AIAA-94-1243-CP

THE 4-D APPROACH TO VISUAL CONTROL OF AUTONOMOUS SYSTEMS

Ernst D. Dickmanns
Universitaet der Bundeswehr München
D-85577 Neubiberg, Germany

Abstract

Based on experience with real-time image sequence processing systems in the application areas of vehicle docking, road vehicle guidance, AGV's on the factory floor, aircraft landing approaches and of dynamical grasping of free floating objects in space with remote control from the ground including long delay times, an efficient, distributed, expectation- as well as object-based general dynamic vision system architecture has been developed. Parallelization is structured according to physical objects, the characteristics of which with respect to 3-D shape, motion behavior, visual appearance and all other significant properties in the task context are represented internally in generic form. Both differential representations for state estimation as well as behavior control, and integral ones for mission planning, mission control and monitoring are being used. References to detailed reports on all application areas mentioned are given.

Introduction

The sense of vision is ***the*** predominant source of information for intelligent motion control in biological systems; why has it been missing in technical systems almost entirely until very recently? There are at least two basic reasons: First, human visual capabilities are well developed, and similar real-time performance on the technical side requires computing capabilities not nearly available until about a decade ago; the data flow in a color video signal is of the order of magnitude 10^7 Bytes per second (10 MB/s) while clock rates of computers are between 10 and 100 MHz. Assuming 10 to 100 operations per data point (or 'picture element') in the image (this will be abbreviated in the sequel as 'pel' or 'pixel') it is immediately seen that many parallel processors are needed for real-time performance just for the image sequence processing part, let alone dynamic scene understanding, control computation and mission monitoring.

The second reason is, that -unlike in biological systems- digital image processing started from static single image evaluations as in remote sensing applications. Until the early 80ies, when researchers from the field of control engineering moved into this newly developing field of image sequence evaluation, the approach to this field has been dominated by 'quasi-static' thinking; time has been introduced through the backdoor by differencing between images and starting from so-called 'optical flow' information.

However, from linear systems theory in the field of trajectory reconstruction based on noise corrupted measurements, recursive estimation techniques have been known since the early 60ies[1] which allow to substitute knowledge about the real-world processes to be observed for missing or poor-quality data. These so-called 'dynamical models' represent temporal dependencies explicitly as side constraints for data interpretation during process evolution over time. Exploiting these constraints systematically in conjunction with spatial shape characteristics resulted in increased image sequence processing efficiency by orders of magnitude[2].

This is due to the fact that temporal predictions using the dynamical models allow to control both assignments of image regions to parallel processors and the extractions of features by special algorithms in limited search regions depending on the situation encountered; all of this is geared to objects of specific classes for which corresponding generic knowledge is represented in 'object processor groups'. This leads to efficient local communication structures and to a modular system design[3].

At UniBwM this approach has been applied to half a dozen different guidance and control tasks. It became well known in the field of visual guidance for autonomous land vehicles where surprising performance levels have been achieved with very moderate computing power. Over the last several years the approach has enjoyed increasingly widespread use worldwide; there are many variants documented in the literature and their number is increasing rapidly [4-13]

In this paper, a survey is given on the general method as it presently stands at UniBwM. The following section covers the method proper featuring the basic ideas like 4-D representation, orientation towards physical objects, expectations and prediction error feedback as well as the central role the Jacobian matrices play for the relationship between image features and object state components; with increasing numbers of sensors and of objects in the scene analysed, the management scheme of the perception system had to become capable of dealing with occlusions, partial observability due to aspect conditions and with model switching on top of the basic distinction between initialisation and tracking phases.

The applications to mobile robots discussed in a survey fashion, following this display of the method, are: rendezvous and docking (planar reaction controlled satellite docking, and spatial autonomous landing approaches of aircraft), surface vehicle guidance with local reactions (road runner including obstacle avoidance or 'intelligent cruise control') and global mission control (landmark navigation both on the factory floor and on an outside road network).

The dynamical grasping experiment during Spacelab mission D2 in May 1993 concludes this application survey; with sensors and the robot arm as effector on board the Space Shuttle Columbia, and with the computers on the ground this teleoperated ('remotely brained') system has achieved the first capture of a free-floating object in space by delayed visual feedback (5.8 seconds !).

The 4-D approach

The 4D approach to dynamic machine vision exploits dynamical models in the form of state transition and control effect matrices for sampled data systems with cycle times as multiples of the basic video cycle time (20 ms in Europe, 16 2/3 ms in the USA); the recursive state estimation methods well known in systems dynamics for smoothing both process and measurement noise are combined here with 3-D shape representations of objects through well visible edge features and with perspective projection of these spatial edge elements into the image plane. The corresponding Jacobian (sensitivity) matrix of feature positions in the image plane with respect to changes in object-state components in 3-D space is used for bypassing the ill-posed nonlinear problem of direct perspective inversion. Instead, the least squares Kalman filter algorithm for noise reduction indirectly also performs this inversion in a sufficiently accurate approximate manner, recovering the third dimension (depth) lost during perspective projection, by temporal continuity conditions in conjunction with the dynamical model used for prediction.

Due to the relatively high temporal frequency of 12.5 to 25 Hz for image sequence interpretation, the underlying linearizations of all nonlinear relationships are sufficiently good; only the last image of the sequence needs be worked with, thereby avoiding storage and retrieval problems with previous ones. This enormously alleviates the data handling problem; no optical flow needs be computed. However, the *spatial* velocity components are obtained by smoothing numerical operations.

This prediction error feedback scheme for each object leads to a servo-maintained internal representation duplicating all essential aspects of the real-world subprocesses being individually observed and analysed. The integral interpretation in 3-D space and time has led to the name '4-D approach'. Figure 1 shows the resulting block diagram for a single imaging sensor and multiple objects besides the own vehicle to be controlled.

In parallel to the real world (shown in the upper left block) with motion processes happening in 3-D space and time, an internal representation, also in 3-D space and time but limited to the most essential aspects for the actual task at hand, is built up in the interpretation process by prediction error feedback ('internally represented world' in upper right block of fig.1). There is a fundamental difference between the initialisation phase when a new object is being discovered, and the tracking phase when temporal continuity conditions yield a good guideline for understanding the evolution of the dynamical scene observed. The basic ideas will be discussed in turn.

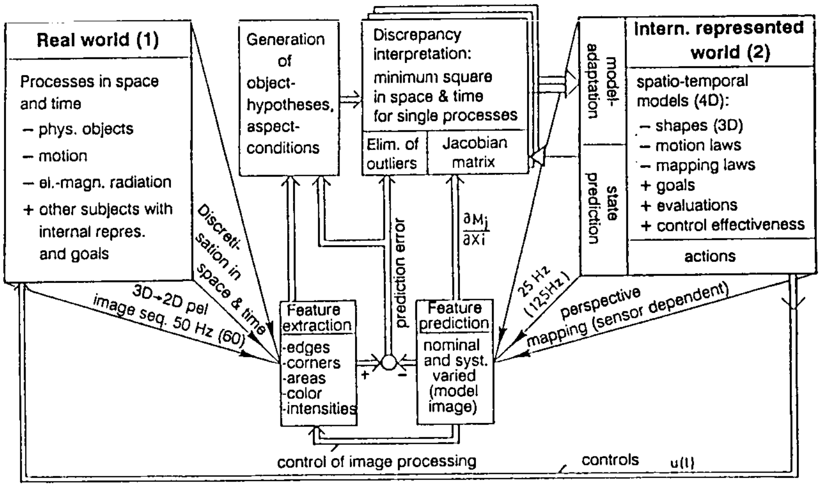

Real world (1)
Processes in space and time
– phys. objects
– motion
– el.-magn. radiation
+ other subjects with internal repres. and goals
Discreti-sation in space & time
3D→2D pel
image seq. 50 Hz (60)
Generation of object-hypotheses, aspect-conditions
Discrepancy interpretation: minimum square in space & time for single processes
Elim. of outliers
Jacobian matrix
prediction error
$\frac{\partial M_j}{\partial x_i}$
Feature extraction
-edges
-corners
-areas
-color
-intensities
+
–
Feature prediction
nominal and syst. varied (model image)
model-adaptation
state prediction
Intern. represented world (2)
spatio-temporal models (4D):
– shapes (3D)
– motion laws
– mapping laws
+ goals
+ evaluations
+ control effectiveness
actions
25 Hz (125Hz)
perspective mapping (sensor dependent)
control of image processing
controls u(t)

Fig. 1: Basic prediction error feedback scheme for single object and multiple imaging sensor

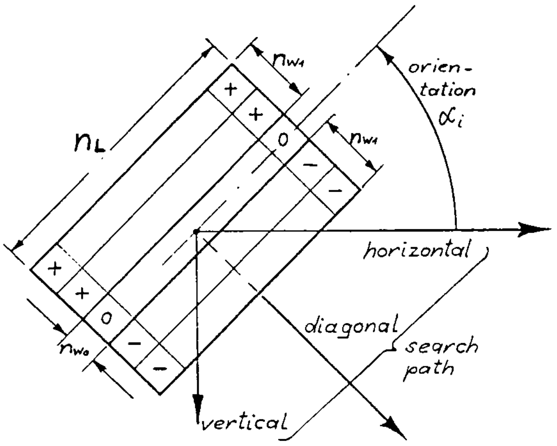
Basic ideas

Five essential elements constitute the base of the approach during continuous observation of a moving object (tracking phase):

Edge element features: The most basic element to efficient image sequence processing in a steadily changing environment is to fully exploit the difference between data and information. A uniformly grey image contains the same amount of data as a page with text and pictures from complex scenes; however, in the former case a human observer completely describes the information content by two words: 'uniformly grey', while for the latter one he may have to talk for several minutes in order to convey at least the most essential aspects. If the image would contain two differently colored areas with a curved boundary, the most economical way of capturing the information content in a symbolic description would be to formulate the boundary between the colored regions by the geometry of a line (straight segments or curves with given curvature along the arc length) and by specifying the homogeneous areas by two color symbols; again, the information exhaustively describing the image can be coded in orders of magnitude less data as compared to the pixels involved. Using tangent direction information and points of discontinuity (corners) seems to be an efficient coding scheme for boundaries in an image. Supposedly, this is the reason behind nature having developed the capability of doing just this in some of the high performance biological vision systems (striate cortex in V1).

Looking for dark-to-bright transitions (or vice-versa) irrespective of the absolute brightness level or spectral information content makes these systems less dependent on threshold values and thus more robust. Confining these tangent direction measurements to closely spaced discrete points in the image plane yields a natural discretisation allowing to construct both smooth curves from assumed linear changes of curvature along arc length in a differential geometry interpretation (corresponding to third order polynomials in Cartesian space over not too large arc lengths) and sharp corners when tangent directions are far apart even though their centers are closely spaced[14]. By these tangents (edge elements) any shape can be represented by corresponding feature groupings; in a multiple scale concept, mask operators for feature extraction can detect dark-to-bright transitions on several scales thereby allowing object detection and characterization with different resolutions; for example, for many practical purposes in vehicle guidance it is sufficient to characterize obstacles by the encasing rectangle or box. In bifocal vision with different focal lengths evaluated simultaneously, this gives an easy choice for either fast tracking or more precise shape determination.

Therefore, the family of ternary edge feature extractors as shown in fig. 2 is used as the predominant image processing tool in the context of the 4-D interpretation scheme; it has been developed over a decade of work in real-time dynamic scene understanding[15-17]. The mask parameters are dynamically controlled by the object recognition process taking the 4-D representation and perspective mapping into account, thereby realizing a fast feedback loop from high-level interpretation to low-level feature extraction; it is especially this feature which makes the tracking phase so efficient.



$$\alpha_i = i \cdot \frac{\pi}{n_{L-1}}\,;\ i = 0, \ldots, n_{L-2}\,;\quad n_L = 3, 5, 9, 17$$

Fig. 2: Operator family for edge feature extraction

Representation in 3-D space and time directly: No basic representations are performed in the image plane; at the earliest time possible it is tried to jump from a feature distribution supposed to belong to a single object to an object hypothesis in 3-D space and time. Since perspective projection is the link between spatial shape, relative orientation as well as position, and the shape in the 2-D image plane, always both object shape and aspect conditions have to be hypothesized in conjunction. Shape invariants of moving rigid bodies are in 3-D space and not in the image plane; simple motion descriptions also are more easily encountered in 3-D space than in the image plane where both motion and shape *together* yield relative image feature distribution from frame to frame. In addition, motion behavior in 3-D space may be as characteristic for an object as its shape; this leads to the third essential element:

Orientation towards physical objects: Knowledge about the real world is attached to objects which serve for structuring complex scenes. Similar properties or shapes lead to the definition of object classes characterized by generic forms and functions; other attributes may be appended depending on the task at hand (e.g. color, texture). For subjects, defined as objects with the capability of self-initiated locomotion[18], stereotypical motion characteristics may give independent cues to recognition besides static shape. In general, the centroid of features from an object yields information for translational motion, while rotational motion and shape may be derived from systematic changes of feature positions around the centroid, that is from differences between feature positions in the image.

Once an object and its motion has been recognized, the continued observation can be made much more efficient by the fourth basic element:

Expectations and prediction error feedback: The dynamical model of a process (e.g. a moving rigid body) may be given to first order by the vector difference equation

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

where x is the state vector of dimension n, k is the time index for the actual state, T is the cycle time, A is the state transition matrix $(n{\cdot}n)$, B the control effect matrix $(n{\cdot}r)$, u the r-vector of control variables, and v represents process noise with covariance matrix Q. Predictions, of course, are made disregarding the noise term. After prediction the time index k is increased by 1 and from the predicted state x^* in combination with the shape description the features to be measured in the next image are obtained from applying the forward perspective projection equations; for this purpose, first, the spatial positions and orientations of edge elements have to be computed by combining position and angular orientation of the body-fixed coordinate system having its origin at the object center with the shape description in these coordinates.

Then, the perspective mapping equations containing all translations and rotations between the body-fixed and the camera coordinate system (in x^*) as well as the camera parameters p have to be applied to all well visible edge features in order to obtain the predicted (horizontal and vertical) feature positions in the image:

$$y^* = h(x^*,p)\,, \qquad (2)$$

where dim. (y) depends on kind (edge or corner) and number of features. It is assumed that the error between predicted (y^*) and actually measured feature positions (y) is so small that a linear approximation to eq.(2) captures the essential part of the dependencies between y and x:

$$del\_y = y - y^* = dh(x^*,p)/dx^*{\cdot}(x - x^*) = C{\cdot}del\_x, \quad (3)$$

where C is the Jacobian matrix of all first order partial derivatives linking state component changes to feature shifts in the image plane. Because of the richness in information contained in this matrix and the central role it plays in the 4-D approach, it will be discussed below as the fifth basic element.

The actual measurement data y from feature extraction will be corrupted by measurement noise w (both from the video signal and from image processing); this noise is assumed to be unbiased and white with covariance matrix R so that the measurement model may be written

$$y = h(x,p) + w. \qquad (4)$$

In order to adjust the internal 4-D representation to the process being observed in the real world, prediction error feedback is used according to the recursive estimation techniques[19] derived from the Kalman filter[1] and its extensions[21]. The new best estimate for the relative object state $\hat{x}$ is obtained by adding to each predicted state component weighted elements depending on the measured prediction error; the weights are determined by the so-called Kalman gain matrix K (or its equivalents in the sequential scheme discussed below) taking the noise characteristics Q and R (confidence in both the underlying process model and the measurements) into account:

$$\hat{x} = x^* + K{\cdot}(y - y^*). \qquad (5)$$

Note that no special provision is made for perspective inversion; the least squares core of the algorithmic procedure for computing the elements of K takes care of perspective inversion hidden in the prediction step and the Jacobian matrix C. Gain computation is not detailed here for brevity; because of occlusions, varying aspect conditions and perturbations in the imaging process, the length of the measurement vector will change steadily. In order to accomodate this easily, the measurement update is made sequentially for each component; this also saves computing time and has been a standard feature of the 4-D approach from the beginning[21].

From this possibility it can be seen immediately that an update of all state components can be made from just one single measurement input; this may look like magic for people grounded in direct perspective inversion. Though this capability is true - substituting knowledge about the real process for missing data -, too few measurements over an extended period of time will lead to drifts in some state components poorly observable from this measurement, or due to model errors as compared to the actual process observed. However, in spite of this fact the value of this property based on the 4-D model can hardly be overestimated for bridging short periods with insufficient measurements for what cause soever. Even periods without any measurements may be bridged by pure predictions (vanishing second term on right hand side of eq. (5)).

An important point resulting from prediction is the capability to efficiently direct image processing by confining attention to smaller subareas of the image, and by providing information on which algorithms may be most economical in the next image (e.g. mask orientation for edge element extraction).

Central role of the Jacobian matrix: Wünsche[21] developed methods for exploiting the entries into the Jacobian matrix for other purposes beyond simple recursive state estimation. When computing power is limited, there usually are many more features available for extraction and state estimation than can be handled by the system available or from a price / performance point of view; in this situation it is very essential to be able to automatically select the most rewarding set of features yielding best estimation results in limited time. The entries into the Jacobian matrix, balanced for poorly compatible units for the state variables (like positions in meters and angles in degrees), allow to concentrate computing power for image evaluation in those areas of the image and onto those features by which best accuracy is achieved.

Small values of elements in the balanced Jacobian indicate that the corresponding state component hardly effects the corresponding feature position; if an entire column has small values this means that the corresponding state component hardly affects any measurement quantity; therefore, it can not be expected that this state component can be recovered accurately from the values measured. Likewise, if an entire row has small entries this feature is almost constant and independent of state changes; this feature is not well suited for updating the state vector and may well be eliminated from further processing.

For example, for a rectangular box looked at along a center line almost parallel to four edges (so that the image is a rectangle) it is not possible to recover the exact viewing angle because of the cosine-effect involved. If the size of the box (width, height and length) has to be determined in addition to the relative state, the dimension in viewing direction, of course, cannot be recovered; if the box is turned by 90° this size component is well determinable now while another one , now pointing in viewing direction, can no more be iterated. In the general case, the actual aspect conditions determine which state components or shape parameters can be observed and which ones have to be frozen until the corresponding entries in the Jacobian matrix become large enough again. This determines the perception strategy and -management.

In summary it can be stated that the efficiency in image sequence understanding by the 4-D approach is due to the frequent bottom-up and top-down traversal of the representation hierarchy; this is done each cycle of rather short duration so that the correspondence problem is not too hard. The linear differential models allow to tap well proven system theory. However, this only works for the continuous tracking phase.

Initialisation versus tracking

It is almost impossible to say something meaningful in general to the initialisation problem since it depends very much on the task domain and on the knowledge available to the system. For this reason, the tracking phase for specific task domains has been developed first; as expected, once this capability has been available to a certain extend, it turned out to be relatively easy to jump from feature aggregations discovered in an initial search phase (with very low cycle times) to an object hypothesis for which the tracking capabilities are given[22]. If stable tracking can be established and the prediction errors converge below certain thresholds it is claimed that a motion process involving an object of the class instantiated has been discovered; if this is not the case the hypothesis is rejected and a new one has to be tried until the set available is exhausted. Surprisingly many cases can be handled successfully by this procedure[23,24]; however, many unresolved problems remain, especially when occlusions are involved.

Again, for certain task domains like vehicle recognition on highways, many of these problem situations have been resolved by creating the capability to recognize partially occluded objects, even those appearing from full occlusion like in lane changes of one of two vehicles in front[25]; since only partial information is accessible in these cases, model based recognition involving knowledge about normal sizes of objects, about part hierarchies and about likely motion states is essential.

One often hears the call for more systematic bottom-up hypothesis generation in the initialisation phase; this, of course, would be nice to have. However, considering the discrepancy in computing power required for this type of initialisation in a somewhat complex realistic scene, and the one needed lateron for the tracking phase it is conjectured that - even in the long run - the approach of jumping to (maybe several parallel) hypotheses relatively early and then do a critical evaluation over time exploiting 4-D models may be a sensible way to go. More experience in several task domains is necessary in order to answer this question in a solid way.

The 4-D solution for complex tasks

The basic principles discussed above for the single sensor, case carry over to multiple objects and multi-sensor systems[23,24]. A general scheme for these types of complex real-time systems is given in fig.3.

As in the single object, single sensor case there is just one unifying mental representation for recognizing the situation and for controlling action; however, for each object observed there is a specific process (presently implemented on a dedicated group of processors) with access to a corresponding knowledge base for this object class. In this knowledge base generic background knowledge is stored; besides physical properties with respect to motion (the elements of the dynamical model) specific properties with respect to the different measurement processes are stored.

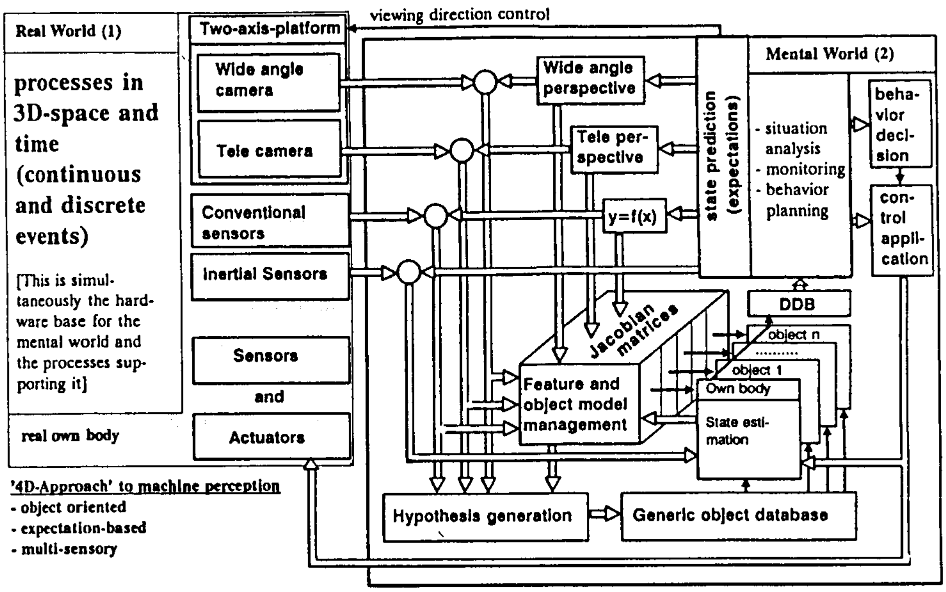


Fig. 3: Multi-sensor, multi object recognition scheme by prediction error feedback

The figure shows provision for four types of sensors, two imaging ones on a pan and tilt viewing direction platform, and two other sets, one for inertial data about the egomotion (lowest group in center) and one for other conventional sensors like odometer, tachometer, steering angle, throttle setting, brake pressure and range sensors or the like.

The inertial sensors may directly feed dynamical models for separating egomotion from visually observed relative motion with respect to another object. Otherwise, vision does allow this separation only when besides this object also a third, static one can be observed simultaneously; accuracy and robustness may be much poorer without inertial data.

For measurement signal interpretation, a Jacobian matrix has to be computed for each pair of object and sensor; in figure 3 this is indicated by the vertical arrows to the diagonal block in the lower right center.

Contrary to the conventional sensors, image sequence evaluation may yield measurement data on several objects in parallel; in fact, one of the difficulties in machine vision is the assignment problem of features extracted to objects in the scene. Grouping image sequence processing according to objects observed, as in the 4-D approach, therefore, requires a perception management subsystem shown also in the lower right center. This is an area of actual research and development; good solutions to this problem will be crucial for high performance machine perception systems.

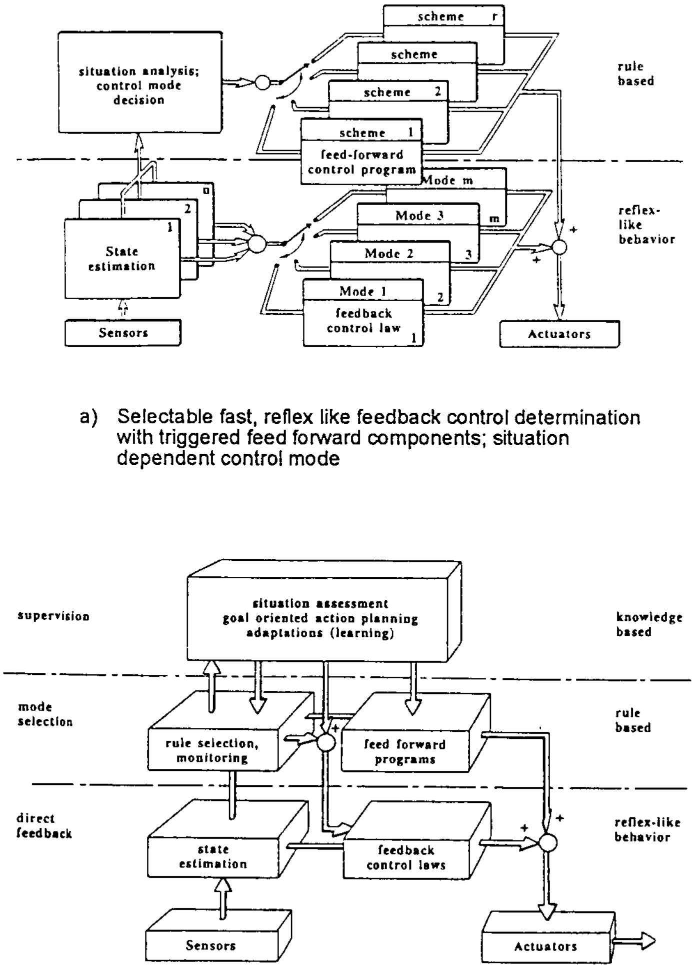
In order to further decouple real-time fast and safe control from slower activities for situation recognition, two different time scales for image sequence interpretation have been introduced in our system. Besides the fast tracks for estimation of relative position, one each for each object of relevance, based on rather few features per object and working at 25 Hz in the new TIP-system, there is one subsystem, attention controlled from the higher levels, which runs at about 5 Hz and is capable of recognizing objects on a more detailed level; specialists for recognition of typical road vehicles[25] (trucks, vans, passenger cars) and of moving humans[26] (walking, running, bicycling and arm waving) are under development.

Perception management

Vision and inertial measurements do have nice complementary properties: Vision incurs long delay times between data collection and object state estimation in general (100 to 300 ms typically, both in biological and in technical systems). In addition, because of the signal integration in the· basic sensing element, motion blur will occur in the image during faster rotation, yielding rate signals derived from image sequences unreliable; on the other hand, inexpensive inertial sensors may yield rather accurate rate signals with almost no time delay. These inertial sensors become expensive when provision has to be made for low drift rates (long term stability). Combining inertial sensors with vision allows to build a flexible system with good overall properties: viewing direction may be stabilized by controlling the platform with the negative angular rate from an inertial sensor, thereby improving the conditions for image evaluation. Visual fixation of the viewing direction onto a well visible set of stationary features allows to solve for the inertial drift problem.

In bifocal vision, two cameras are mounted fixed relative to each other on a gaze control platform; with one wide angle lens for a large viewing angle covered and one tele-lens for high resolution capability within a subarea of the wide angle image, there is a need for viewing direction control in a saccadic mode in addition to the smooth pursuit mode for object tracking: If the tele camera tracks an object, and in the wide angle image a new object of higher interest is discovered, viewing direction should be changed as fast as possible in order to center the new object in the tele-image. With the viewing direction control platform developed[27] a 20° saccade can be performed in about 150 ms; during this fast turn of four evaluation cycles (4·40 ms = 160 ms) no useful information can be derived from the blurred images. Therefore, within these periods the internal representations have to be updated according to the 4-D model exclusively; only after slowdown below a certain angular rate in the vicinity of the coordinates aimed at, image feature extraction will start again with new predicted positions and search ranges. Via a status bit this information is broadcast to all 'object processes' together with the actual viewing direction.

Another important point in perception management on the object level is handling of occlusions. It is not yet clear, how much of this should be done on the 'object process' level and how much on the situation level; both have to deal with the problem in parallel. On the situation level (upper right corner in fig.3) the semantics in the task context have to be taken into account; on the lower object level the problem is to decide in each evaluation cycle which features belong to which object, how this attribution affects relative state estimation and what is the most likely separation line between the two objects.



a) Selectable fast, reflex like feedback control determination with triggered feed forward components; situation dependent control mode

b) Hierarchical scheme for adaptable fast control determination

Fig. 4: Situation dependent, knowledge based intelligent control[28]

In addition, during this process the question has to be answered which state variables and which shape parameters are presently observable; the interpretation models have to be adjusted correspondingly. Especially the relative viewing angle (in azimuth) of vehicles ahead changes observability frequently in typical highway traffic situations[25]; the problem of vehicle length estimation has already been referred to above.

Intelligent control

The own body carrying the camera is now always represented as an object of the real world (number 1 in fig.3, lower right). Since Newtonian motion is of second order in each degree of freedom the state vector also contains all velocity components in 3-D space; this allows state vector feedback for achieving some goal function in an optimal way. Figure 4 shows the general scheme adopted in the 4-D approach to intelligent autonomous systems based on state feedback for fast reactive counteraction to perturbations and event-triggered feed-forward control adaptation to a new situation, both managed by knowledge based situation assessment and behavior selection; by purpose, this is not called behavior planning since the generic, well proven behavioral capabilities are available (or may be learned in more advanced versions) and are just invoked with the right set of parameters. The actual control laws, of course, are specific to the task at hand.

The rest of the paper gives a survey on the different application areas to which the 4-D approach has been successfully applied; it is grouped according to the task fields: Rendez-vous and docking, surface vehicle guidance, and dynamical grasping in 3-D space.

Rendez-vous and docking

Most of the methodical developments of the 4-D approach have been performed by Wuensche[29] on the single sensor, single object problem of controlling planar motion in three degrees of freedom of a tabletop aircushion vehicle with reaction jet control relative to another 3-D body (satellite model plant in the laboratory). The real-time system has been controlled by a VAX 750 combined with a custom-made 8-bit image sequence processing system BVV1[2]. The system performed a self-calibration of the horizontal mounting direction of the camera relative to a docking rod for final mechanical fixture. All six state variables relative to the docking partner of known polyhedral shape, and the vertical mounting direction of the camera have been estimated continiuously by tracking four corner features. Which ones of the usually eight visible features should be selected for tracking in order to achieve optimal accuracy of relative position estimated, has been decided by the system itself exploiting the entries into the Jacobian matrix. While the usage of modified Kalman filters has found very wide acceptance in the vision community in the meantime, the more detailed exploitation of the information in the Jacobian matrix does not seem to have been appreciated correspondingly.

A somewhat different type of rendez-vous with one relative state component (horizontal speed) appreciably different from zero (about 200 km/h in the actual example flown) is the landing approach of an aircraft; in this spatial maneuver the number of state variables is doubled to twelve and there are four analog control variables instead of the three discrete ones with the satellite. Large perturbations may occur due to wind gusts; therefore, inertial sensors (rate and position gyros as well as accelerometers) have been important for robust recognition of the relative position to the runway over the last 1 to 2 Km during approach. For initialisation, signals from a Differential Global Positioning System DGPS have been used[30,31]. Thus, the guidance system may be classified as multi-sensor, single object (besides the own body, of course).

The system has been developed over a period of more than a decade[32], starting from simple all-software-simulations. Lateron, in moving base simulations (three rotations) with computer generated imagery and the real sensor and computer hardware in the real-time loop, fully automatic, on-board autonomous landing approaches (including sidewinds and gusts) until touch down have been demonstrated; this type of flight simulator for machine vision autopilots seems to be the only one in operation up to now.

Real flight experiments have been performed in 1991 with a twin turbo-prop aircraft D0 128 of the University of Braunschweig; the human pilot was in control, but the perception system estimated the complete state vector at a rate of 16 Hz.

In the meantime, the hardware base for the system has been changed to transputers and a bifocal camera arangement; new flight experiments are planned for spring 94.

Surface vehicle guidance

Contrary to the developments in the US-DARPA ALV program, without having knowledge about these activities at all, we started from a behavioral approach based on the 4-D method for continuous vision processes. No higher level AI-components have been involved on our side initially; the system was capable of recognizing local road environments and of reacting in such a way that certain predescribed behavioral parameters like speed, offset from a line or maximum lateral accelerations were observed. The capability of performing (elements of, or full) missions developed over time on this base.

Local reactions for motion control

Road runner: Taking the guideline model for the construction of high-speed roads as the essential knowledge component for recognizing a road recursively while driving on it, a substantial gain in efficiency for image sequence processing has been realized[33]. Mysliwetz[16,34] extended this to robust road recognition, including the general case of hilly terrain, by applying the 'Gestalt'-idea -known from psychology- to road shape recognition from a large number of (approximate) tangent elements. In the classification scheme discussed, this work till the end of the 80ies belonged to the single sensor, single object group. For more demanding real world applications, especially high speed driving, it turned out that a combination of cameras with both small and large focal length, termed 'bifocal vision', is desirable; this combination has been in use for years, but with separate signal evaluation for different objects and purposes. Recently, the signals from both cameras have been used for recognizing the one object road in a joint evaluation[35] ('two sensors, one object'- case; upper central part in fig.3).

Intelligent cruise control: Adding to this lane keeping capability the one for obstacle recognition, relative state estimation and relative state control[23,24], within the same framework of event-triggered feedback and feed-forward behavior selection as shown in fig.4a, remarkable performance sufficient for driving on 'Autobahnen' in normal traffic situations has been achieved[24].

The *reactive feedback* scheme (lower level in fig.4a) is used for realizing:

- lane- and speed-keeping with speed adjusted to horizontal curvature such that a preset lateral acceleration level is not exceeded;
- convoy driving behind another vehicle with distance depending on speed driven (2 seconds rule); this includes 'stop & go' driving in traffic jam as a special case.

The *event-triggered feed-forward scheme* (upper level in fig.4a) provides the capabilities for:

- transition from the unconstrained cruise, lane keeping mode to the convoy driving mode (including stop in front of an obstacle);
- lane changing to left and right. (At present, the human driver has to check whether the intended lane is free; in response to an inquiry he triggers the lane change by hitting a key on the board.)

More than 2 000 km have been travelled autonomously in normal Autobahn-traffic since September 1992 with the two vehicles **VaMoRs** of UniBwM and **VITA** of our industial partner Daimler-Benz.

These results on the two lower levels of fig.4b are achieved essentially with differential representations and local considerations; a different situation occurs when global points of view in a task context come into play. The upper level in 4a (the medium one in 4b) forms the transition from strict, local reactive control to global mission performance.

Global performance (mission control)

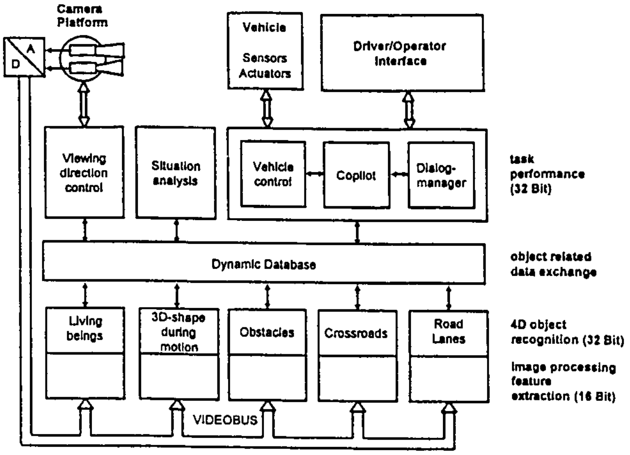
The result of an application of a stereotype feed-forward control time history is a state change with roughly predictable differences between initial and final conditions; this so-called 'maneuver element' may be labeled by a symbol and stands for the finite state transition as an integral representation of this control sequence (e.g. 'lane change': Same driving conditions except for a lateral offset of one lane width).

Maneuver elements may be generic by specifying some parameters which modify the control sequence; for example, lane change may be specified as smooth or rough by fixing the maximum lateral acceleration limit in an otherwise structurally fixed control sequence; this is equivalent to specifying the maneuver time allowed. On the Autobahn, all navigation is done by proper lane changes and lane following; this makes mission performance rather simple. Besides the well structured environment, this was one of the reasons for choosing Autobahn-driving as the first field of application for practical machine vision at the end of the 70ies.

When the capability of reading traffic and navigation signs on the Autobahn is added to the existing system, this goal will be achieved.

Landmark navigation on the factory floor is much more demanding, though in case of failures the damage possible is much less because of the low speeds driven. Taking doors, well visible features on workbenches and other rectangular markers with special height-to-width ratios as landmarks, the suitability of the 4-D approach for real-time visual landmark navigation has been demonstrated in 1991 with moderate computing performance available[36] with an AGV in a laboratory environment and in a factory hall.

Driving on road nets with an automobil in an autonomous mode performing an abstractly defined mission is the most demanding task demonstrated, though yet far from robust real-life applicability. The system shown schematically in fig.5 is in the final stage of development for this purpose: the lower line of blocks is formed by object-specific processor groups each consisting of 16-bit processors for feature extraction and 32-bit processors with floating point units for recursive state estimation; usually, these groups work on separate areas of the image for which they make their own predictions. Occlusions have to be dealt with in cooperation between such processor groups.



- Bifocal camera pair: tele, wide angle with active viewing direction control
- Object-related, intelligently controlled feature extraction
- Recognition of moving objects exploiting spatio-temporal models (4D)
- Situation analysis in task context (AI-Aspects)

Fig. 5: Transputer based system for autonomous mission performance

All object related data are exchanged via a dynamic data base (DDB) which always contains the most recent estimates of object states and parameters. The higher levels of the system shown above the DDB incorporate situation- and control- specific knowledge for finding the best behavioral mode in the actual situation. Both fast reacting state feedback control laws and event triggered feedforward control time histories (as discussed for lane changing but also for turning off onto a cross-road) may be applied. A landmark navigation component has been added on top of this for fully autonomous mission realisation[37].

The system, in a different mode of operation, may also be used for monitoring and warning when the human driver is in control of the vehicle[38].

Dynamical grasping in 3-D space

Quite a different application of the 4-D approach has been demonstrated in May of 1993 during the Spacelab mission D2 with the Space-Shuttle Columbia. One of the experiments on board was the RObot Technology EXperiment ROTEX of DLR, Germany; in this set of tasks under the direction of G.Hirzinger one of the tasks was to grasp an object freely floating in space in a confined workcell for safety reasons. The relative position between the object and the six-degree-of-freedom robot arm was to be determined from a camera in the hand of the robot; one of the difficulties was that the computers for visual interpretation and control had to be on the ground. Due to the routing via three geostationary satellites and quite a few groundstation computers the lumped delay time from measurement taking until the control signal derived from these data again arrived on board the Spacelab was around six seconds!

This-time delay has been compensated by exploiting the dynamical models for the object to be caught and for the robot arm. On May 2nd, 1993, this maneuver has been performed by 'remotely-brained machine vision' automatically after initialisation of visual tracking by a human operator[39].

Conclusions

The development of the 4-D approach to dynamic machine vision continues to be successful. Spatio-temporal models oriented towards physical objects together with the laws of perspective projection in a *forward*-mode (and as approximate linear relationships between the states or parameters of the physical objects and the features by which these objects may be visually recognized) are the core elements of the method. The spatio-temporal models as invariants for object recognition also serve for integrating multi-sensory measurement data.

By prediction error feedback an internal symbolic 4-D representation of processes involving these objects is being maintained allowing situation assessment and longer term predictions.

For specific tasks, behavioral capabilities can easily be realised by state feedback or feed-forward control. The internal feedback loops from state prediction to measurement activities in the image plane make interactions between the higher and lower processing levels very efficient.

References

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

[2] E.D.Dickmanns, V.Graefe: a) Dynamic monocular machine vision, b) Application of dynamic monocular machine vision. J. Machine Vision & Application, Springer-Int., Nov. 1988, pp 223-261

[3] AGARD Lecture Series 185 'Machine Perception', Hampton, VA, USA; Munich; Madrid, Sept. 1992, Chapter 6, 7

[4] U. Franke, H. Fritz, A. Kühnle, J. Schick: Transputers on the road. Proc. World Transputing Conference '93, Aachen, Sept. 1993

[5] A. Blake, A. Yuille (eds.) 'Active Vision', MIT-Press, Cambridge MA, 1992

[6] K. Kluge: YARF: A System for Adaptive Navigation of Structured City Roads. Ph.D.Thesis, CMU School of Comp. Sci., Febr. 1993

[7] D. Koller, K. Daniilidis, T. Thorhallson, H.-H. Nagel: Model-Based Object Tracking in Traffic Scenes. In Proc. ECCV '92, S. Margherita, Italy. Lect. Notes in Computer Science 588, Springer-Verlag, Berlin,1992

[8] D. Terzepoulos, D. Metaxas: Dynamic 3D models with local and global deformations: Deformable superquadrics. IEEE Trans. PAMI, Vol. 13, No. 7, 1991, pp 703-714

[9] R. Koch: Dynamic 3-D Scene Analysis through Synthesis Feedback Control. IEEE Trans. PAMI, Vol. 15, No. 6, 1993, pp 556-568

[10] H.S. Sawhney, J. Olinsis, A.R. Hanson: Image Description and 3-D Deconstruction from Image Trajectories of Rotational Motion. IEEE Trans. PAMI, Vol. 15, No. 9, Sept. 1993, pp 885-898

[11] J. Weng, N. Ahuja, T.S. Huang: Optimal Motion and Structure Estimation. IEEE Trans. PAMI, Vol. 15, No. 9, Sept. 1993, pp 864-884

[12] T.N. Tan, K.D. Baker, G.D. Sullivan: 3D Structure and Motion Estimation from 2D Image Sequences. Image and Vision Comp., Vol. 11, No. 4, May 1993

[13] B. Sridhar, V.H.L. Cheng, A.V. Phatak: Kalman filter based range estimation for autonomous navigation using imaging sensors. IFAC Symposium 'Automatic control in aerospace, Tsukuba, 1989, Pergamon Press, 1990, pp 45-50

[14] 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

[15] K.D. Kuhnert: Zur Echtzeit-Bildfolgenanalyse mit Vorwissen. Diss., UniBw München, Fakultät LRT, 1988.

[16] B. Mysliwetz: Parallelrechner-basierte Bildfolgen-Interpretation zur autonomen Fahrzeugführung. Dissertation, UniBw München, Fakultät LRT, 1990

[17] Dirk Dickmanns: KRONOS-Users Guide. UniBwM/ Inf./1992

[18] E.D.Dickmanns: Subject-object discrimination in 4D dynamic scene interpretationfor machine vision. Proc. IEEE-Workshop on Visual Motion, Newport Beach, 1989, pp 298-304

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

[20] T.J. Broida, R. Chellappa: Estimating the kinematics and structure of a rigid object from a sequence of monocular images. IEEE Trans. PAMI, Vol. 13, 1991, pp 497-513

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

[22] E.D.Dickmanns, T. Christians: Relative 3D-state estimation for autonomous visual guidance of road vehicles. In: Kanade, T. e.a. (ed.): 'Intelligent Autonomous Systems 2', Amsterdam, Dec. 1989, Vol.2 pp. 683-693; also appeared in: Robotics and Autonomous Systems 7 (1991), Elsevier Science Publ., pp 113 - 123

[23] E.D.Dickmanns, B.Mysliwetz, T.Christians: Spatio-temporal guidance of autonomous vehicles by computer vision. IEEE-Trans.on Systems, Man and Cybernetics, Vol.20, No.6, Nov/Dec 1990, Special Issue on Unmanned Vehicles and Intelligent Robotic Systems, pp 1273-1284

[24] E.D.Dickmanns, R.Behringer, C.Brüdigam, D.Dickmanns, F.Thomanek, V.v.Holt: An All-Transputer Visual Autobahn-Autopilot/Copilot. Proc. ICCV'93, Berlin, May 1993. Also in TAT/WTC'93, Aachen, Sept. 1993

[25] M.Schmid: 3D-Erkennung von Fahrzeugen in Echtzeit aus monokularen Bildfolgen. Dissertation, UniBw München, Fakultät LRT, 1993

[26] W. Kinzel; E.D. Dickmanns: Moving Humans Recognition using Spatio-Temporal Models. XVII-th Congress of the Int. Soc. for Photogrammetry and Remote Sensing (ISPRS), Washington D.C., 1992

[27] J.Schiehlen, E.D. Dickmanns: Two-Axis Camera Platform for Machine Vision. 57th AGARD-GCP-Symposium, Seattle, 12.-15. Okt. 1993

[28] E.D.Dickmanns: 4D Dynamic Vision for Intelligent Motion Control. In C. Harris (ed.): Int. Journal for Engineering Applications of AI (IJEAAI), Special Issue 'Intelligent Autonomous Vehicles Research', Vol.4, No.4, pp.301-307, 1991

[29] H.J. Wünsche: Bewegungssteuerung durch Rechnersehen. Fachberichte Messen, Steuern, Regeln, Bd. 20, Springer-Verlag, Berlin, 1988

[30] R.Schell: Bordautonomer automatischer Landeanflug aufgrund bildhafter und inertialer Meßdatenauswertung. Dissertation, UniBw München, Fakultät LRT, 1992

[31] Dickmanns, E.D.; Schell, R.: Autonomous Landing of Airplanes by Dynamic Machine Vision. Proc. IEEE Workshop on 'Applications of Computer Vision', Palm Springs, Nov./Dec. 1992

[32] G.Eberl: Automatischer Landeanflug durch Rechnersehen. Diss., UniBw München, Fakultät LRT, 1987

[33] 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

[34] E.D.Dickmanns; B.Mysliwetz: Recursive 3D Road and Relative Ego-State Recognition. IEEE-Trans. PAMI, Vol.14, No.2, Special Issue on 'Interpretation of 3D Scenes', February 1992, pp. 199-213.

[35] R.Behringer; V. v.Holt; D. Dickmanns: Road and Relative Ego-State Recognition. In: Intelligent Vehicles, IEEE, SAE, Detroit, 1992

[36] C.Hock; E.D.Dickmanns: Intelligent Navigation for Autonomous Robots Using Dynamic Vision. XVII-th Congress of the Int. Soc. for Photogrammetry and Remote Sensing (ISPRS), Washington D.C., Aug. 1992

[37] C.Hock: Wissensbasierte Fahrzeugführung mit Landmarken für automome Roboter. Dissertation, Fakultät LRT, 1993

[38] M.Kopf: Ein Beitrag zur modellbasierten, adaptiven Fahrerunterstützung für das Fahren auf deutschen Autobahnen. Diss., UniBw München, Fakultät LRT, 1993

[39] C.Fagerer, Dirk Dickmanns, E.D.Dickmanns: Visual Grasping with Long Delay Time of a Free Floating Object in Orbit. UniBwM / LRT / WE13 / FB 93-3