Contributions to Visual Autonomous Driving

A Review

Ernst D. Dickmanns

Part III: Expectation-based, Multi-focal, Saccadic (EMS-) Vision (1997 – 2004)

Abstract (Part III): The third-generation system for dynamic vision based on spatio-temporal models and COTS-hardware is described. It has been developed in a joint effort with US-American partners contributing full-frame real-time stereo-vision and overall architectural aspects. Vehicles equipped with the system have been VaMoRs and Wiesel_2 on the German side as well as HMMWV and XUV on the American side. Driving in networks of minor roads, off-road, and performing transitions between these modes avoiding both positive and negative obstacles (ditches) has been demonstrated in 2001 (with a separate computer for stereo-vision) and in 2003 with an American stereo system sized as single European-standard-board integrated into one of the four Dual-PC forming the entire EMS-vision system.

III.1 Introduction

In parts I and II the basic approach to real-time computer vision with spatiotemporal models using custom-designed systems of standard digital microprocessor and a specific window-scheme for reducing storage requirements have been discussed. Part II described the 'transputer'-system developed in the framework of the PROMETHEUS-project and the transition to 'Commercial-Off-The-Shelf'-(COTS-) components. Surprising high-performance results have been shown both at the final demonstration of the PROMETHEUS-project in October 1994 on Autoroute_1 near Paris with the full transputer system running at 12.5 Hz, and in the fall of 1995 by the long-distance drive (> 1600 km) with four Motorola-Power-PC (MPC) replacing many transputers for object detection and tracking in the vision system. Due to the tenfold increase in computing power of the MPC over transputers, the image evaluation rate had been doubled to 25 Hz (40 ms cycle time) while the number of processors was reduced by a factor of five.

During the long-distance test drive over more than 1600 km to a project meeting in Odense, Denmark, the goal was to collect systematic information on situations when the system with standard black-and-white edge feature extraction failed and had to be reset. From these statistical data the necessary next steps had been derived for most efficient improvement of performance by

adding new components, now in reach at reasonable costs. The results given in section 4 of Part II (II.4) were the following:

- Add color processing capability for at least one of the cameras of the multi-focal system.
- Install error detection capabilities on all levels from feature detection over object tracking to situation asessment; improve automatic reset capabilities by increased exchange of information between the levels.
- Increase viewing range L_5 to 250 to 300 m (one pixel corresponds to 5 cm orthogonal to the viewing direction, ~ 0.2 mrad / pixel); this entails the need of inertial stabilization at least for the tele camera. A large simultaneous field of view is needed in parallel.
- Once fast active gaze control is available, it should be used also for gaze fixation on sets of features of special interest and for saccadic shifts of attention.

These requirements led to the concept of "**E**xpectation-based, **M**ulti-focal, **S**accadic" (EMS-) vision to be discussed after in section III.2 the hardware base selected has been discribed.

III.2 Selection of the hardware base

Since no reduction in growth potential for digital microprocessors (about a factor of ten every four to five years) was in sight, also the third-generation dynamic vision system was conceived on the basis of general PC-components. At about that time the American effort with respect to specially designed massively parallel vision processors faded [Roland and Shiman 2002]. Even the initially successful computer system for vision processing 'The Connection Machine' declined [Hillis 1992]. Developments on the market for 'general-purpose-processing' were so fast that special processing systems needing new operating systems and application software were not able to compete. In view of this fact and of the upcoming local communication networks between processors, UniBwM decided to use industry-standard PC-processors from one of the main providers, after the previously announced pin-compatibility for the next-generation MPC had not been realized.

Based on previous good experience with industrial Intel-processors, four Dual-Pentium_4 units with clock rates between 300 and 700 MB were chosen in the mid-1990s in conjunction with the 'Scalable Coherent Interface' (SCI) for synchronization and data exchange between these

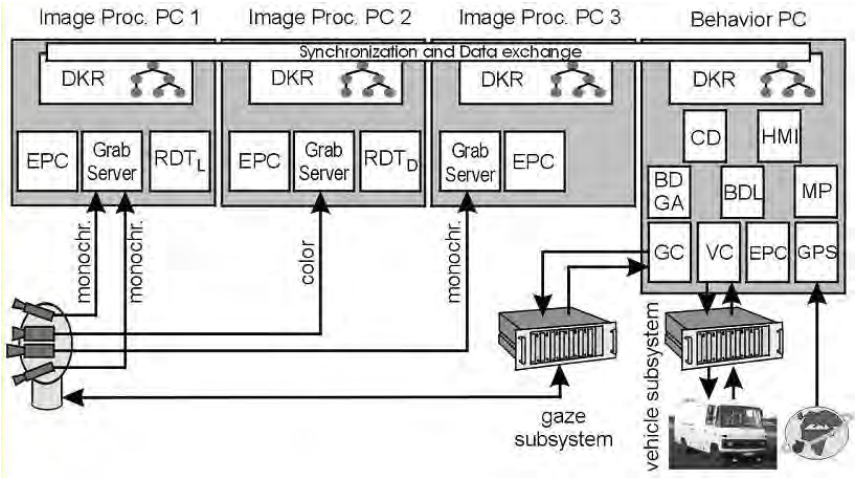

Figure III.1: Third-generation (Commercial-Off-The-Shelf, COTS) vision system of UniBw Munich realizing '**E**xpectation-based, **M**ulti-focal, **S**accadic' (EMS)-vision. (Details may be found in www.dyna-vision.de)

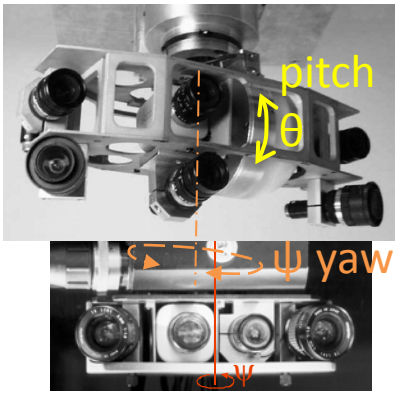

Figure III.2: ‚Multi-focal Eye' for VaMoRs (top, two axes: Yaw ψ and pitch θ) and for VaMP (bottom, yaw ψ only).

units. Figure III.1 shows that three of these systems were used for image processing. The fourth one, dubbed 'Behavior PC', handled all interface operations with the gaze control (GC) and vehicle control (VC) sub-systems, which continued to be transputer-based, and with the Global Positioning System (GPS). The 'vehicle-eye' (Figure III.2) with up to five active video cameras has been designed differently for VaMoRs with two degrees of freedom and for VaMP with gaze control in yaw only. While VaMP drives on smooth roads, pitch control has been left off for smaller overall size and reduced requirements in electrical power for the passenger car. VaMoRs was expected to also drive in rough terrain and to handle transitions between road and off-road driving that deviate pronouncedly from planar surfaces, in general; power requirements were of no concern in VaMoRs.

The transputer subsystems for gaze- and vehicle-control allowed sticking with the conventional operating system for the Dual-PC and the SCI. Synchronization was achieved through SCI that also provided to all processors the actually valid best estimates for objects observed and for their states; Dynamic Knowledge Representation (DKR) in all four processors was thus kept identical (top row in Fig. III.1). The **B**ehavior PC then had to come up with **D**ecisions for **G**aze and **A**ttention (BDGA) and for **L**ocomotion of the own vehicle (BDL); in rare cases of conflicting decisions on this level, a special '**C**entral **D**ecision' process (CD) on the next higher level had to resolve the situation in the mission context. HMI is a process for handling Human Machine Interface.

III.3 Goals for the third-generation vision system

In the early 1990s both in the Prometheus-project with the test vehicle VaMP and in activities of the western defense system with VaMoRs and other test vehicles from industrial partners, road (lane) running on different types of roads and performance of special maneuvers like detection of cross roads with estimation of the parameters of the intersection, and turning off to the left- and right-hand side have been developed [Mysliwetz 1990; Dickmanns and Mysliwetz 1992; Dickmanns et al. 1993; Hock 1993; Schmid and Thomanek 1993; Brüdigam 1994; Dickmanns et al. 1994; Schmid 1994; Thomanek et al. 1994; Schiehlen 1995; Behringer 1996; Müller 1996; Thomanek 1996]. Usually, these capabilities existed as separate software packages; one of the goals of the new system design was to integrate all these perceptual, decisional and control capabilities into one general framework with the potential for learning by doing. Especially saccadic gaze control, (limited) color recognition, turning off onto previously unknown cross roads, and error detection capabilities on all levels from feature detection over object tracking to situation asessment were considered essential components for more advanced vision systems.

In addition, binocular stereo vision in the near range (up to 10 to 15 m) and gaze fixation onto a moving object when the base of the camera (the own vehicle) also moves, should become newly available capabilities. In military applications with phases of driving off-road, beside ('positive') obstacles above the ground also larger (pott-) holes and ditches constitute ('negative') obstacles to be avoided; this class of obstacles with a wide variety of potential appearances should become detectable by the new system. For this reason, area-based features like linearly shaded gray values or colors as well as certain textures should be evaluated when computer performance has become sufficiently high to allow this in real-time.

Each time before a capability is activated for real-world application, the system has to check whether all components needed are available and whether a functional test yields reasonable results. The distinction between 'objects proper' *without* -, and 'subjects' *with* the capability of sensing and acting should be applied uniformly throughout all higher system levels. The simple term 'object' is used for both, when a distinction is not necessary or not yet possible. Subjects have certain classes of capabilities with respect to perception, memory, decision making, and control application to a

variety of specific sub-systems [Dickmanns 1989]. In general, they all are used in cooperation for achieving some goals; beside recognizing the environment for avoiding dangers and for moving towards some destination point, the learning of properties of other 'objects proper' and of typical behaviors of other subjects in certain situations is valuable for improving the proper decision in future situations to be encountered. Therefore, maneuvers (sequences of actions) and how they evolve over time as well as when they are used, are important knowledge elements.

For own actions, it is not the abstract notion of absolute trajectories in 3-D space which is of major importance, but the combination of 1.) feed-forward control time histories to be applied to relevant actuators and of 2.) feedback control laws for deriving superimposed control components from relative state deviations observed concurrently. The control variables are the only means in dynamical systems that allow taking influence on the further evolution of state variables over time; therefore, these should be at the center for counteracting unforeseeable perturbations that do not correspond to the model underlying the feed-forward control time histories chosen.

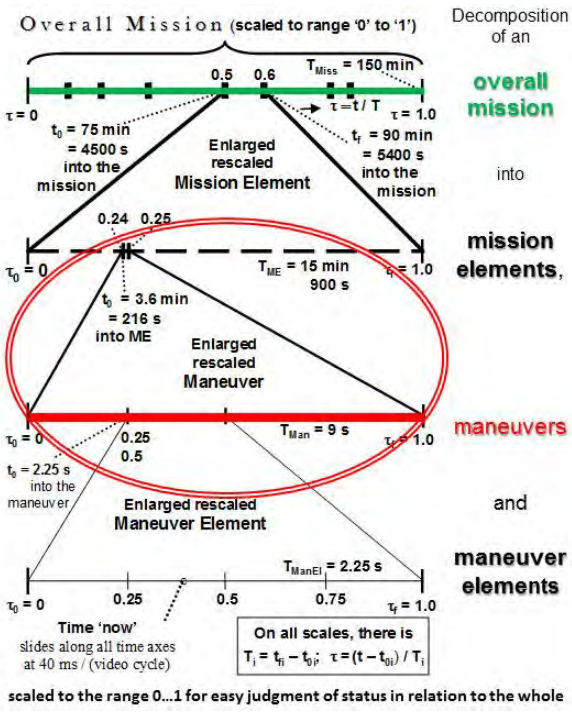

Figure III.3: Decomposition of a mission into mission elements, maneuvers and maneuver elements, all scaled to the range 0 - 1.

In real-world applications to both ground and air vehicles perturbations are frequent and mostly unforeseeable, in general; therefore, reasonable nominal conditions have to be chosen for the so called 'nominal trajectory', and proper linear feedback control laws are determined in parallel to the resulting 'nominal feed-forward control time histories', usually remembered by a few parameters only (over time); no entire state histories of the (optimal) trajectory need be recalled [Dickmanns 2007, sections 3.4 and 14.6]. Therefore, capabilities for handling temporally extended processes have to be essential knowledge elements of smart subjects. Since missions to be performed, usually, are made up of sequences of typical general mission elements like cruise phases, intermediate maneuvers for transitions, and obstacle avoidance (see Fig. III.3), it is efficient to store in the corrsponding knowledge base 1.) typical maneuver elements, 2.) maneuvers built from these by situation-dependent sets of parameters, and 3.) cruise phases with proper feedback control laws for counteracting perturbations and for achieving the next point of transition in the mission context.

Before details of the resulting overall system design for the joint German / American project **AutoNav** will be discussed in section III.5, a brief look at the past developments in the US is given (see also Part I.2).

III.4 The joint US - German project 'AutoNav'

When high-speed driving on well-structured roads using the "4-D"-approach of UniBw Munich in conjunction with inexpensive standard computer hardware was demonstrated, the interest in the slow-driving test vehicles of DARPA with massively parallel processors vanished. Especially the

keynote talk [Dickmanns 1989] at IJCAI in Detroit is reported to have brought a change in the US-policy of developing autonomous ground vehicles. The crawling vehicles were given up and the new '**H**ighly **M**aneuverable **M**anned **W**heeled **V**ehicle' (**HMMWV**) was selected as base for future activities in autonomous driving. The lead position for funding was transferred from DARPA to the **A**rmy **R**esearch **L**aboratory (ARL) at Aberdeen, Delaware. A survey on the US-effort on Strategic Computing may be found in [Roland and Shiman 2002]. Beside the former players the 'Intelligent Systems Division' at the Manufacturing Engineering Laboratory of the '**N**ational **I**nstitute of **S**tandards and **T**echnology' (NIST) under James Albus also became involved. They received an HMMWV similar to that of several other universities (Fig. III.4). The basic equipment with **L**aser **R**ange **F**inders (LRF) and cameras was available to, but not used by all. Besides driving on networks of minor roads, also driving cross country and avoiding both positive and negative obstacles (like ditches) was the major new aspect since the mid-1990s. The regular demonstrations were dubbed 'Demo I (tele-operated), Demo II, and III' (autonomous vehicles: HMMWV's and **Ex**perimental **U**nmanned **V**ehicles XUV).

Figure III.4: NIST Robotic HMMWV funded by Army Research Lab.

The development of special hardware for massively parallel processing of image sequences continued for a while, though at a much reduced level. The very fast development of general-purpose hardware for digital microprocessors and of communication networks on the general market allowed arranging standard hardware for real-time understanding of image sequences using the 4-D approach. This had been demonstrated in 1994 by the transputer systems in the European Prometheus project (see Part II). When the second-generation of European transputer systems failed to appear in the mid-1990s, the commercial-off-the-shelf (COTS) US Power-PC by Motorola with tenfold the computing power of the early transputers allowed switching from half to full video frame rate (of 25 Hz, 40 ms cycle time). This development has led to a joint project dubbed '**AutoNav**' in the framework of a '**M**emorandum **o**f **U**nderstanding' (MoU) between the US- and the German departments of defense. On the US-side it became part of Demo III.

3 The joint US-German project 'AutoNav'

The Army Research Lab. (ARL, Delaware) coordinated the activities of the partners: 1. A newly formed robotic group at General Dynamics Robotic Systems, responsible for industrial system integration and testing, 2. the stereo-vision group at the former Sarnoff Research Center (SRI-Princeton) under Peter Burt, responsible for the first to be developed real-time full-frame-video stereo system based on specifically designed hardware, and 3. the NIST-group in charge of overall system architecture including both scene understanding, decision making in the overall strategic military situation, and of control application; they also had to contribute the Laser Range Finder (LRF), a standard component on the US-side. All of this activity is included in the book [Albus and Meystel 2001].

On the German side, the industrial partner was Dornier System GmbH, Friedrichshafen, in charge of equipping the tracked vehicle Wiesel_2 (a recent prototype with digital electronics) with additional sensors and actuators for autonomous driving. UniBw Munich was responsible for

integrating the first real-time vision system based on 'Commercial-Off-The-Shelf' (COTS) components: standard industry-PC-microprocessor components out of the Intel Pentium series with 0.3 to 0.7 GHz clock rate and on a newly available communication network. This became the third generation vision system of VaMoRs. Dornier GmbH was responsible for the transfer onto the Wiesel_2.

Figure III.5: German tracked test vehicle Wiesel_2 in a clockwise sequence of images demonstrating obstacle avoidance by evading off the dirt road and re-entering it after passing.

Fig. III.5 shows the vehicle while performing autonomous obstacle avoidance on a dirt road; it leaves the road and enters back onto it after passing the (symbolic) obstacle made up of light (but difficult to detect) wooden structures. The upper left image shows the horizontal search paths for detecting dirt roads with jagged boundaries and regions of grass on the road by especially large oblique edge masks. Only by introducing higher level knowledge early in the phase of edge feature extraction (very flexible by parameterized operators) these difficulties could be handled [Behringer 1996].

The project nominally ran from 1997 till 2001, the official retirement date of E. D. Dickmanns; however, the stereo subsystem of the American partners capable of full-frame real-time image

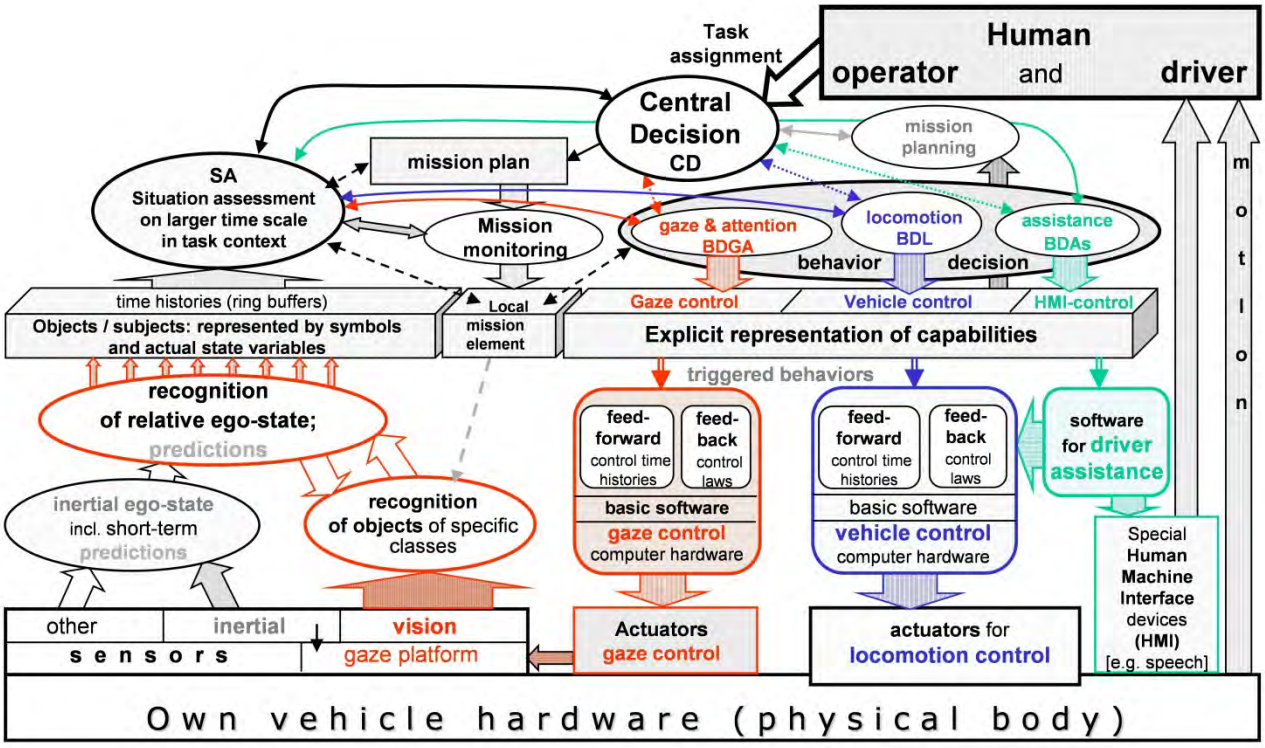

Figure III.6: Coarsely resolved system architecture of an EMS-vision system with gaze control and inertial gaze stabilization. A mission is performed by situation-dependent activation of behavioral capabilities, not by a pre-specified sequence of actions or trajectory segments.

sequence evaluation in 2001 consisted of a separate computer system of about thirty liters in volume. Therefore, a prolongation till 2003 has been arranged. In this year, by DARPA-funding, the volume of the stereo system had been shrunk in size to fit onto one European-standard-size PC-board and worked in one of the four PC-systems making up the third-generation vision system of UniBwM, dubbed EMS-vision system (see Fig. III.1).

III.5 The general concept of the EMS-architecture

All actions of the behavior decision system with the three subsystems for 1.) gaze & attention (**BDGA** in red), 2.) own locomotion (**BDL** in blue), and 3.) driver assistance (**BDAs** in green) in the upper right corner of Fig. III.6 are based on sensor data (lower left corner of the figure) and their interpretation by hypothesizing objects and subjects and estimating their relative state (left part of vertically central beam {or bar} for knowledge representation). A variable number of n '**objects proper**' and m **subjects** are detected and tracked in parallel. Their states relative to the also estimated ego-state in connection with the mission element actually under performance (central piece of the 'knowledge bar') yield the actual situation; taking the capabilities of the corresponding subsystems (right part of the knowledge bar) into account, the actual behavioral outputs are triggered, respectively confirmed (red, blue, and green downward arrows right).

The higher decision levels do not compute the actually valid control output; each subsystem on the lower central level (rounded squares) has both feed-forward and feedback control output capabilities available that allow fast responses avoiding the delay times through the upper levels. For this reason special measurement data needed are directly fed into these subsystems (not shown in Fig. III.6).

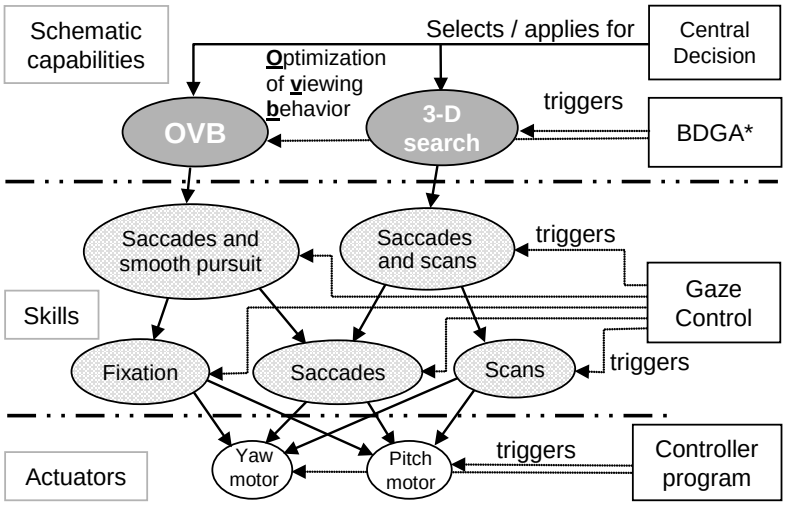

Figure III.8: Network of perceptual capabilities with MarVEye structured on three levels in order to show dependencies; before activation, all levels are checked for complete availability (after [Pellkofer 2003]).

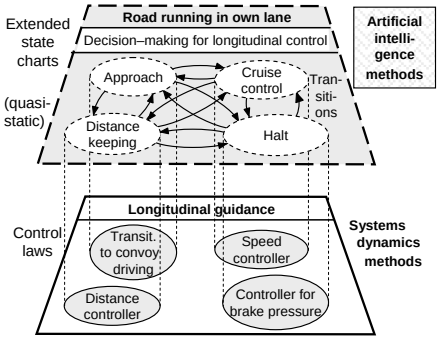

Figure III.7: Dual representation of behavioral modes: 1. Decision level (dashed), quasi-static AI-methods, extended state charts [Harel 1987] with conditions for transitions between modes. 2. Realization on (embedded, distributed) processors close to the actuators through feed-forward and feedback control laws [Maurer 2000; Siedersberger 2000, 2004].

Fig. III.7 gives a more detailed example of this dual representation of behavioral modes. Most vehicles today already have special processors for addressing the physical actuators. Therefore, the split into two levels is more convenient; in addition, minimal reaction time (least delays) improves system behavior, in general. Abstract behavioral capabilities on the higher levels may be linked to the physical actuators by so called capability networks for

visualizing the dependencies. They easily allow checking all components needed for realizing a specific mode of operation. Fig. III.8 shows the case for two-degree-of-freedom (2-*dof*) gaze control by a yaw and a pitch motor. BDGA means Behavior Decision for Gaze and Attention. The arrows downward show the skills and actuators needed for realization of schematic capabilities.

A somewhat more complicated network results for capabilities in locomotion of ground vehicles; there are three actuators involved: For braking, for throttle activation, and for turning the steer wheel. The 3-*dof* motion possible on a plane allows several more skills to be learned or taught (see Fig. III.9). It may be seen that most capabilities on the upper level require access to both lateral and longitudinal control. On the medium level of skills we find the maneuver elements needed for performing an entire maneuver. For example, the maneuver '**Avoid obstacle**' (center top) may need a) the skills 'Keep distance', which in turn may use 'Decelerate' either by control output a_1) to the throttle (reduce actual setting), or a_2) to the brakes; in some cases it may be necessary to b) accelerate by throttle activation for obstacle avoidance. If lateral deviation is needed, path c) is activated that allows the control **steering rate** to be activated to c_1) either achieve a d esired steer angle λ_{com} , or c_2) the steer angle zero for going straight ahead. If λ_{com} is larger than a p reset maximal value (e.g. $\lambda_{max(V)}$) for avoiding excessive lateral acceleration, the phase c_1 is terminated at $\lambda_{max(V)}$ and a circular arc c_3) with constant radius corresponding to $\lambda_{max(V)}$ is inserted until the next maneuver element will lead to the desired heading change. In the case of a t urn-off maneuver onto a crossroad with 90° intersection angle, the next maneuver element has to be c_2 for driving straight ahead in the new direction of the cross road. Since during this maneuver element also a certain change in heading angle $\Delta\Psi_{(\lambda\text{-dot})}$ will be achieved, this maneuver element c_2 has to be started at $(90 - \Delta\Psi_{(\lambda\text{-dot})})$. Shortly before the new driving direction is achieved, a switch to the new mode of road running (lane keeping in the cross road) may be done to counteract any errors that might have accumulated during the turn-off maneuver. The distribution of these activities among the two levels usually is a point of discussion between the AI- and the engineering people.

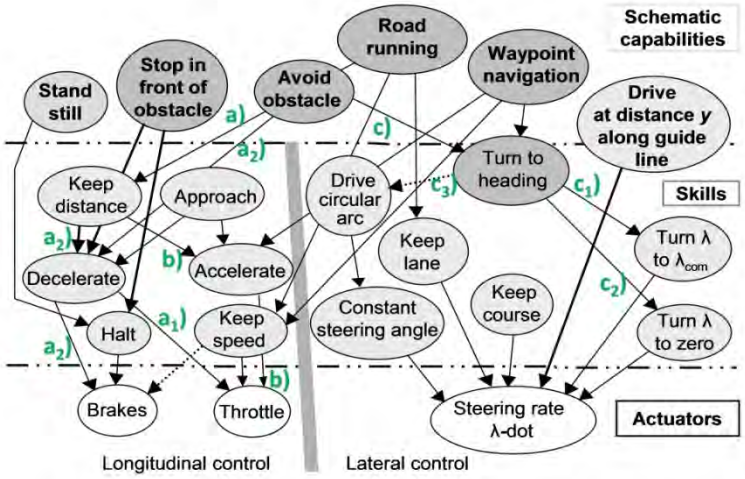

Figure III.9: Network of behavioral capabilities of a road vehicle: Longitudinal and lateral control is fully separated only on the hardware level with three actuators; many basic skills are realized by diverse parameterized feed-forward and feedback control schemes. On the upper level, abstract schematic capabilities as triggered from "central decision" (CD) are shown [Maurer 2000, Siedersberger 2004]

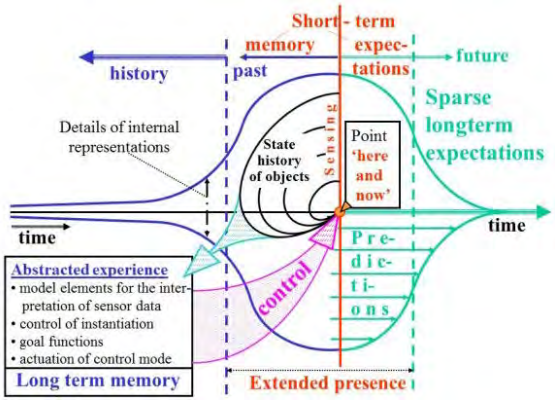

Figure III.10: Volume of representational details over time relative to the point 'now'.

These examples show that the development of consistent and yet flexible networks of capabilities

for mission performance and their activation is rather involved. During active use, the development of situations has to be tracked in order to check whether the mode running may be continued or whether new aspects require a ch ange. Figure III.10 visualizes the steadily changing internal representation around the point 'here and now': The number of objects detected and tracked as well as their relative states, and the intentions of subjects hypothesized in the environment actually perceived yield the situation determining the background for behavioral decisions. These also have to consider the recent past in short-term memory as well as the future to be expected from spatiotemporal models stored in the long-term memory. Experience made steadily should be used to improve both these models and knowledge about the situations, when to use them.

A more precise example of this capability-based approach to mission performance is given in Figure III.11. 'Central Decision' (CD, top) is the main agent; it c alls the routine for mission planning (MP) with the indication of all side constraints to be observed (top left). Exploiting its knowledge about the behavioral capabilities available to the system, MP delivers back a sequential list of mission elements, each of which can be performed by a special set of capabilities; conditions

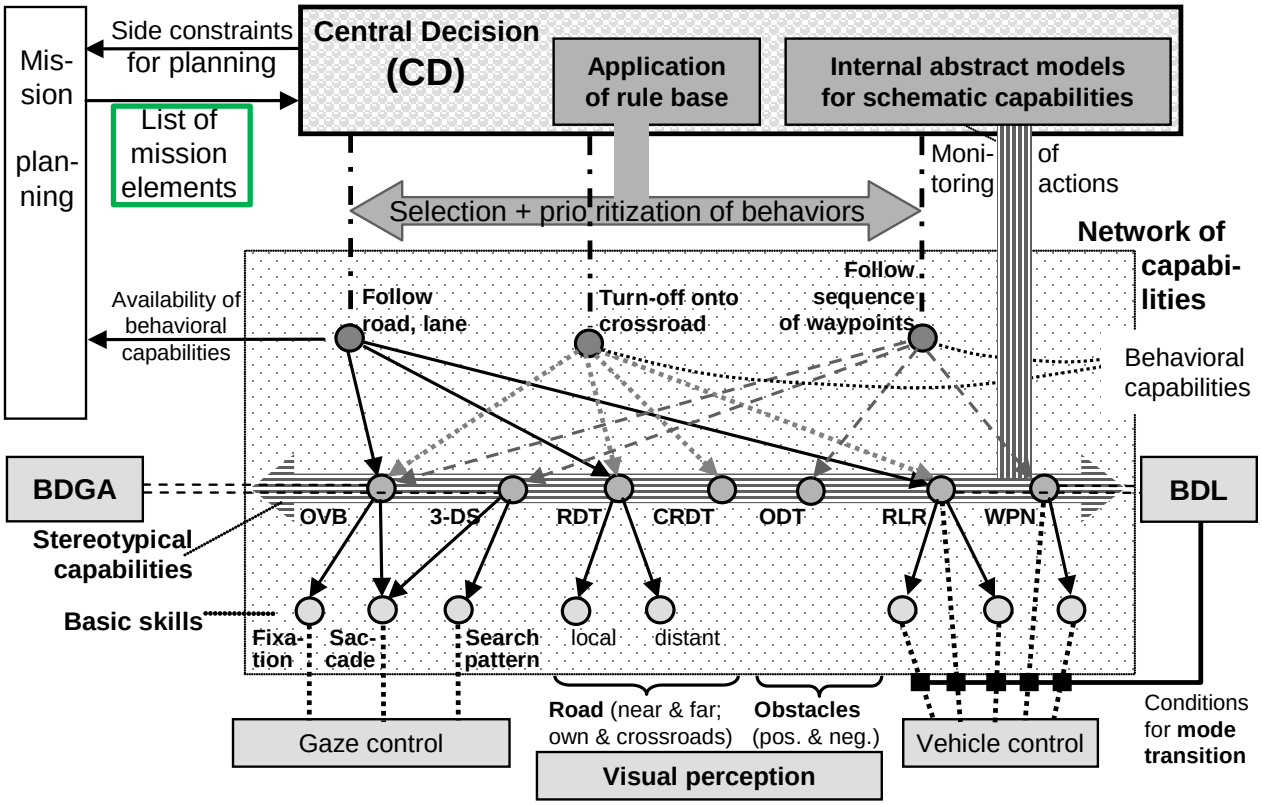

Figure III.11: Activation, prioritization, and monitoring of stereotypical capabilities by Central Decision (CD) exploiting the capability network and special decision units for Gaze and Attention (BDGA) and for Locomotion (BDL). Perceptual capabilities for roads and obstacles are provided by specialists for certain classes of objects from the image stream delivered (not detailed here). [Gregor *et al*. 2002, Gregor 2002, Pellkofer 2003, Siedersberger 2004]

***Legend** for **B**ehavior **D**ecision (central horizontal bar **BDGA** to **BDL**): **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 (potholes and ditches = negative obstacles); **RLR** = Road or Lane Running (lateral guidance at appropriate speed); **WPN** = WayPoint Navigation when driving off-road (based on GPS).*

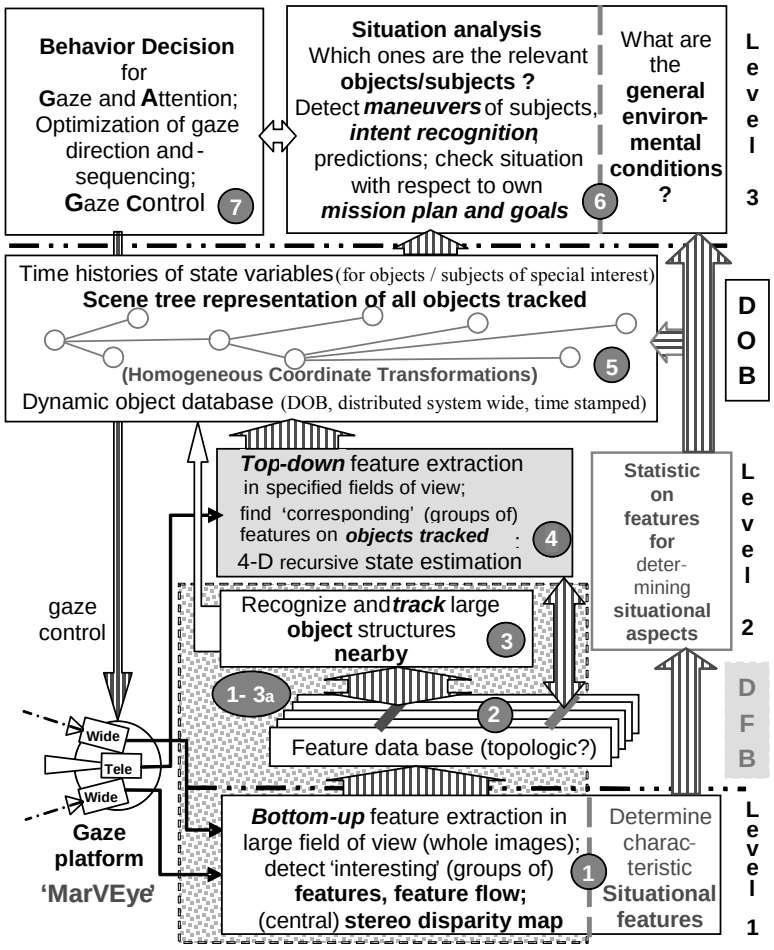

Figure III.12: Three knowledge levels for efficient visual scene analysis and decision making in dynamic vision. The scene tree (5, DOB) is filled from the bottom by hypothesis generation and used from the top for decision making.

for transition between mission elements are part of the list. A rule base on level CD controls the selection of capabilities to be used actually. According to Fig. III.7, activation of capabilities and behaviors is performed distributed in the system (lower part of the figure and right-hand part of Fig. III.6). A differently structured visualization of the data flow between the three levels for knowledge representation of: 1.) features, 2.) objects & subjects, and 3.) situations is shown in Fig. III.12. On the lowest level (1) features are extracted from every wide-angle image purely bottom-up without regard to time; this is essential for detecting new objects over time. On the right-hand side over all levels a special path is shown for features characterizing the general conditions (like overall light intensity, fog, rain, or snow fall).

All these features may be stored in a dynamic feature base (DFB, (2)). On the object-level 2 (dark cirles (3) to (5)), hypotheses of objects / subjects are formed and tracked (4); on the one side this leads to grouping of features in (2) and of tracking them over time. On the other side, new features may be hypothesized and searched for as well as tracked in higher-resolution images (tele). Object hypotheses in the 4-D approach allow estimating both pose and velocity components of objects; best estimates of their states are stored in the Dynamic Object data Base (DOB (5)) making use of the powerful homogeneous coordinate transformations linking the entries in the 'scene tree' [Roberts 1965]. The transition from sets of features to objects reduces the amount of data needed for situation assessment by two to three orders of magnitude.

III.6 State of EMS−Vision at retirement (2001 – 2004)

The step from a collection of separate capabilities for special applications to an integrated system allowing software-controlled transitions between sets of capabilities needed for certain application domains and mission elements turned out to be much more challenging than expected. In addition, the next generation of PhD-students had to learn what was behind the different approaches chosen; 'easy improvements' proposed, sometimes had to be given up again due to rare occurrences that none-the-less disallowed using them. Main emphasis in this phase was developing the third-

generation (EMS-) vision system for VaMoRs because this was more easily transferred to the vehicles of the industrial partners on both sides of the Atlantic Ocean; the results will be discussed in Section III.6.1. A small effort was done to provide a much simpler EMS-vision system for a hybrid obstacle avoidance system with another German industrial partner using radar and dynamic vision for improved robustness. This will be discussed in Section III.6.2.

III.6.1 VaMoRs

The funding contract for the last four years 1997 till 2001 had as goal developing perceptual and behavioral capabilities for performing missions on networks of roads including minor ones with unsealed surfaces, and phases of off-road driving. In all situations both positive (above ground) and negative obstacles (ditches, potholes) should be detected and avoided. Precise maps and GPS-signals were not allowed to be a 'must'; the system should be able to orient itself using imprecise maps (or even without any map information) and lacking GPS-signals every now and then. This approach may therefore be dubbed "scout-vision".

Perceptual capabilities: Beside odometry and inertial sensing of three linear accelerations and three angular rates, the 'vehicle eye' (Fig. III.2 top) with five active cameras was the main source of data. Two parallel cameras were needed specifically for real-time full image stereo interpretation by the US-partners. The divergently looking wide-angle cameras of UniBwM are mounted at the lower side of the structure of the 'eye' (see Fig. III.13, next page, top-right in the image). The pitch angle of the eye is stabilized by direct angular rate feedback from a miniature sensor on the platform; there is no need for an expensive inertial platform. Gaze stabilization reduces the amplitudes of perturbations while braking or driving on non-flat terrain by more than one order of magnitude. Gaze fixation is also possible, *e.g.* the nearest corner of a ditch is kept at the center of the image when driving around it. The divergent wide-angle cameras allow viewing the entire crossing when approaching it; with the central tele-camera the crossroad may be watched at distances further away from the crossing to determine the angle of intersection of the two roads [Schiehlen 1995, Rieder 2000, Luetzeler and Dickmanns 2000, Luetzeler 2002, Pellkofer 2003, von Holt 2004].

The vision system is capable of determining small and wide roads, certain types of lane markings if available, cross roads as T-junctions or full crossings, positive obstacles (above the ground) and negative ones. It was an unexpected experience that early detection of a candidate for a negative obstacle could be done better at a larger distance by area-based features in a monocular image than by the stereo system. Because of the relatively low elevation of the cameras above the ground (~1.8 m) a narrow ditch (even a deep one) shows only a small stripe of its almost vertical remote side slope. Since these regions, usually, have a color or gray value different from the almost horizontal surface with grass or plants, they are detected easily by corresponding operators. The stereo system concentrates on small features and was disturbed by single grasses in front of the ditch crossing the ditch region in the image. During approach the 3-D structure of the ditch becomes more pronounced, and stereo vision yields more precise data [Siedersberger et al. 2001].

Behavioral capabilities: Road and lane running are the basic capabilities of vehicles for locomotion in networks of roads. Lane changing and obstacle avoidance by braking or lateral maneuvering are other standard ones. On highways, all navigation has to be done by proper lane selection; on standard roads, turning-off to the right- or left-hand side onto cross roads are the corresponding maneuvers. The more involved one is to the side where oncoming traffic has to be crossed (left-hand side for driving on the right); in this case (with no regulation by traffic lights), large look-ahead distances are requested for detection and state estimation of oncoming vehicles. Stopping and waiting may be the correct behavior required.

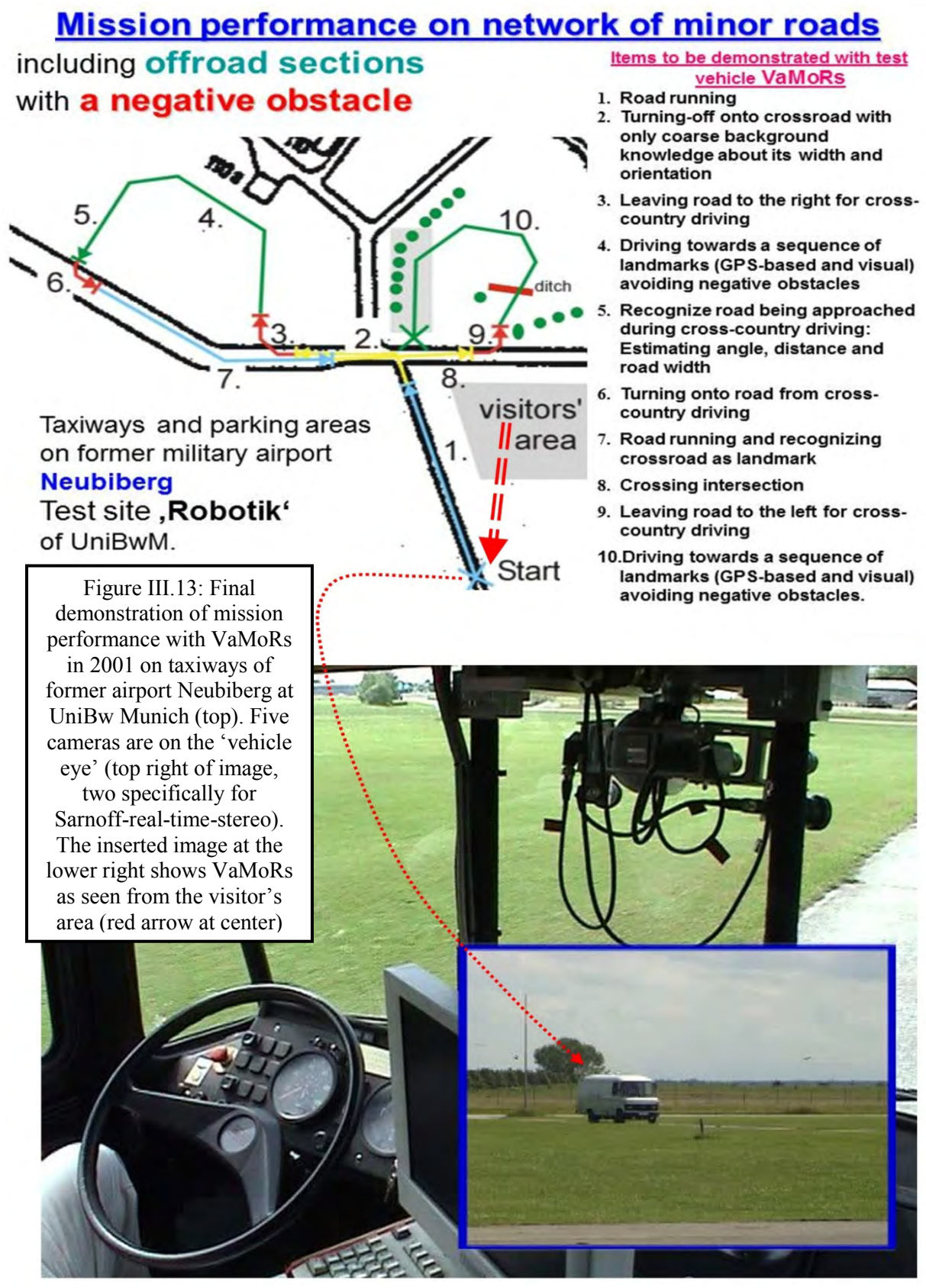

Figure III.13: Final demonstration of mission performance with VaMoRs in 2001 on taxiways of former airport Neubiberg at UniBw Munich (top). Five cameras are on the ‘vehicle eye’ (top right of image, two specifically for Sarnoff-real-time-stereo). The inserted image at the lower right shows VaMoRs as seen from the visitor’s area (red arrow at center)

For making the transition from road running to driving off-road (or vice versa) the flatness of the ground has to be recognized and the type of driving has to be adapted. Usually, there will be slopes of different types and steepness involved that have to be negotiated. Most essential is the decision whether a transition is too dangerous or impossible. The top-half of Fig. III.13 shows the sequence of mission elements 1 to 10 of the demonstration; it includes points of transition between driving on roads and off-road at locations 3, 6, and 9. When an obstacle is encountered, the decision has to be made whether to stop or at which side to drive around it [Mandelbaum et al. 1998, Siedersberger et al. 2001]. A video showing the demonstration of the shortened mission with the elements 1, 8, 9, 10 may be found in www.dyna-vision.de under 'Mission performance'.

Mission decomposition into a list of consecutive elements: At the stage of development achieved, mission performance has to be done in a mixed mode partly by the human operator and partly autonomously. The part specified by the human operator was the following: Mission Element **ME1**: At the 'Start' find the road with its parameters 'width' and 'lane markings'; start driving till the next crossroad, which you have to follow to the left. **ME2:** Estimate the angle of intersection, the distance to the (virtual center-) point of intersection, and the width of the crossroad. Properly start the turn-off maneuver according to the parameters found. **ME3:** Drive along new road, disregarding all crossings, until the GPS-waypoint 3 is reached. **ME4:** At this point, leave the road to the right to go cross-country via a sequence of GPS-waypoints, watching out for negative obstacles (phases 4 and 5 in Fig. III.13, top left); look for a road that you will encounter under an almost right angle. **ME6:** Enter this road to the left taking care of non-planar ground. **ME7:** Follow this road disregarding all crossings (2, 8) till GPS-waypoint 9. **ME9:** Turn-off to the left taking care of non-planar ground. **ME10:** Follow a sequence of GPS-waypoints (10) watching out for positive and negative obstacles. The ditch (red bar between 9. and 10.) is detected, but when the parameter values of the estimation process become sufficiently stable, only the right-hand side of the ditch can be seen in the stereo images. From here on, the mission elements are determined fully autonomously by the system for perception and control.

Mission performance: At the start, road width is detected as small (a single lane) with no lane markings. From the tele-image and the gaze angle of the platform in yaw, the angular offset $\Delta\psi_{RV}$ between road direction and vehicle orientation is determined. If required, in a first lateral maneuver $\Delta\psi_{RV}$ is reduced to zero with VaMoRs at the center of the road; then standard road running is performed, while periodically a saccade to the left is made for detecting the crossroad in the tele-image. When detected repeatedly, short sequences of fixation phases onto the distant crossroad and the road driven are performed for determining the parameters of the crossroad and the distance to the intersection. When estimation errors become small, gaze is fixated onto the crossroad at some distance during the final approach. At a properly selected distance to the point of intersection on the own road the feed-forward control time history for the steer rate nominally leading to the turn-off desired, is activated. Starting at about 70% of this maneuver a feedback component for the lateral offset in the new road is superimposed to correct perturbations that might have occurred.

After achieving an acceptable state on the new road the turn-off maneuver is ended and pure road running is continued to the GPS-waypoint 3. Shortly before arriving there, the vertical structure of the ground for transiting to off-road driving is checked by stereo vision. If acceptable, the transition is started and 'off-road-driving with additional checking for negative obstacles' is activated. Towards the end of phase 5, the system starts searching for visual features indicative of a road in almost orthogonal direction to the path driven. When these are found repeatedly, the system again starts saccadic attention control in order to estimate the relative direction of the road and the distance to the point for entering it. At the proper distance, the maneuver for entering a road (off-

road-driving to road-entry feed-forward control time history for steering) is started keeping viewing direction fixated at a constant range on the new road. Then again, road- (or lane-) running is initiated depending on the width of the new road. In Fig. III.13 the wide road (blue track, 6 and 7) ends just before the two T-junctions from left and right occur; these have to be crossed, and at GPS-point 9 a turn-off to the left has to be done (similar to 3 to the right). In the final off-road section both positive (bushes & small trees) and negative obstacles are present at unknown locations. The ditch is about 0.8 m wide over a length of ~ 4 m (depth ~ 0.6 m). The path shown in green is the prescribed one.

After transition, the vehicle slowly starts following this path until a candidate for a ditch is detected; when the parameter values of the estimation process become sufficiently stable, only the right-hand side of the ditch can be seen in the stereo images. Therefore, the vehicle has to stop and to comprehend the extension of the ditch as well as its relative orientation by doing saccadic gaze changes; by combining the results the decision is then made. In the case given, the decision is to go around the right-hand corner of the ditch, which is now fixated by gaze control. When the direction of the left-hand front wheel passes the corner at a sufficient distance (~ 1 m), this distance is kept constant by reducing the steer angle. When the front wheel reaches the position of the corner, the steer angle is increased in the direction of the next GPS-waypoint, while gaze direction is commanded into the direction of the front wheels for detecting other obstacles in the path driven.

With the successful demonstration in 2001 of the first "mobile real-time, full frame stereo system" the **AutoNav** project was finished [Luetzeler and Dickmanns 2000, Gregor et al. 2000, Gregor and Dickmanns 2000, Hofmann et al. 2000, Pellkofer and Dickmanns 2000, Siedersberger 2000, Gregor et al. 2001a, 2001b, Siedersberger et al. 2001, Gregor et al. 2002, Gregor 2002, Luetzeler 2002]. In the German system an additional laser-system was available only for Wiesel_2. This allowed exploring the advantages of the different approaches.

With regard to system integration up to the military decision levels, the main contributions came from the American side [Albus and Meystel 2001]. They have been implemented by the industrial partners on the test vehicle XUV (Experimental Unmanned ground Vehicle) in Demo III (see Fig. III.14). As one of the consequences of this success, the US-Congress, when deciding future military funding in 2001, set as one goal that by 2015 one third of the new combat ground vehicles for the US-forces should have the ability of autonomous driving. This was an essential decision in the USA that brought the funding agency DARPA back into the business of autonomous ground vehicles.

Figure III.14: Test vehicle XUV in the Demo III program of USA.

They set an award of one million dollar for an autonomous vehicle capable of demonstrating a long-range autonomous drive in a semi-desert environment: The 'Grand Challenge' of the year 2004 stimulated quite a bit of new efforts in the US (see Part IV). However, since maintenance and service missions in secured areas became of major importance, the system design for autonomous vehicles changed: The precise geography was considered to be known, and signals from the Global Positioning System GPS were assumed to be steadily available. The main task for this type of 'autonomous driving' was to follow a sequence of GPS-waypoints and to avoid obstacles above the driving plane; the GPS-points were given rather tightly, and the trajectories through these points were prepared to contain no negative obstacles. In sharp curves, several waypoints defining the curvature were given.

III.6.2 VaMP

While the main emphasis in developing EMS-vision was on the ability to perform rather complex missions in a network of roads including off-road sections with the test vehicle VaMoRs, the second test vehicle at UniBwM was used to apply EMS-vision to an assistance system for highway driving. For improving reliability in detecting positive obstacles, a radar system has been provided by an automotive supplier company; this was to be used in connection with bifocal vision. Radar is known to miss almost no positive obstacle, but to yield relatively many false alarms depending on the type of environment.

One difficult task in designing assistance systems is to keep the driver fully attentive in the loop. The proposal made by our partner was to let the driver be fully responsible for lateral control, while longitudinal guidance (obstacle detection and avoidance) at the desired speed was to be done by the hybrid assistance system. Fig. III.15 shows the hardware components used and their interconnection in the scene tree. 6DOF means full six-degrees-of-freedom transformations exploiting homogeneous coordinates. The radar sensor is installed beneath the front license plate; it has good

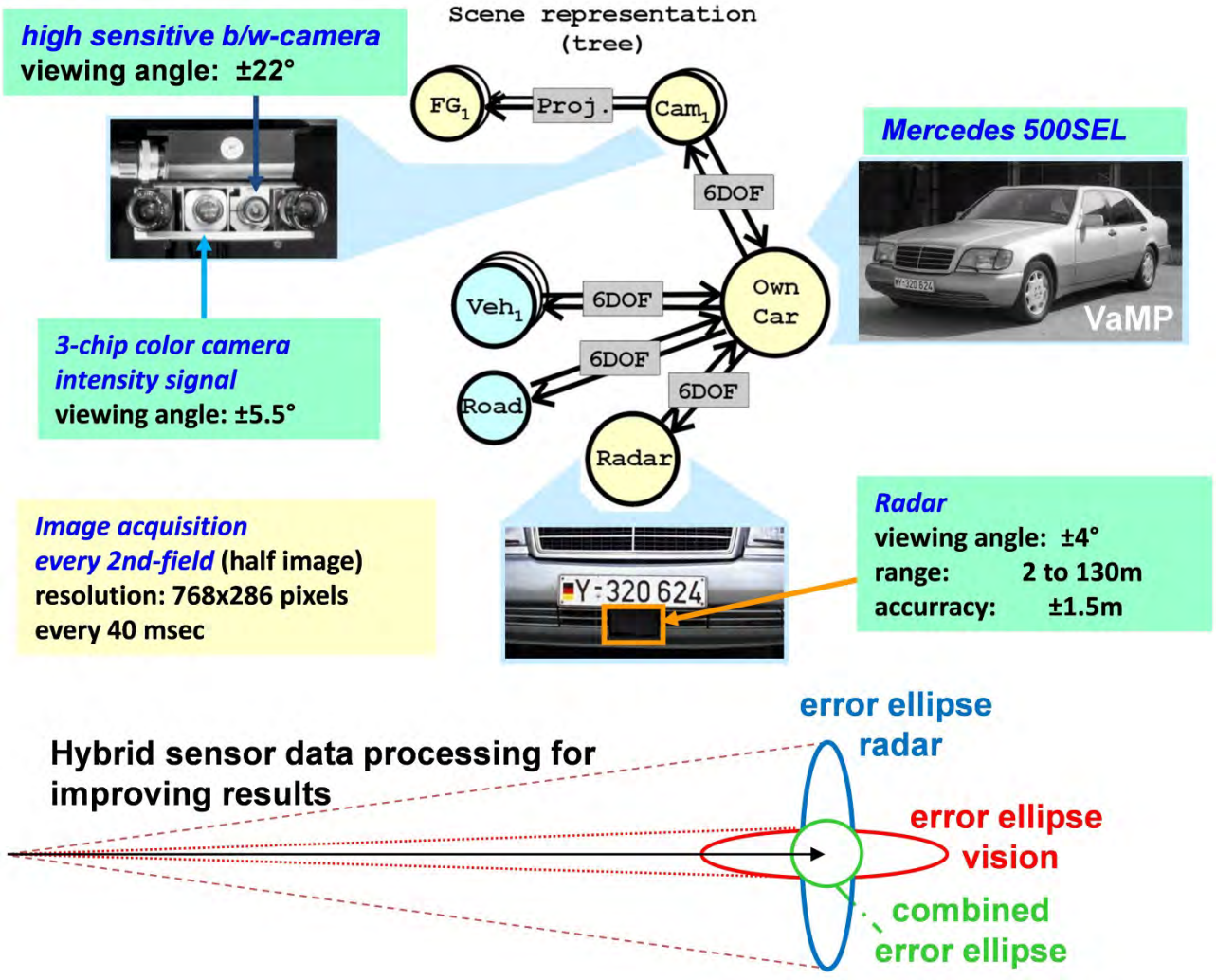

Figure III.15: EMS hybrid vision (combining bifocal vision with radar) in VaMP for reduction of false alarms and for improving accuracy (lower part with principle error ellipses); FG = frame grabber; Cam = camera (two); 6DOF = 6 degrees of freedom transformations; Veh = vehicles observed (n in parallel).

range resolution but poor resolution normal to it (blue ellipse). On the contrary, vision has good lateral resolution but poor resolution in range direction (red ellipse). Hopefully, the combined characteristic looks closer to the green circle.

Since radar misses almost no positive obstacle, the preprocessed radar output is taken as the starting point for 4-D image sequence interpretation. Objects are hypothesized at the center of each radar target found. The vision system then looks for collections of features that might justify the hypothesis of an object. For false radar alarms, usually, no corresponding stable visual features are found over a short sequence of images (e.g. 5 images meaning 0.2 seconds); the corresponding candidates are then removed from the list of vehicles tracked. For obstacles substantiated by visual features, from their arrangement in the image the lateral extension of the object is determined much more precisely than possible from radar. Fig. III.16 shows a typical road scene with objects detected

Figure III.16: Bifocal image evaluation of VaMP exploiting EMS-vision with COTS microprocessors in 2000. The red rectangle shows the reference vehicle for convoy driving. A video-clip "Hybrid Adaptive Cruise Control VaMP 2000" may be found on the web-site mentioned.

marked by rectangles; only the lower body up to about 1 m height is tracked since the upper part of cars very often show fast moving features stemming from reflections in windows that are more disturbing than helpful in object tracking. The right-hand part of the figure is the tele image, the full area of which is marked in the left-hand image by the orange rectangle. Lane markings detected are shown in both images by short red, yellow, and blue line elements and/or a line in light blue. The long horizontal yellow lines with the center part moving up and down relative to the equal lines at each side give an indication of the vertical curvature estimated of the road; the center part below the outer yellow lines means an upward curvature (increasing path angle with range) and vice versa.

Characteristic for almost all road vehicles are the dark area underneath the body and the dark tires; the latter ones show either as rectangles, when seen from behind, or as segments of ellipses, when seen from the side [Hofmann et al. 2000; Hofmann 2004] (see Fig. III.17). Only small sections of the ellipses may be seen in general due to coverage by the fenders and the body of the car; for efficient definition of search regions, trapezoidal sections have been used for lateral aspect angles. Wheels are the elements that make the distinction between a car and a large box situated on the road.

Figure III.17: Recognition of wheels of vehicles in a road scene.

III.7 Conclusion (Part III)

The third-generation dynamic vision system of UniBwM "Expectation-based, Multi-focal, Saccadic Vision" (EMS-vision) combines all capabilities previously developed separately with the 4-D approach in a unified framework. 'Capabilities' of 'subjects' (agents) are the core notions for structuring the system. Beside 3-D objects under motion appearing in 2-D image sequences, maneuvering capabilities and goal-oriented actions of subjects over time (including the own ones) are the essential knowledge elements. Early jumps to hypotheses − maybe several ones in parallel for the same collection of features − allow fast and efficient scene understanding. Pruning of invalid hypotheses is improved by checking additional features derived from the models instantiated. This seems to be similar to approaches used in biological cognitive systems.

The implementations realized in the US-German joint project 'AutoNav' with VaMoRs on the German side, and in the VaMP-project for 'Hybrid Adaptive Cruise Control' (HACC) with an industrial partner seem to validate the generality of the approach. Much work remains to be done for implementing learning capabilities on all levels. The continuing increase in computing power and communication capabilities of onboard microprocessors will yield small systems in the near future allowing the realization of complex cognitive systems at reasonable costs. The 'scout-type' vision system is considered to be economically more competitive in the long run than the ones mainly in use today, based on high-precision actual maps and localization via GPS or other procedures/devices ('confirmation-type' vision).

References (Part III)

Albus J.S., Meystel A.M. 2001: Engineering of Mind. – An Introduction to the Science of Intelligent Systems. Wiley Series on Intelligent Systems

Behringer R. 1996: Visuelle Erkennung und Interpretation des Fahrspurverlaufes durch Rechnersehen für ein autonomes Straßenfahrzeug. Diss. UniBw Munich, LRT; also: Fortschrittberichte VDI, Reihe 12, Nr. 310

Brüdigam C. 1994:Intelligente Fahrmanöver sehender autonomer Fahrzeuge in autobahn-ähnlicher Umgebung. Diss. UniBw Munich, LRT.

Dickmanns E.D. 1989: Subject-Object Discrimination in 4-D Dynamic Scene Interpretation by Machine Vision. Proc. IEEE-Workshop on Visual Motion, Newport Beach, pp 298-304

Dickmanns E.D. 1989. Detroit, Aug. 23. Invited keynote talk **IJCAI on ‚4-D approach'** (11:10 – 12:40 Uhr); several videos!

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

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

Dickmanns E.D.; Behringer R.; Dickmanns D.; Hildebrandt T.; Maurer M.; Thomanek F.; Schiehlen J. 1994: The Seeing Passenger Car 'VaMoRs-P'. In Masaki (ed): Proc. of Int. Symp. on Intelligent Vehicles '94, Paris, Oct. 1994, pp 68-73

Dickmanns E.D. 2007: Dynamic Vision for Perception and Control of Motion. Springer-Verlag, April 2007, (474 pages.)

Gregor R, Dickmanns E.D. 2000: EMS-Vision: Mission Performance on Road Networks. Proc. Int. Symp. on Intelligent Vehicles (IV'2000), Dearborn, (MI),

Gregor R., Luetzeler M., Pellkofer M., Siedersberger K.H., Dickmanns E.D. 2000: EMS-Vision: A Perceptual System for Autonomous Vehicles. Proc. Int. Symposium on Intelligent Vehicles (IV'2000), Dearborn, (MI),

Gregor R., Luetzeler M., Pellkofer M., Siedersberger K.-H., Dickmanns E.D. 2001a: A Vision System for Autonomous Ground Vehicles With a Wide Range of Maneuvering Capabilities. Internat. Workshop on Computer Vision (ICVS), Vancouver, Kanada, Juli

Gregor R., Luetzeler M., Dickmanns E.D. 2001b: EMS-Vision: Combining on- and off-road driving. Proc. SPIE Conf. on Unmanned Ground Vehicle Technology III, AeroSense '01, Orlando (FL), April 16-17

Gregor, R., Luetzeler, M., Pellkofer, M., Siedersberger, K.H. and Dickmanns, E.D. 2002: EMS-Vision: A Perceptual System for Autonomous Vehicles. IEEE Trans. on Intelligent Transportation Systems, Vol.3, No.1, March, pp. 48 – 59

Gregor R. 2002: Faehigkeiten zur Missionsdurchfuehrung und Landmarkennavigation. Diss. UniBw Munich, LRT

Harel D.1987: State charts: A Visual Formalism for Complex Systems. Science of Computer Programming, 8: 231–274

Hillis W.D. 1992 (6th printing): The Connection Machine. MIT Press, Cambridge, MA

Hock C. 1993: Wissensbasierte Fahrzeugfuehrung mit Landmarken fuer autonome Roboter. Diss. UniBw Munich, LRT

Hofmann U., Rieder A., Dickmanns E.D 2000: EMS-Vision: An Application to Intelligent Cruise Control for High Speed Roads. Proc. Int. Symp. on Intelligent Vehicles (IV'2000), Dearborn, (MI)

Hofmann, U. 2004: Zur visuellen Umfeldwahrnehmung autonomer Fahrzeuge.: Diss. UniBw Munich, LRT

Luetzeler, M., Dickmanns, E.D 2000: EMS-Vision: Recognition of Intersections on Unmarked Road Networks. Proc. Int. Symp. on Intelligent Vehicles (IV'2000), Dearborn, (MI).

Luetzeler M. 2002: Fahrbahnerkennung zum Manoevrieren auf Wegenetzen mit aktivem Sehen. Diss. UniBw Munich, LRT

Mandelbaum R., Hansen M., Burt P., Baten S. 1998: Vision for Autonomous Mobility: Image Processing on the VFE-200. In: IEEE International Symposium on ISIC, CIRA and ISAS

Maurer M. 2000: Flexible Automatisierung von Straßenfahrzeugen mit Rechnersehen. Diss. UniBw Munich, LRT

Mueller N. 1996: Autonomes Manoevrieren und Navigieren mit einem sehenden Fahrzeug. Diss. UniBw Munich, LRT

Mysliwetz B. 1990: Parallelrechnerbasierte Bildfolgeninterpretation zur autonomen Fahrzeugfuehrung. Diss. UniBw Munich, LRT.

Pellkofer, M., D ickmanns, E.D. 2000: EMS-Vision: Gaze Control in Autonomous Vehicles. Proc. Int. Symp. on Intelligent Vehicles (IV'2000), Dearborn, (MI)

Pellkofer M. 2003: Verhaltensentscheidung fuer autonome Fahrzeuge mit Blickrichtungssteuerung. Diss. UniBw Munich, LRT

Rieder A. 2000: Fahrzeuge sehen – Multisensorielle Fahrzeugerkennung in einem verteilten Rechnersystem fuer autonome Fahrzeuge. Diss. UniBw Munich, LRT

Roberts L.G. 1965: Homogeneous matrix representation and manipulation of n-dimensional constructs. MS-1405, Lincoln Laboratory, MIT

Roland A., Shiman P. 2002: Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993. MIT Press

Schiehlen J. 1995: Kameraplattformen fuer aktiv sehende Fahrzeuge. Diss., UniBw Munich, LRT. Also as Fortschrittsberichte VDI Verlag, Reihe 8, Nr. 514

Schmid M., Thomanek F. 1993: Real-time detection and recognition of vehicles for an autonomous guidance and control system. Pattern Recognition and Image Analysis 3(3): 377–380

Schmid M. 1994: 3-D-Erkennung von Fahrzeugen in Echtzeit aus monokularen Bildfolgen. Diss. UniBw Munich, LRT.

Siedersberger, K.-H. 2000: EMS-Vision: Enhanced Abilities for Locomotion. Proc. Int. Symp. on Intelligent Vehicles (IV'2000), Dearborn, (MI)

Siedersberger K.-H., Pellkofer M., Luetzeler M., Dickmanns E.D., Rieder A., Mandelbaum R., L. Bogoni: Combining EMS-Vision and Horopter Stereo for Obstacle Avoidance of Autonomous Vehicles. Internat. Workshop on Computer Vision (ICVS), Vancouver, Kanada, Juli 2001

Siedersberger K.H. 2004: Komponenten zur automatischen Fahrzeugfuehrung in sehenden (semi-) autonomen Fahrzeugen. Diss. UniBw Munich, LRT.

Thomanek F, Dickmanns ED, Dickmanns D 1994: Multiple Object Recognition and Scene Interpretation for Autonomous Road Vehicle Guidance. Proc. Int. Symp. on Intell. Vehicles'94, Paris, Oct. 1994, pp231 - 236.

Thomanek F. 1996: Visuelle Erkennung und Zustandsschaetzung von mehreren Straßenfahrzeugen zur autonomen Fahrzeugfuehrung. Diss. UniBw Munich, LRT.

Von Holt V. 2004: Integrale Multisensorielle Fahrumgebungserfassung nach dem 4-D Ansatz. Diss. UniBw Munich, LRT.

www.dyna-vision.de : Website with many details on the development of real-time dynamic machine vision systems at UniBw Munich, including dozens of video-clips in a variety of applications.

Bibliography

Bertozzi M., Broggi A., Fascioli A. 2000: Vision-based intelligent vehicles: State of the art and perspectives. Robotics and Autonomous Systems 32: 1–16

Davis L., Kushner T.R., Le Moigne J.J., Waxman A.M. 1986: Road Boundary Detection for Autonomous Vehicle Navigation. Optical Engineering, 25(3): 409–414

Dickmanns E.D. 1986: Computer Vision in Road Vehicles – Chances and Problems. ITCS-Symposium on ‚Human Factors Technology for Next-Generation Transportation Vehicles', Amalfi, Italy, June 16-20.

Dickmanns E.D. 1987: 4-D-Dynamic Scene Analysis with Integral Spatio-Temporal Models. 4th Int. Symposium on Robotics Research, Santa Cruz. In: Bolles R.C.; Roth B. (1988). Robotics Research, MIT Press, Cambridge, pp 311-318.

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

Dickmanns E.D. 2001: Fahrzeuge lernen sehen. Broschüre (139 S.) und CD (mit ~ 40 Minuten Videoclips) zu 25 Jahren Forschung und Lehre an der UniBwM,

Dickmanns E.D 2002. Vision for ground vehicles: history and prospects. Int. J. of Vehicle Autonomous Systems (IJVAS), Vol.1, No.1, pp. 1 – 44.

" 2002: Expectation-based, Multi-focal, Saccadic (EMS) Vision for Ground Vehicle Guidance. Control Engineering Practice 10, pp.907 – 915

Kiy K.I., Klimontovich AV, Buyvolov GA (1995) Vision-based system for road following in real time. Int Conf on Advanced Robotics ICAR'95 1:115–124.

Thorpe C., Kanade T. 1986: Vision and Navigation for the CMU Navlab. In: SPIE Conf. 727 on 'Mobile Robots', Cambridge, MA

Tsugawa S., Yatabe T., Hirose T., Matsumoto S. 1979: An Automobile with Artificial Intelligence. Proc. 6th IJCAI, Tokyo: 893-895

Tsugawa S. 2008: A History of Automated Highway Systems in Japan and Future Issues. Proc. IEEE Int. Conf. on Vehicular Electronics and Safety, Columbus, OH, USA. Sept. 22-24

Weber M. 2015: Where to? A History of Autonomous Vehicles. Computer History Museum. http://www.computerhistory.org/atchm/where-to-a-history-of-autonomous-vehicles

Weems CC, Rana D, Hanson AR, Riseman EM, Shu DB, Nash JG 1990: An Overview of Architecture Research for Image Understanding at the University of Massachusetts. Proc. ICPR 1990, Vol-II, pp. 379-384.

Zimdahl W., Rackow I., Wilm T. 1986: OPTOPILOT – ein Forschungsansatz zur Spurerkennung und Spurführung bei Straßenfahrzeugen. VDI Berichte Nr. 162, pp 49-60