wünsche developed several important implementational details and demonstrated its performance and efficiency in fully autonomous docking maneuvers with a model control plant in the laboratory <Wünsche 87>. Fig. 5 shows the aircushion vehicle with a computer controlled reaction jet propulsion system on a planar table together with several docking partners. The convex prismatic shapes of the bodies are assumed to be known. They are represented in the computer by wire frame models. The

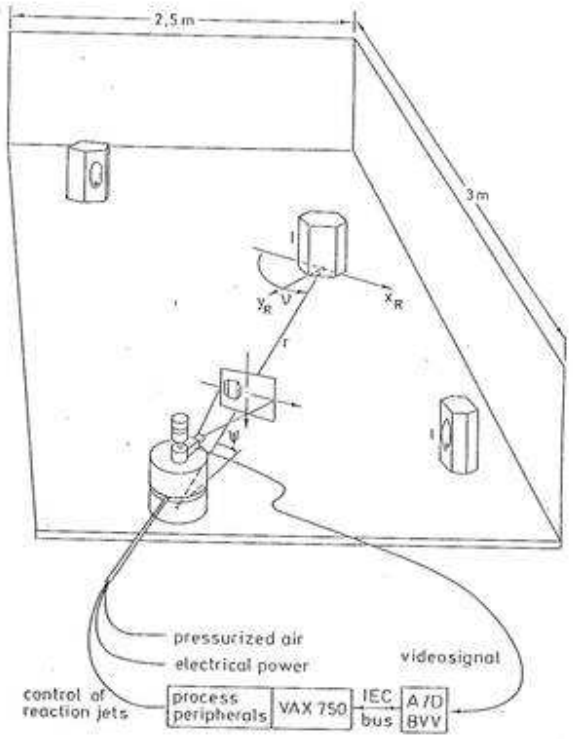

Figure 5: Satellite model plant for visually controlled docking.

controlled vehicle carrying a CCD-TV-camera has the task to recognize its docking partner, drive towards it, recognize the precise position of the docking port, position itself axially relative to it and finally to maneuver towards the partner until mechanical lock-in is achieved. During this maneuver relative position and velocity have to be evaluated steadily in order to guarantee safe process control. Partial occlusions of the target body over a finite period of time may occur but are not allowed to disrupt the docking procedure.

In <Wünsche 86; 87; Dickmanns, Wünsche 86b)> application specific details are given. Corners in the perspective projection are chosen as features to be tracked; the complete state vector of the vehicle relative to the docking partner is estimated recursively by tracking a varying number of these features in the monocular image sequence over time. The state vector consists of two translatory positions and speed components and one angular orientation and rate; together with the camera pitch angle and a rotatory disturbance acceleration, eight state components are steadily estimated.

Kalman filter techniques in a sequential stabilized formulation are used based on tracking up to four feature positions. By continuously checking a performance index, those features are automatically selected which yield the best estimation results. Occlusions due to changes in aspect angle are predicted and feature tracking is redirected autonomously. If a sudden occlusion of a feature by an unexpected object occurs, after a short period of repeated trials at the old feature position a new combination of features is selected yielding the next best performance index. Due to the strong perspective distortion of imaged features at small distances, the algorithms have to be tolerant against changes in feature shape. Fig. 6 shows a block diagram form of fig. 1 (lower part) for this application. In figure 7 results of a test run over 90 seconds are given, in which the vehicle first turns towards the docking partner (ψ-reduction for $t<9s$, first row) then moves towards it (R-reduction for $13<t<20s$, 5-th row), circumnavigates it at constant R (for $20<t<60s$, last two rows, ν being the polar angle and VT the tangential velocity component) and finally closes in for docking ($t>60s$, see R and VR).

The approach developed in this application for 3D-object recognition and tracking is generally applicable and is presently being transferred to the task of recognizing other vehicles, distance keeping and collision avoidance in road

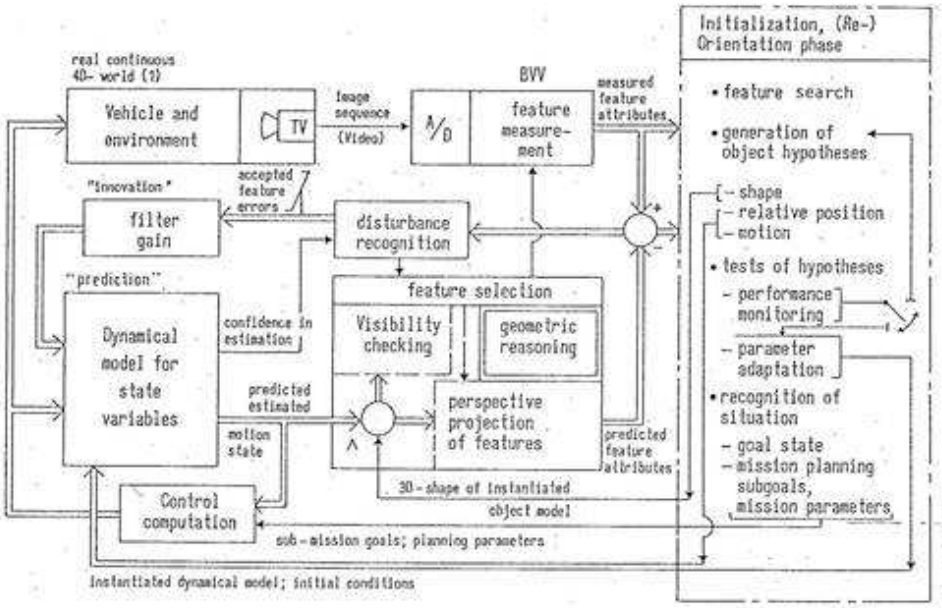

Figure 6: Block diagram for the "dynamic-" or "4D-" vision concept

scenes. Here, the object shape and its relative motion state have to be determined simultaneously. Background knowledge on road-vehicle shapes and their mobility parameters is exploited to alleviate the task of generating simultaneously both spatial and temporal interpretation hypotheses.

In this case, both shape- and dynamical model-instantiations from a long term memory model-base have to be tested against feature aggregations from the image sequence; simultaneous reasoning in space and time is required. Shape models will be based on the normalized-curvature-function (NCF) representation proposed in <Dickmanns 85>.

In order to alleviate the solution of this task, the next generation of our image sequence processing system BVV provides the capability of active (fast) control of the viewing direction (see <Mysliwetz, Dickmanns 86>) and of analyzing video data of two TV cameras in parallel (e.g. for tele- and wide angle lenses yielding images of different resolution).

5. Landing approach by computer vision

This is the most complex real-time motion control task solved by computer vision up to now. Aircraft motion occurs simultaneously in six degrees of freedom: three translatory and three rotary ones. In each degree of freedom, according to Newton's law, one differential equation of second order is required to model the dynamical behaviour, so twelve state variables are necessary to describe the rigid body motion.

An aircraft is controlled by selecting four control variable time histories: Elevator for pitch and altitude, aileron for roll, rudder for yaw and sideslip and throttle for thrustlevel control; in addition diverse flaps may be set for certain flight regimes (start and landing). To direct such a vehicle in a well controlled maneuver requires skill and concentration even for a trained human pilot; he has to acquire this capability in an extended learning process. In case that he would like to switch to a different type of aircraft, he has to submit to a type rating procedure. Here, he not only has to become acquainted with the positions of knobs and dials but he has to develop a feeling for the normal dynamic behaviour of the vehicle following certain control inputs. Typical motion sequences in time are essential knowledge elements with respect to his task of reacting correctly in given situations.

Exactly this knowledge, coded in differential equations as side constraints to the development of trajectories, should be available to an automatic system for recognizing and controlling landing approaches by computer vision. Though it is possible to obtain the actual position and relative orientation of a camera (up to a scaling factor) from a single image containing the rectangular landing strip and the horizon, it is hard to extract velocity components from a sequence of noise corrupted images by differencing in the conventional manner, especially when control activation is based on these noisy data.

Simultaneously exploiting spatial and temporal models as shown in section 2

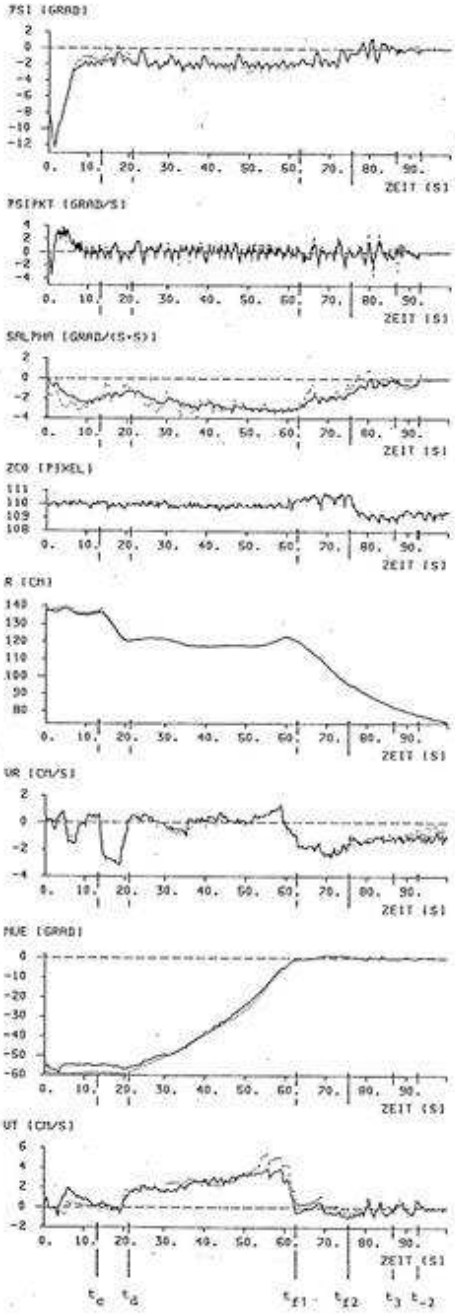

Figure 7: Time histories of state variables for a docking maneuver by computer vision in real time (from <Wünsche 87>).

and figure 6, G. Eberl has shown that the problem of controlling landing approaches by visual feedback to the computer may be tackled successfully relying on present day microprocessors <Eberl 87>. In a six-degree-of freedom simulation with real-time image sequence processing hardware in the loop, complete landings starting from 2 km distance have been performed fully autonomously with aerodynamic speed V being the only quantity not determined from vision. Twelve state variables and four control time histories have to be determined depending on the perspective distortions and its rates in the image of a rectangular planar landing strip. The problem has been solved using three cooperating Kalman filters (of sixth, fifth and second order) on a Perkin Elmer computer PE 3252 which was able to compute both the motion simulation and the state- and control-evaluation in 100 ms cycle time (fig. 8). It seems unlikely that such a complex task can be handled by computer vision without using integrated spatio-temporal process models.

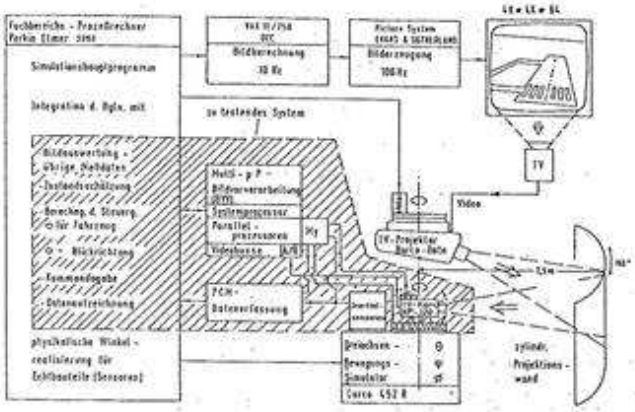

Figure 8: Hardware in the loop simulation for automatic visual landing approach of an aircraft

6. Conclusions

By unifying dynamical models, 3D-shape representation and (forward) perspective projection and by using this integrated model in the sense of observer-/filter-methods of modern control theory, a formulation of the recognition process simultaneously in space and time has been achieved. Though only data of the last image of a sequence are being processed, by exploiting spatio-temporal models of the process in the scene being watched, also speed components in 3D-space can be determined using smoothing integration (summation) operations over time. The inversion of the perspective projection is bypassed by extrapolating the motion-state with the dynamical model over time in order to obtain the model-state variables at the next measurement point in time. For this state the features to be evaluated in the real image are computed in the "impoverished model image" by forward projection. The measured differences between these feature positions and the actually measured ones are used to improve the model state in space and time (velocity components) directly; only the partial derivative matrix of the perspective projection equations with respect to the

state variables of the dynamical model is needed here; the error decay rate may be selected fixing the corresponding feedback gain matrix as known from modern control theory.

This approach offers a sufficiently rich imbedding for the interpretation process in space and time and is computationally efficient. The models may be time-varying (e.g. vehicle dynamics as function of speed); the method even then works reasonably well provided the model rate of change is slow as compared to the image frequency. The results achieved up to now in four application areas are encouraging; they have been obtained with sampling rates from 8 to 25 Hz. Development steps are being done presently towards the capability of handling more complex situations where (several) objects of unknown shape may occur having unknown dynamical models.

[1]Dr.-Ing., Prof. for Control Engineering
This research has been partially supported by BMFT and DFG.

Literature:

<Broida, Chellappa 86> T.J. Broida, R. Chellappa: Estimation of Object Motion Parameters from Noisy Images. IEEE Trans. PAMI Vol.8 No.1, Jan.1986, pp 90-99.

<Dickmanns 85> E.D. Dickmanns: 2D-Object recognition and representation using normalized curvature functions. In M.H. Hamza (ed.): Proc. IASTED Int. Symp. on Robotics and Automation '85, Acta Press, 1985, pp 9-13.

<Dickmanns, Zapp 85> E.D. Dickmanns, A. Zapp: Guiding Landvehicles Along Roadways by Computer Vision. Proc. Congres Automatique 1985, AFCET, Toulouse.

<Dickmanns, Zapp 86> E.D. Dickmanns, A. Zapp: A Curvature-based Scheme for Improving Road Vehicle Guidance by Computer Vision. In: "Mobile Robots", SPIE-Proc. Vol. 727, Cambridge, Mass., Oct. 1986, pp 161-168.

<Dickmanns, Wünsche 86a> E.D. Dickmanns, H.-J. Wünsche: Regelung mittels Rechnersehen. Automatisierungstechnik (at) 34, 1/1986, pp 16-22.

<Dickmanns, Wünsche 86b>: Satellite Rendezvous Maneuvers by Means of Computer Vision. Jahrestagung DGLR München, Okt. 86. Jahrbuch 1986 Bd 1, DGLR, Bonn, pp 251-259.

<Dickmanns, Zapp 87> E.D. Dickmanns, A. Zapp: Autonomous High Speed Road Vehicle Guidance by Computer Vision. Preprint 10th IFAC-Congress, München, July 1987.

<Dickmanns, Eberl 87> E.D. Dickmanns, G. Eberl: Automatischer Landeanflug durch maschinelles Sehen. (To appear in DGLR-Jahrbuch 1987) Jahrestagung der DGLR, Berlin, Oct. 1987.

<Eberl 87> G. Eberl: Automatischer Landeanflug durch Rechnersehen. Dissertation UniBw München, LRT, 1987.

<Gennery 82> D.B. Gennery: Tracking Known Three-Dimensional Objects. American Assoc. for AI, AAAI, Pittsburgh, Aug. 1982, Proc., pp 13-17.

<Graefe 84> V. Graefe: Two Multi-Processor Systems for Low-Level Real-Time Vision. In: J.M. Brady e.a. (eds.): Robotics and Artificial Intelligence, Springer-Verlag, 1984, pp 301-308.

<Kalman 60> R.E. Kalman: A new Approach to Linear Filtering and Prediction Problems. Trans. ASME, Series D, J. Basic Eng., 1960, pp 35-45.

<Kailath 80> Th. Kailath: Linear Systems. Englewood Cliffs, N.J., Prentice Hall, 1980.

<Meissner 82> H.G. Meissner: Steuerung dynamischer Systeme aufgrund bildhafter Informationen.. Dissertation HSBw München, LRT, 1982.

<Meissner, Dickmanns 83> H.G. Meissner, E.D. Dickmanns: Control of an Unstable Plant by Computer Vision. In: T.S. Huang (ed.): Image Sequence Processing and Dynamic Scene Analysis. Springer-Verlag, 1983.

<Mysliwetz, Dickmanns 86> B. Mysliwetz, E.D. Dickmanns: A Vision System with Active Gaze Control for Real-Time Interpretation of Well Structured Dynamic Scenes. Conf. On Intelligent Autonomous Systems, Amsterdam, Dec. 1986, Proc., pp

<Popper 77> K.R. Popper, J.C. Eccles: The Self and Its Brain - An Argument for Interactionism. Springer Internat., Berlin, 1977.

<Rives e.a. 86> P. Rives, E. Breuil, B. Espian: Recursive Estimation of 3D Features Using Optical Flow and Camera Motion. Conf. on Intelligent Autonomous Systems, Amsterdam, Dec. 1986, pp 522-532.

<Wünsche 86> H.-J. Wünsche: Detection and Control of Mobile Robot Motion by Real-Time Computer Vision. In: "Mobile Robots", SPIE-Proc. Vol. 727, Cambridge, Mass., Oct. 1986, pp 100-109.

<Wünsche 87> H.J. Wünsche: Erfassung und Steuerung von Bewegungen durch Rechnersehen. Dissertation UniBw München, LRT, 1987.

<Yonas 83> A. Yonas: Development Stages of Depth Perception in Human Infants. In: D. Ingle etal. (eds.): Brain Mechanisms and Spatial Vision, NATO-ASI, Lyon 1983, Springer Verlag.