Engineering of Intelligent Systems (EIS 2004)
Febr.29 – March 3, 2004, Madeira

Three-Stage Visual Perception for Vertebrate-type Dynamic Machine Vision

Ernst D. Dickmanns

Universitaet der Bundeswehr Munich (UBM)
85577 Neubiberg, Germany
Ernst.Dickmanns@unibw-muenchen.de

Abstract: Efficient real-time visual perception in civilized natural environments (e.g. road networks) has to take advantage of foveal – peripheral differentiation for data economy and of active gaze control for a number of benefits. 1. Inertial gaze stabilization considerably alleviates the evaluation of image sequences taken by cameras with stronger tele-lenses; it allows a reduction in angular disturbances from rough ground by at least an order of magnitude with simple negative angular rate data feedback. 2. Visual tracking of fast moving objects reduces motion blur for these objects. – 3. In the near range, a large field of view is mandatory, however, only coarse angular resolution is sufficient; with a field of view (f.o.v.) > ~ 100°, both the region in front of and to the side of the vehicle may be viewed simultaneously. For own behavior decision, motion behaviors of objects both in the wide f.o.v. nearby and in several regions of special interest further away have to be understood in conjunction.

In order to achieve this efficiently, three distinct visual processes with specific knowledge bases have to be employed in a consecutive way. In the wide f.o.v., bottom-up feature extraction has to answer the question: 'Is there anything of special interest?' The corresponding feature extraction operators are domain-specific. On initialization, they have to give indications of objects of interest all over the image. Stable feature aggregations over several cycles have to trigger object hypotheses for the second stage; these regions may then be discarded for stage 1. Stage 2 works on single objects, however, on multiple of these in parallel. When looking almost parallel to the ground this is necessary for proper scaling, since each line in the image represents a different distance on the ground plane. For this reason, feature extractors and state estimators for each object have to be tuned specifically. Representing 3-D objects in 3-D space and time allows exploiting the first order derivative matrix of the perspective mapping process (the so-called 'Jacobian') and spatial interpretation despite the fact that range has been lost in each single image point. The parallel processes of stage 2 yield best estimates of the relative state 'here and now' for all objects observed. They are symbolically represented in a scene tree known from computer graphics. Stage 3 works on time series of these symbolic data in order to more deeply understand motion sequences on larger spatial and temporal scales (maneuvers, mission elements) affecting own decision making. For more autonomous, flexible use, both perceptual and behavioral capabilities are represented explicitly. Experimental results will be given for the test vehicles **VaMoRs** (a 5-ton van) and **VaMP** (Mercedes 500 SEL).

Introduction

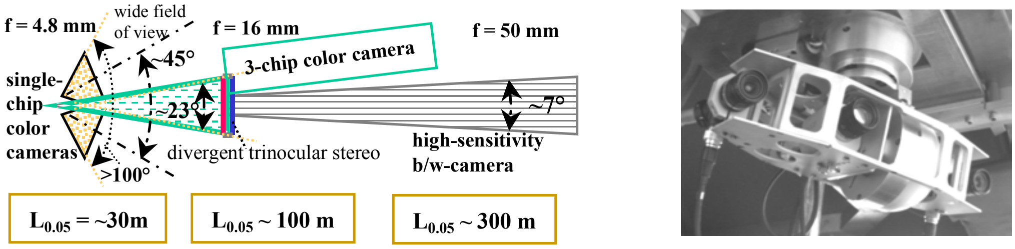
Nature has spent countless numbers of iterations for developing in the vertebrate eye (and the correspondingly tuned control and data evaluation systems) a vision capability unparalleled by any other approach up to now. For guidance of human-built vehicles, all these capabilities in the form available in humans have also to be provided for technical counterparts in machine vision, at least in the long run if they are intended to really alleviate vehicle operation. For this purpose, it does not make a difference whether the technical systems are to be assistance systems for human operators or are intended to work on their own (autonomous driving). Acceptance depends on a similar (or improved) performance level both in accuracy, in range of applicability, and in reliability.

Since about two decades, simple vision systems for ground vehicle applications are under development [Dav et al. 86; DiZ 86; KPH 86; Tho et al. 86; Tur et al. 87]. A review is given in [Dic 02]. Almost all systems depend on one or two cameras mounted fix onto the vehicle body; thus, they are fully susceptible to angular motion perturbations from rough ground, which is especially detrimental for objectives with large focal lengths. Driving relatively slowly with small look-ahead ranges and on smooth surfaces, this disadvantage is not yet felt in most cases up to now. Going cross-country on rough ground or driving at high speeds with the need for long look-ahead ranges (up to several hundred meters), active gaze control seems unavoidable. For this reason, UBM has developed an

Expectation-based, **M**ulti-focal, **S**accadic (EMS-) vision system [InV'00] which mimics vertebrate-type vision. The corresponding complex vision sensor is an arrangement of three to four cameras mounted fix relative to each other on a pointing platform with high dynamic bandwidth [PLD 01].

'MarVEye' as complex vision sensor

Figure 1 shows on the left side the multiple f.o.v. for a four-camera arrangement, while the picture on the right is made from a three-camera arrangement. Three more cameras can be fixed to the ground plane. For a large stereo f.o.v. as used for cross-country driving, a pair of coaxial stereo cameras is mounted directly under the pair with divergent optical axes for a large f.o.v. seen on the left and right.



a) Fields of view and viewing ranges of MarVEye.

b) Realization of MarVEye4 with 3 CCD-cameras on a two-axis platform.

Figure 1: MarVEye system parameters (a), camera set in VaMoRs on yaw and pitch platform, large stereo base and one mild tele-camera (b).

Adaptation of the 4-D approach to perception with MarVEye

The so-called 'scene tree' has been introduced by [DDi 97] for separating a general framework for visual perception from the actual description of a scene observed. It contains all (real and intermediate virtual) stages (coordinate frames) for linking objects in the real world (like roads with lanes and vehicles) to visual features in the image plane. Points on objects are transformed by 4*4 **H**omogeneous coordinate **T**ransformations **M**atrices (HTM) from the real world into the image plane of each camera in the own vehicle (see exploded view on top left with the corresponding coordinate systems in figure 2). The fourth component in HTM's serves scaling, and allows to include perspective projection into this scheme [Rob 77]. All the rest of the 4-D approach as described in [DiG 88; DiW 99] can be left as it stands. However, since saccades introduce heavy motion blur during their time of duration (a few tenths of a second), an 'image valid'-bit has been introduced indicating periods (with value 'true') when image evaluation should be performed. When this bit is set 'false', scene perception is continued based exclusively on model prediction, which is performed every cycle anyway in the recursive estimation process. Trust in results decreases accordingly over time in this phase.

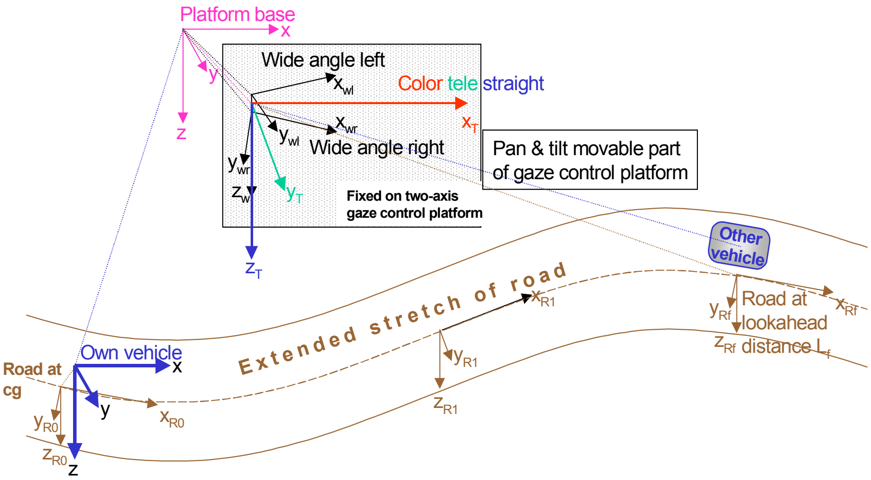


Figure 2: Homogeneous coordinates for forming the scene tree of a road scene.

Three stages of visual perception

According to the data and the specific knowledge levels to be handled, three separate stages of visual perception can be distinguished as indicated in figure 3 (to be read from bottom to top).

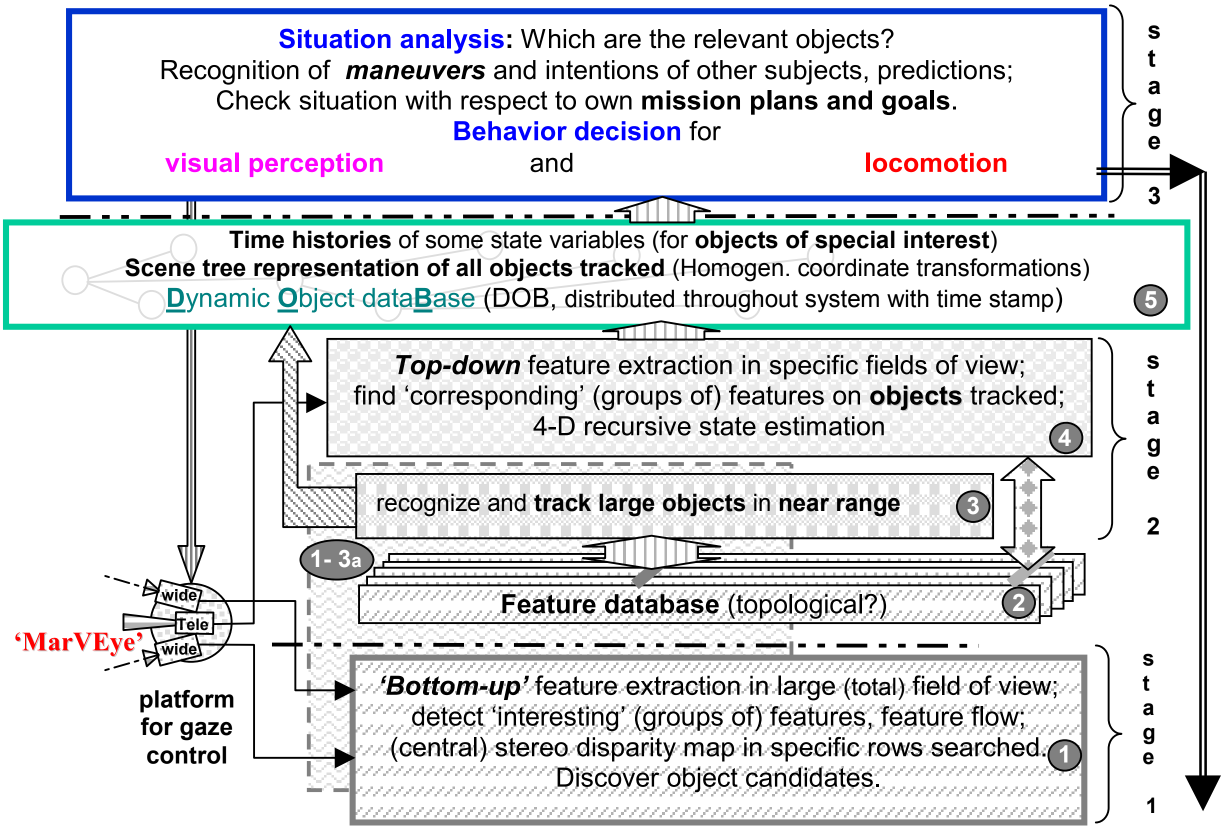

Situation analysis: Which are the relevant objects?
Recognition of maneuvers and intentions of other subjects, predictions;
Check situation with respect to own mission plans and goals.
Behavior decision for
visual perception
and
locomotion
stage 3
Time histories of some state variables (for objects of special interest)
Scene tree representation of all objects tracked (Homogen. coordinate transformations)
Dynamic Object dataBase (DOB, distributed throughout system with time stamp)
5
Top-down feature extraction in specific fields of view;
find 'corresponding' (groups of) features on objects tracked;
4-D recursive state estimation
4
stage 2
recognize and track large objects in near range
3
1- 3a
Feature database (topological?)
2
wide
Tele
wide
'MarVEye'
platform for gaze control
'Bottom-up' feature extraction in large (total) field of view;
detect 'interesting' (groups of) features, feature flow;
(central) stereo disparity map in specific rows searched.
Discover object candidates.
1
stage 1

Figure 3: Three distinct levels of visual perception: 1. Bottom-up detection of standard features; 2. Specific features for individual objects under certain 3-D aspect conditions at certain ranges; symbolic storage of results 'here and now' for many objects in parallel in DOB. 3. Look at time histories of state variables in order to recognize maneuvers and mission elements performed (trajectory and intent recognition at larger spatial and temporal scales).

In a methodologically clean architecture, the sub-stages numbered 1 to 3 in dark circular regions should be completely separated. Due to hardware constraints in frame-grabbing and communication bandwidth, they have been lumped together on one computer in actual implementation for the wide f.o.v.; the elliptical dark field on the shaded rectangular area at the left hand side (1 – 3a) is to indicate that object recognition based on wide-angle images is performed on the corresponding features by the same processor directly, in the actual implementation. In the long run, stages 1 and 2 are good candidates for special hard- and software on plug-in processor boards for PC.

Stage-2 image processing (labeled 3 and 4 in the dark circles) requires a completely different knowledge base. It is applied to (usually quite small) special image regions discovered in the feature database by interesting groups of feature collections. For these feature sets, object hypotheses in 3-D space and time are generated. These 4-D object hypotheses with specific assumptions for range and other aspect conditions allow predictions of additional features to be detectable with specially tuned operators (types and parameters), thereby realizing the idea of 'Gestalt'-recognition. Several hypotheses may be put up in parallel if computing resources allow this; invalid ones will disappear rapidly over time since the Jacobian matrices allow efficient use of computational resources. Each hypothesis consists of a 3-D shape model and a dynamical model for the degrees of freedom in motion available. For subjects with potential control applications (see next but one paragraph) their likely value settings also have to be guessed and iterated.

Determining the angular ego-state from inertial measurements (at least the medium- to high-frequency components) allows a reduction of search areas in the image, especially in the tele-images. All results are stored in the **D**ynamic **O**bject data**B**ase (DOB) which serves the purpose of data distribution; copies of this DOB are sent to each computer in the distributed system. The DOB also isolates the higher system levels from the high visual and inertial data rates to be handled in real-time at the lower levels. Here, only object identifiers and the actually best estimates of state variables and control variables as well as model parameters in the sense of systems dynamics are stored. By this approach, a reduction of several orders of magnitude in data rates is achieved.

By looking at time series of these variables (of limited extension), the actually most interesting objects for own decision-making and their likely future trajectories have to be found. For this temporally deeper understanding of movements, the concept of 'maneuvers' and 'mission elements' for achieving some goals of the subjects in control of these vehicles has to be introduced. Quite naturally, this leads to explicit representation of behavioral capabilities of subjects for characterizing their choices in decision making. Classes of subjects are defined as special classes of objects with the capabilities of sensing (perceiving), decision making and actuation of some control variables. These may activate gaze control, the mode of visual perception, or locomotion control. Realization of these behavioral activities usually is performed on special hardware (real-time processors) close to the physical device. Control engineering methods like parameterized stereotypical feed-forward control time histories or (state- or output-) feedback control schemes are applied here. However, for achieving autonomous capabilities, an abstract (quasi-static) representation of these control modes and the transitions possible between them has been implemented recently on the higher system levels by AI-methods according to extended state charts [Har 87; Mau 00; Sie 03]. This allows more flexible behavior decisions based also on maneuvers of other subjects supposedly under execution. This is part of situation assessment in a defensive style of driving. For more details on **B**ehavior **D**ecision for **G**aze and **A**ttention (BDGA, see [InV'00(f), PLD 01, Pel 03]).

Explicit representation of perceptual and behavioral capabilities

Figure 4 gives a survey on the perceptual capabilities of the **M**ultifocal **a**ctive / **r**eactive **V**ehicle **Eye** 'MarVEye'. All activities are based on two torque motors of the platform for gaze control (lower layer). Five skills are shown on the second level, from which three are the basic ones and two use a fixed sequence of activation. Complex viewing patterns can be activated from the upper level triggered by the **C**entral **D**ecision process CD, which

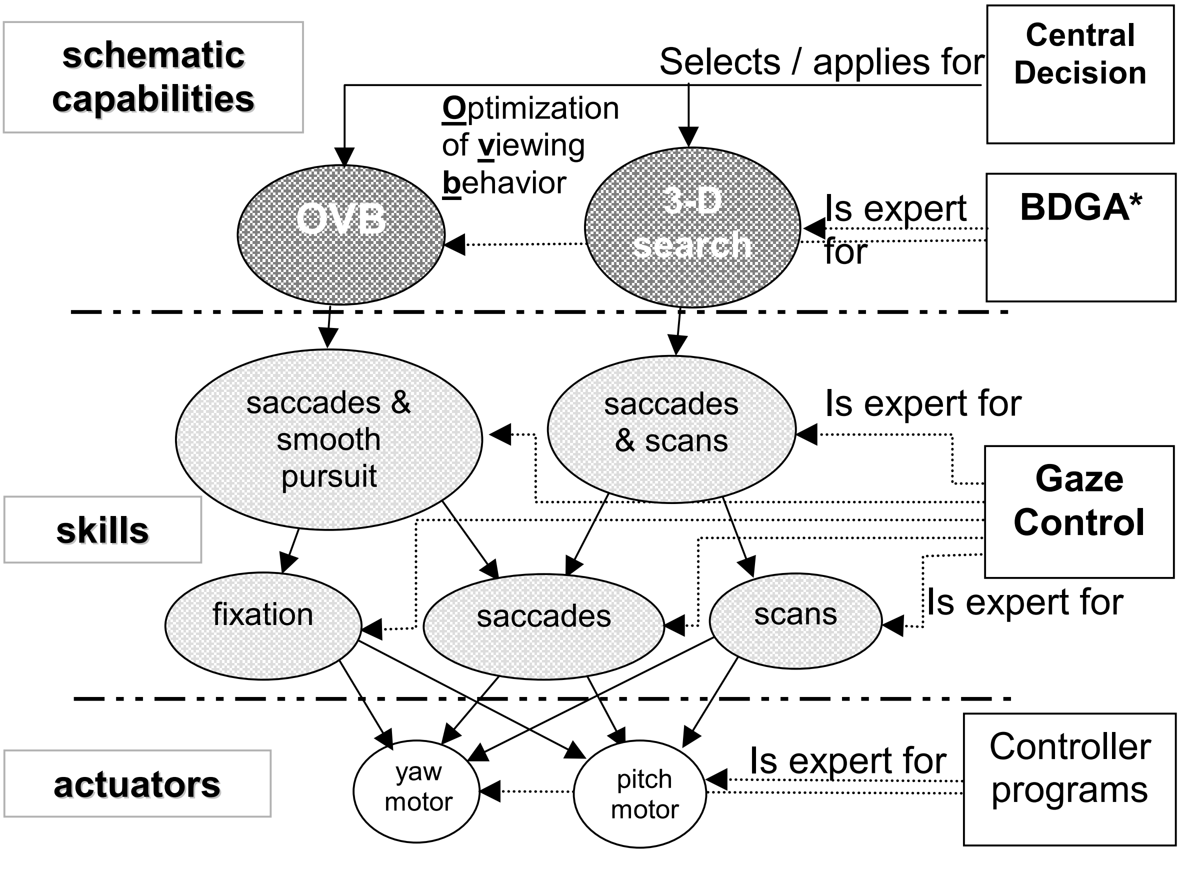


Figure 4: Network of perceptual capabilities with **MarVEye** structured on three levels in order to show dependencies; before activation, all levels are checked for complete availability (after [Pel 03]).

controls the mission with its consecutive elements. **O**ptimization of **V**iewing **B**ehavior is a specialist process for finding an optimal compromise between all viewing demands from the object recognition specialists requested by BDGA. This is a rather involved topic and cannot be treated here; both compromising single viewing directions and saccades between two or three viewing directions with short phases of smooth pursuit can result (for details see [InV'00(f), PLD 01, Pel 03]).

Figure 5 gives a survey on behavioral capabilities for locomotion and their interdependencies in a similar network. Like for gaze control, conditions for transition between the modes and adjustable parameters are again visualized in extended state charts [Har 87]. Space does not allow going into details here. Full details are available in the dissertations [Mau 00; Sie 03], brief outlines in [InV'00 (e and g)]. Proper activation of all these capabilities with the right trigger point in time is the major integration challenge. This will be surveyed after a quick view on the hardware base for system integration.

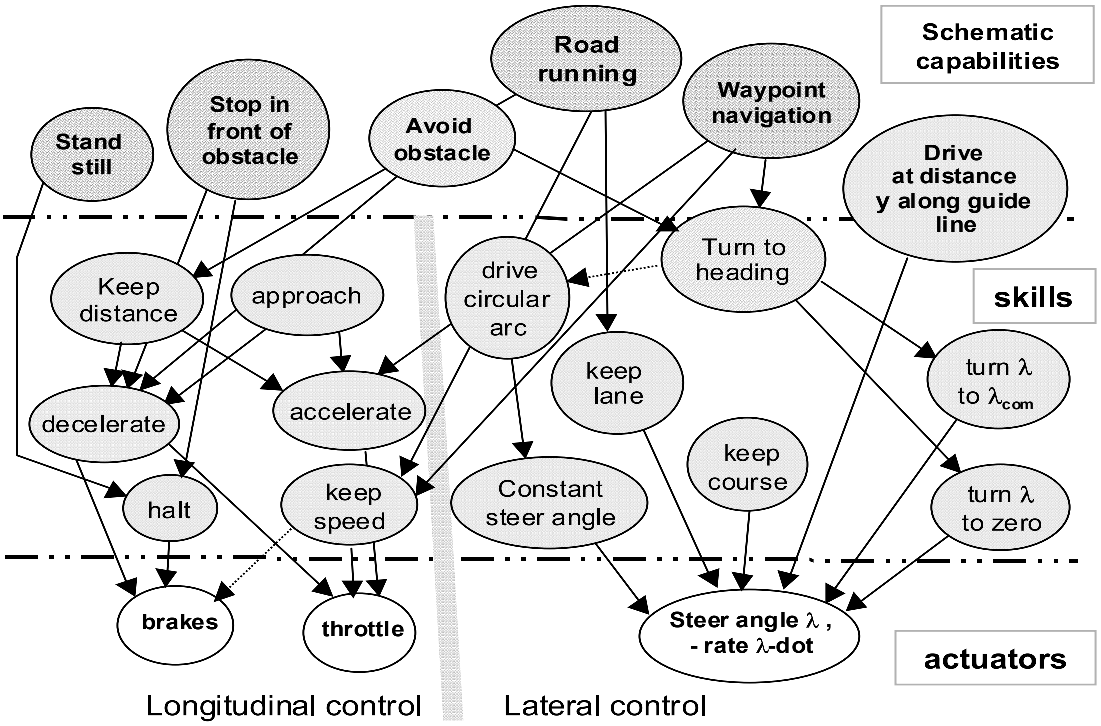


Figure 5: Network of behavioral capabilities of a road vehicle. Longitudinal and lateral control are fully separated only on the hardware level with three actuators; there are many basic skills realized by diverse, parameterized feed-forward and feedback control schemes (layer 2). On the upper level, abstract schematic capabilities as triggered from CD are shown (after [InV'00 (e and f); Mau 00; Sie 03]).

System integration

Figure 6 shows a coarse block diagram of the system in the 5-ton van **VaMoRs** (see lower right corner). Four PC with DualPentium processors provide the computing power for image processing and the higher system levels. In order to cope with the limited data rate capabilities, cameras and frame-grabbers are associated with specific PC (3 to the left). The fourth PC handles the subsystems for gaze and locomotion control implementation as well as for data capture from conventional sensors and from the Global Positioning System (GPS). It also hosts the human-machine interface (HMI) and the three processes for behavior decision (CD, BDGA and BDL). All processors are started and controlled through HMI with the help of an Embedded PC-demon process (EPC) and Ethernet links. Larger data volumes are exchanged between the processors by a Scalable Coherent Interface (SCI) which also serves for synchronization between the devices needing this. **D**ynamic **K**nowledge **R**epresentation (DKR) is the process in charge of communicating all actual data of interest to the distributed processors. Details of this implementation may be found in [Rie 00].

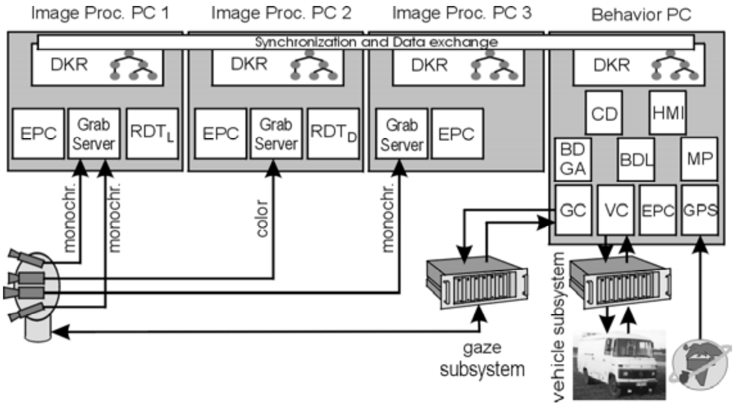


Figure 6: Hardware realization on a cluster of PC's plus two subsystems for hardware interfacing (gaze and vehicle control).

Figure 7 shows the activation, prioritization and monitoring scheme implemented, based on the specific capability networks BDGA and BDL discussed above.

If everything works properly, decisions on the higher levels derived by AI-methods are implemented through the networks and the communication channels by proper timing of control initialization and sequencing on the subsystem levels. The meaning of the acronyms is explained in the legend of the figure.

Experimental results

Avoiding negative obstacles (larger ditches and holes) while driving cross-country is one of the more demanding tasks for autonomous driving. With a size just large enough so that a wheel can be trapped, these obstacles can hardly be detected at larger distances. In the near range, these obstacles can be detected most reliably by stereo vision. However, with grass and low growing, even thin vegetation easily traversable by wheeled vehicles, it is very hard to detect ditches by stereo vision. On the other hand, ditches and holes with steep slopes, usually miss or have different vegetation on these slopes so that they show up in color (and very often even in gray-scale) images by different photometric appearance (color, intensity and texture). For this reason, the real-time stereo system available in **VaMoRs** [Sie et al. 01] has been complemented with a photometric component in the EMS framework for ditch detection at longer distances [PHD 03].

In a video-film a short mission for a ground vehicle will be shown in order to demonstrate the capabilities of the system: 1. Driving on an unmarked minor road with scene interpretation from the wide-angle cameras. 2. Detection of a crossroad with the tele-camera and determination of its width and intersection angle by active gaze control. 3. Making a turn to the right onto the crossroad (105°-intersection angle). 4. Road following along one side of this wider road. 5. Leaving this road to the left for going cross-country on a grass surface; the trajectory is fixed by a sequence of way-points to be followed using GPS-data. 6. Detection of a ditch on this trajectory, the location of which is unknown to the system. 7. For gaining more time to completely determine the size and relative location of the ditch, the vehicle stops and performs a number of saccades, in order to precisely recognize the extension of the ditch. 8. A new part of the trajectory has to be planned avoiding the ditch and going around the nearest corner. 9. This trajectory is now followed with gaze fixation on the nearest corner until the front wheels are tangential to the trajectory near the corner. 10. Now the vehicle resumes way-point navigation by GPS which is a specific behavioral capability (see WPN on the lower dark bar in figure 7 (near BDL)).

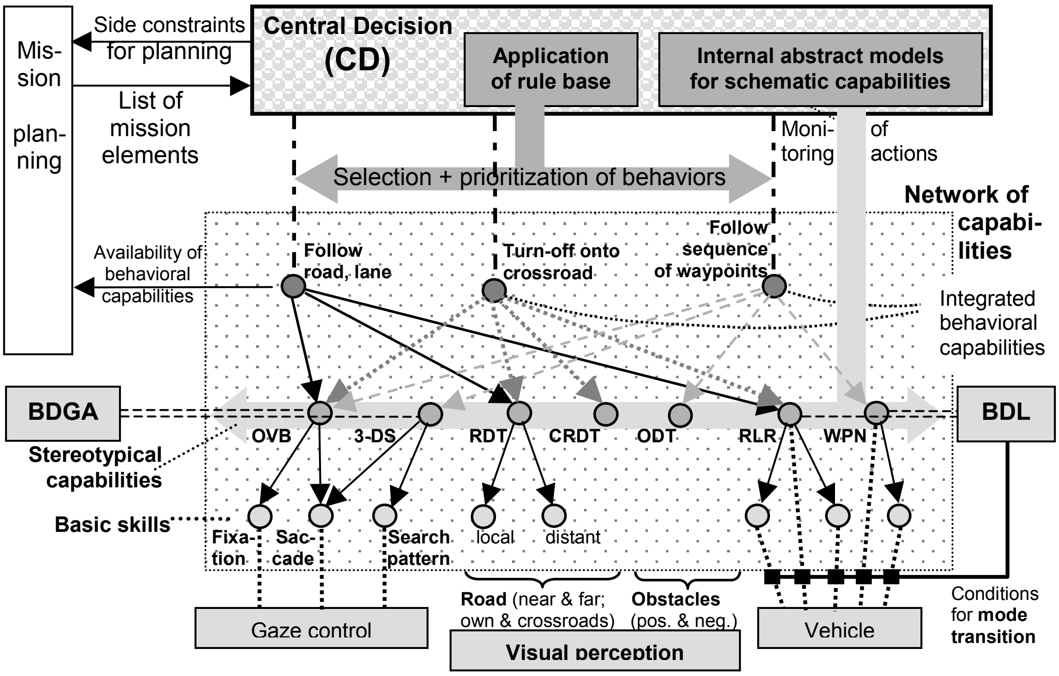

Mission planning
Side constraints for planning
List of mission elements
Central Decision (CD)
Application of rule base
Internal abstract models for schematic capabilities
Monitoring
of actions
Selection + prioritization of behaviors
Network of capabilities
Availability of behavioral capabilities
Follow road, lane
Turn-off onto crossroad
Follow sequence of waypoints
Integrated behavioral capabilities
BDGA
BDL
OVB
3-DS
RDT
CRDT
ODT
RLR
WPN
Stereotypical capabilities
Basic skills
Fixation
Saccade
Search pattern
local
distant
Road (near & far; own & crossroads)
Obstacles (pos. & neg.)
Conditions for mode transition
Gaze control
Visual perception
Vehicle

Figure 7: Activation, prioritization and monitoring of some stereotypical capabilities by Central Decision (CD) exploiting the capability network and special decision units for Gaze and Attention (BDGA) and for locomotion (BDL). Perceptual capabilities for roads and obstacles are provided by specialists for certain object classes from the image stream delivered (not detailed here).
Legend: **OVB** = Optimization of Viewing Behavior; **3-DS** = Search in 3-D space; **RDT** = Road (or lane) Detection and Tracking, this is achieved with different algorithms (basic skills) for the near and the far range; **CRDT** = CrossRoad Detection and Tracking; **ODT** = Obstacle Detection and Tracking, both stationary and moving, above the ground (positive) and missing support for the wheels (deep potholes and ditches = negative obstacles); **RLR** = Road (or Lane) Running (lateral guidance with appropriate speed; **WPN** = WayPoint Navigation when driving off the road (based on GPS). {For a complex maneuvering capability like turning-off onto a crossroad and going around a ditch in cross-country driving, see the video film.}

Conclusions

Explicit representation of perceptual and behavioral capabilities on higher system levels in an abstract form increases flexibility in mission performance and more growth potential towards really autonomous systems. EMS-vision with foveal – peripheral differentiation is a very data-efficient approach mimicking vertebrate vision in the biological realm. Three distinct stages in visual perception with own knowledge bases behind them can be distinguished: 1. Bottom-up extraction of visual features in all images. 2. Multiple parallel tracking of single objects by specialist processes and determining their relative 4-D state ‘here and now’. 3. Collection of these symbolic data in a scene tree for combined analysis on larger scales in space and time with respect to mission performance (situation assessment exploiting locomotion capabilities of other vehicles in a predictive mode). As new knowledge elements, stereotypical maneuvers and behavioral modes realized by state-feedback are being exploited for classes of subjects. - The approach has been realized in two vehicles (5-ton van and sedan). Mission performance with **VaMoRs** both on- and off-road has validated the approach. Explicit representation of all types of capabilities of subjects is considered to be the way to go for really autonomous systems. Gaze control and locomotion control as demonstrated here are just an entry into this promising field for the future.

References

[AlM 01] Albus J.S., Meystel A. M.: Engineering of Mind. – An introduction to the science of intelligent systems. J. Wiley & Sons Publication, New York, 2001, 411 pages

[Dav et al. 86] Davis L.S., T.R. Kushner, J.L. LeMoigne, A.M. Waxman: Road Boundary Detection for Autonomous Vehicle Navigation. Optical Engineering, Vol. 25, No. 3, March 1986, pp. 409 – 414.

[Dic 87] Dickmanns E.D.: 4D-Dynamic Scene Analysis with integral Spatio-Temporal Models. Proc. 4th Int. Symp. on Robotics Research (ISRR-4), Santa Cruz, CA, USA, Aug. 1987

[Dic 02] Dickmanns E.D.: Vision for ground vehicles: history and prospects. Int. J. of Vehicle Autonomous Systems, Vol.1, No.1, 2002, pp. 1 – 44.

[DiG 88] Dickmanns E.D., Graefe V.: a) "Dynamic monocular machine vision", b) "Application of dynamic monocular machine vision". J. Machine Vision & Application, Springer-Int., Vol. 1, Nov. 1988, pp 223-261

[DiZ 86] Dickmanns E.D., Zapp A.: "A Curvature-based Scheme for Improving Road Vehicle Guidance by Computer Vision". In: [SPIE 86], pp 161-168.

[DiW 99] Dickmanns, E.D. and H.-J. Wünsche: Dynamic Vision for Perception and Control of Motion. In: B. Jaehne, H. Haußenecker and P. Geißler (eds.) Handbook of Computer Vision and Applications, Vol. 3, Acad. Press, 1999, pp 569-620.

[DDi 97] Dirk Dickmanns: Rahmensystem für visuelle Wahrnehmung veränderlicher Szenen durch Computer. Dissertation, Universität der Bundeswehr München, Fakultät für Informatik, 1997.

[Gra 84] Graefe V.: Two Multi-Processor Systems for Low-Level Real-Time Vision. In J.M. Brady et al. (eds.): Robotics and Artificial Intelligence, Springer-Verlag, 1984, pp. 301 – 308.

[InV’00] Proc. of the Symposium on ‘Intelligent Vehicles’. Dearborn, MI, Oct. 2000, with the following contributions on EMS-Vision:

a) Gregor, R., Lützeler, M., Pellkofer, M., Siedersberger, K.H. and Dickmanns, E.D.: EMS-Vision: A Perceptual System for Autonomous Vehicles.

b) Gregor, R. and Dickmanns, E.D.: EMS-Vision: Mission Performance on Road Networks.

c) Hofmann, U. and Dickmanns, E.D.: EMS-Vision: An Application to Intelligent Cruise Control for High Speed Roads.

d) Lützeler, M. und Dickmanns, E.D.: EMS-Vision: Recognition of Intersections on Unmarked Road Networks.

e) Maurer, M: Knowledge Representation for Flexible Automation of Land Vehicles.

f) Pellkofer, M. and Dickmanns, E.D.: EMS-Vision: Gaze Control in Autonomous Vehicles.

g) Siedersberger, K.-H.: EMS-Vision: Enhanced Abilities for Locomotion.

[Har 87] Harel, D.: State charts: A Visual Formalism for Complex Systems. Science of Computer Programming. 1987, Vol. 8, pp. 231-274.

[KPH 86] D. Kuan, G. Phipps, A. Hsueh: A Real-Time Road Following Vision System for Autonomous Vehicles. (see [SPIE 86], pp. 152 – 160)

[Mau 00] Maurer M.: Flexible Automatisierung von Strassenfahrzeugen mit Rechnersehen. Diss. UniBw Munich, LRT, 2000.

[Pel 03] M. Pellkofer: Verhaltensentscheidung für autonome Fahrzeuge mit Blickrichtungssteuerung. Diss., UniBwM, LRT, 2003

[PHD 03] Pellkofer M., Hofmann U., Dickmanns E.D.: Autonomous Cross Country Driving Using Active Vision. SPIE-Aero-Sense, Proc. ‘Unmanned Ground Vehicles’, Orlando, April 2003

[PLD 01] Pellkofer M., Lützeler M., Dickmanns E.D.: Interaction of Perception and Gaze Control in Autonomous Vehicles. Proc. SPIE: Intelligent Robots and Computer Vision XX; Oct. 2001, Newton, USA, pp 1-12

[Rie 00] A. Rieder: Fahrzeuge sehen. Diss. UniBwM, LRT, 2001.

[Rob 77] L.G. Roberts: Machine Perception on Three-Dimensional Solids. In: J.K. Aggarwal et al. (eds): Computer Methods in Image Analysis. IEEE Press, New York, 1977, pp. 285 – 323.

[Sie et al. 01] Siedersberger K.-H.; Pellkofer M., Lützeler M., Dickmanns E.D., Rieder A., Mandelbaum R., Bogoni I.: Combining EMS-Vision and Horopter Stereo for Obstacle Avoidance of Autonomous Vehicles. Proc. ICVS, Vancouver, July 2001

[Sie 03] Siedersberger K.-H.: Komponenten zur automatischen Fahrzeugführung in sehenden (semi-) autonomen Fahrzeugen. Diss. UniBw Munich, LRT, 2003 (to appear).

[SPIE 86] SPIE Internat. Conf. on ‘Mobile Robots’, Vol. 727, Cambridge, MA, USA, Nov. 1986.

[Tho et al. 88] Thorpe C., M. Hebert, T. Kanade, S. Shafer: Vision and Navigation for the Carnegie-Mellon Navlab. IEEE Trans. PAMI, 1988, Vol. 10, No. 3, pp. 401 – 412

[Tur et al. 87] Turk M.A., D.G. Morgenthaler, K.D. Gremban, M. Marra: Video Road-Following for the Autonomous Land Vehicle. Proc. IEEE Int. Conf. on Robotics and Automation, Raleigh, NC, USA, March 31 – Apr. 3, 1987, pp. 273 – 280.