AN ADVANCED VISION SYSTEM FOR GROUND VEHICLES

Ernst Dieter Dickmanns

UniBw Munich, LRT, Institut fuer Systemdynamik und Flugmechanik
D-85577 Neubiberg, Germany

ABSTRACT

'**E**xpectation-based, **M**ulti-focal, **S**accadic' (EMS) vision has been developed over the last six years based on the 4-D approach to dynamic vision. It is conceived around a '**M**ulti-focal, **a**ctive / **r**eactive **V**ehicle **Eye**' (MarVEye) with active gaze control for a set of three to four conventional TV-cameras mounted fix relative to each other on a pointing platform. This arrangement allows both a wide simultaneous field of view (> ~ 100°) with a central region of overlap for stereo interpretation and high resolution in a central 'foveal' field of view from one or two tele-cameras. Perceptual and behavioral capabilities are now explicitly represented in the system for improved flexibility and growth potential. EMS-vision has been implemented in two test vehicles **VaMoRs** and **VaMP** on sets of three to four dual-processor PC plus microprocessors for hard real-time data processing. Test results on autonomous mission performance on- and off-road have been demonstrated.

1. INTRODUCTION

The basic difference between stationary vision systems and those on-board a vehicle is the fact that in the latter case the cameras do not only move tangential to the trajectory driven but are also subjected to (rotational and translational) perturbations. These may stem from non-smooth surfaces, from aerodynamic forces or from using the control devices (throttle, brakes and steering in case of road vehicles). Fast rotational motion will lead to motion blur in the image, especially under poor lighting conditions like dawn and dusk because of the extended integration time needed for accumulating light intensity. The same rotational rate will be the more detrimental to image processing the larger the focal length of the camera.

For perceiving the environment from aboard a vehicle properly, three **t**asks have to be solved in **v**ision: 1. Is there something of interest in the hemisphere into which the vehicle is moving? (later on referred to as VT1) 2. What is it that attracted attention by the previous process; where is it relative to me and how does it move? (VT2) 3. What is the likely future motion and (for subjects) intention of the object/subject tracked? (VT3). These vision tasks have to be solved by different methods and on different time scales in order to be efficient. Also the fields of view required for answering the first two questions are quite different. Question 3 may be answered more efficiently by building on the results of many specific processes answering question 2, than by resorting to image data directly.

Vision systems for a wider spectrum of tasks in ground vehicle guidance have been addresses by a few groups only. The Robotics Institute of Carnegie Mellon University (CMU) [1-6] and DaimlerChrysler Research in Stuttgart/Ulm [7-12] (together with several university groups) are the most active ones. A survey on research activities in vision for ground vehicles is given in [13, 14]; the latter one of these shows in table 2 the number of institutions working on specific tasks in road vehicle guidance by machine vision, grouped according to the most active countries. At least about 50 groups worldwide are or have been active in this field. A good survey may be obtained from looking at the proceedings of the yearly International Symposium on 'Intelligent Vehicles'xx' [15]. To the best of our knowledge, none of these groups has developed and demonstrated an integrated approach similar to EMS-vision [16-23]. A good example of conceptual work in a similar direction is [24]; however, there are fundamental differences in the approach.

The parameters for the design of EMS-vision have been derived from road traffic scenarios and typical tasks to be solved during a complex mission on a network of roads from minor dirt roads to freeways (German Autobahn).

Assuming limited speed of other objects/subjects in the environment, it is immediately clear that for avoiding collisions, only the near environment is of special interest. However, it is almost the full hemisphere, from which for example other subjects in 'Stop & Go'-traffic may move into the trajectory planned for ego-motion.

Driving oneself at high speed (say 40 m/s = 144 km/h = 90 mph) the distance for stopping the vehicle may be considerable; with one second reaction time and –5 m/s² deceleration (~ half of Earth gravity acceleration) a stopping distance of ~200 m will result. Since objects of

dimension 0.1 m in their smaller direction of extension may be harmful to vehicles, this size should be covered by 2 to 3 pixels for reliable detection. This yields an angle of about 0.16 milli-radian per pixel (mrad/pel) or an image size of about 7° according to CCIR standards (~ 760 pel/line). The requirements resulting for a 'vehicle eye' from these considerations are discussed in section 2.

It will turn out that for reasons of data economy, foveal-peripheral differentiation in the field of view (f.o.v.) and active gaze control will be advantageous, as has been found out by 'Nature' when developing the vertebrate eye over millions of years in evolution. Once a separate unit for imaging with gaze direction control is available (the eye), small relative to the body size, coupling with inertial sensing for rotational ego-motion may lead to inertial gaze stabilization, which considerably alleviates image processing onboard vehicles. - Feedback of deviations in position between the desired and the actual center of visual features, observed in image processing, allows smooth pursuit (fixation). In addition, gaze control increases visual search range (overall f.o.v.).

This rather complex type of visual sensing is way too expensive for solving a single task. However, with proper gaze control and background knowledge, many different tasks are solvable just by switching mode and by resorting to specific knowledge bases geared to the problem domain at hand. Our human vision system demonstrates the wide range of applications possible once sufficient computing power, communication bandwidth and storage capacity are available. The EMS vision system developed has been conceived as a starting point for technical vision systems with growth potential for general applications according to the availability of computational resources. As a guideline, long-term technical capabilities for vehicle guidance may develop similar to the growth process of a human individual from early childhood to grown-up proficiency after special training, for example as a test pilot.

The structure of the cognitive vision and motion control system developed is supposed to be rather general, and stable in the meantime. Considerable effort still has to go into knowledge representation for feature extraction, hypothesis generation, situation assessment and behavior decision in different domains as well as for learning and cooperative behavior. Explicit representation of perceptual and behavioral capabilities has recently been introduced into the system for improving flexibility in mission performance and more easy adaptation to new developments.

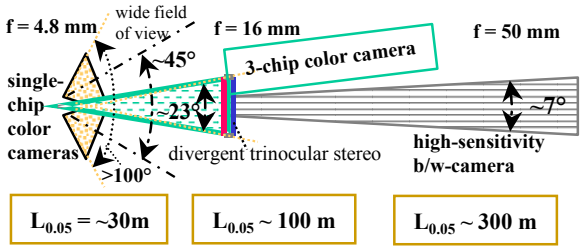
After section 2 with the requirements leading to 'MarVEye' for road vehicle guidance, integrated inertial/visual perception will be briefly discussed in section 3. Section 4 gives some answers to problem area VT3 for visual perception as discussed above; this topic is solved exploiting the object-oriented symbolic representations in the **D**ynamic **O**bject data**B**ase (DOB) by the 'scene tree', relying on both **H**omogeneous **C**oordinate **T**ransformations (HCT) and spatio-temporal models including stereotypical behavioral capabilities for subjects. Section 5 gives a brief survey on behavior decision and its two-level implementation. One hardware realization of the system is discussed in section 6, and section 7 presents some experimental results for turning off onto a crossroad in a network of minor roads without lane markings.

2. REQUIREMENTS FOR A TECHNICAL VISION SYSTEM

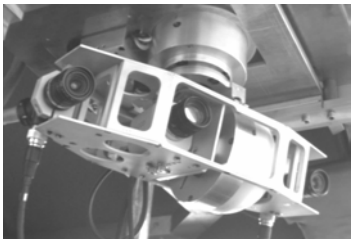
Since in road vehicle guidance human visual performance is the standard for comparison, the technical vision system has to be scaled according to this standard. Humans are able to simultaneously observe a f.o.v. ranging over more than 180° horizontally. The foveal high-resolution part of the human eye is approximately elliptical with ~ 2° horizontally and ~ 1° vertically at the center of the overall f.o.v. The eye can be turned in a very fast manner (up to several hundred °/s) by so-called saccades. An internal representation of large objects may even be obtained by 'saccading' and piecing together the details of the object over space and time. This can be done both along body contours and by saccading back and forth between feature sets, which cannot be mapped simultaneously. The framework necessary for achieving this type of performance is spatio-temporal modeling and keeping track over the own perceptual activities.

2.1. Fields of view and resolution

In road vehicle guidance on an advanced level it should be possible to look to the side (of intended motion) and to the front (actual motion) simultaneously. From this a required f.o.v. > ~ 100° is derived. Depending on the situation, this f.o.v. has to be gaze controllable to both sides. For driving on networks of minor roads with such a system, a yaw (pan) range required for the eye of about ± 70° results. This depends on the look-ahead range requested for the tele-camera looking into the crossroad shortly before the turn-off maneuver begins. It has turned out in long-distance driving experiments on German high-speed roads that for perceiving passing vehicles cutting into your lane directly in front of the own vehicle (normally not allowed, but happening rather frequently) it is advantageous to have binocular stereo vision capability available at short ranges (< ~ 10 m). At larger ranges, monocular distance estimation exploiting motion stereo in the 4-D approach has proven sufficient [25, 26]. Also for tight maneuvering in connection with parking and in 'stop & go'-traffic,



a) Fields of view and viewing ranges of MarVEye



b) Realization with 3 CCD cameras on two axis platform

Fig. 1: MarVEye system parameters (a), camera set in VaMoRs on yaw and pitch platform, large stereo base but no strong tele-camera (b).

stereo vision has its merits. For this reason, the camera configuration as given in figure 1a has been selected.

The wide f.o.v. is achieved by two cameras with their optical axes in a common plane, but under diverging angles with a region of central overlap (about 10° - 15°). In this vertical stripe, also the f.o.v. of the mild tele-camera is positioned. Vertical gaze can be controlled by pitch (tilt) of the platform head in a range of about –20° to + 30°. A ratio of 3 to 4 in focal length has proven favorable for easy recognition of objects in both images. For covering the range of focal lengths of 10 to 15 requested in traffic situations as discussed above, two tele-cameras with different tele-lenses are required.

2.2. MarVeye

In figure 1a this concept is shown with a high-sensitivity black-and-white camera carrying the strong tele-lens; this has been chosen for achieving complementary properties with respect to lighting conditions. Most experiments have been performed with the three-camera system shown figure 1b (right part of fig.1). The results given in section 7 have been obtained with this system in the 5-ton van VaMoRs.

Fig. 2: MarVEye5 with central motor-zoom camera, focal length from 4.1 to ~ 73 mm.

An advanced version is a system with a compact motor-zoom camera as shown in figure 2 (with an optical zoom range of 1:18). It has a large stereo base of about 30 cm, but is rather bulky.

Fig. 3: MarVEye6 for cars with pitch control through a mirror.

In the latest develop-ment step, the whole platform is moved only in yaw direction. As a compromise between mechanical and optical properties, all cameras are mounted fix on the yaw platform; the tele-camera is with its optical axis in the yaw axis. Vertical gaze control is done (for this camera only) by a mirror reflecting the horizontal view into the optical axis of the vertical camera (Figure 3); due to a low moment of inertia, high bandwidth in pitch can be achieved. The stereo-base has been reduced to about 15 cm and space needed for the imaging devices is small so that the system fits into the center of the windshield of a car without obscuring the driver's view.

2.3. Active gaze control

This capability is essentially a means for data economy. Since high resolution is required in areas of special interest in the real 3-D world only, the f.o.v. needed for imaging decreases with increasing distance. A three-lane driveway has a lateral extent of about 15 m. At 200 m distance, this corresponds to an angular f.o.v. of 4.3°. Allowing a factor of three for inaccurate pointing and other perturbations on the vehicle body, a f.o.v. of ~ 13° is appropriate. With ~760 pel/line according to CCIR standard, a spatial resolution of about 0.3 mrad/pel results (which corresponds to ~6 cm per pixel at 200 m distance).

With a radius of curvature of 200 m (lower limit for high-speed roads in mountainous terrain in Europe), this f.o.v. has to be shifted in pan (yaw angle) by plus or minus 30° from the road tangent direction in order to have the driveway mapped into it. This means that without active gaze control, the high-resolution f.o.v. has to be about 60° (~ 5 times the value with gaze control). Taking the effects of vertical curvature of the road surface by the

same factor into account results in a factor of ~ 25 for image data to be handled without gaze control as compared to the case with this feature. The price to be paid for this 25-fold data rate reduction is a delay time resulting from a 60°-saccade, which is in the order of magnitude of a few tenths of a second.

Requesting the same high resolution over a 100° x 45° f.o.v. (the simultaneous f.o.v. of MarVEye without gaze control) would increase the data rate by a factor of more than 40. For covering the total f.o.v. of 'MarVeye' with gaze control by sets of cameras mounted directly onto the vehicle body, an increase in data rate by about two orders of magnitude would result. In absolute terms this would mean several Gigabyte/s data rate, which is not realizable in a vehicle at present.

Beside data economy, the effects of inertial view stabilization and of smooth pursuit of a moving object with corresponding reduction of motion blur should not be underestimated. In summary, high performance machine vision should be active if it is expected to develop to a state somewhere near the capabilities of human vision.

3. VISUAL / INERTIAL PERCEPTION

Due to the high data rates, vision takes time before objects in the real world and their relative states are recognized. Starting a new hypothesis in visual recursive estimation typically requires about ten to twenty cycles (i.e. 0.4 to 0.8 seconds) before the transients have faded away. If especially in the initial phase uncertainties from angular ego-motion are superimposed (without knowing them!), recognition will be very poor. Inertial sensors are able to provide good information on ego-motion including the effects of perturbations. From these data, just by exploiting the integral relationships between accelerations, velocities and positions from general dynamical models for 3-D motion, the short-term state of the own body can be reliably determined. Low-frequency drift problems occurring with inertial integration have to be handled separately; for these purposes, the effects of time delays in visual interpretation are negligible. Therefore, inertial and visual data processing has to be done in an integrated way (like in vestibular-ocular interaction in vertebrate systems).

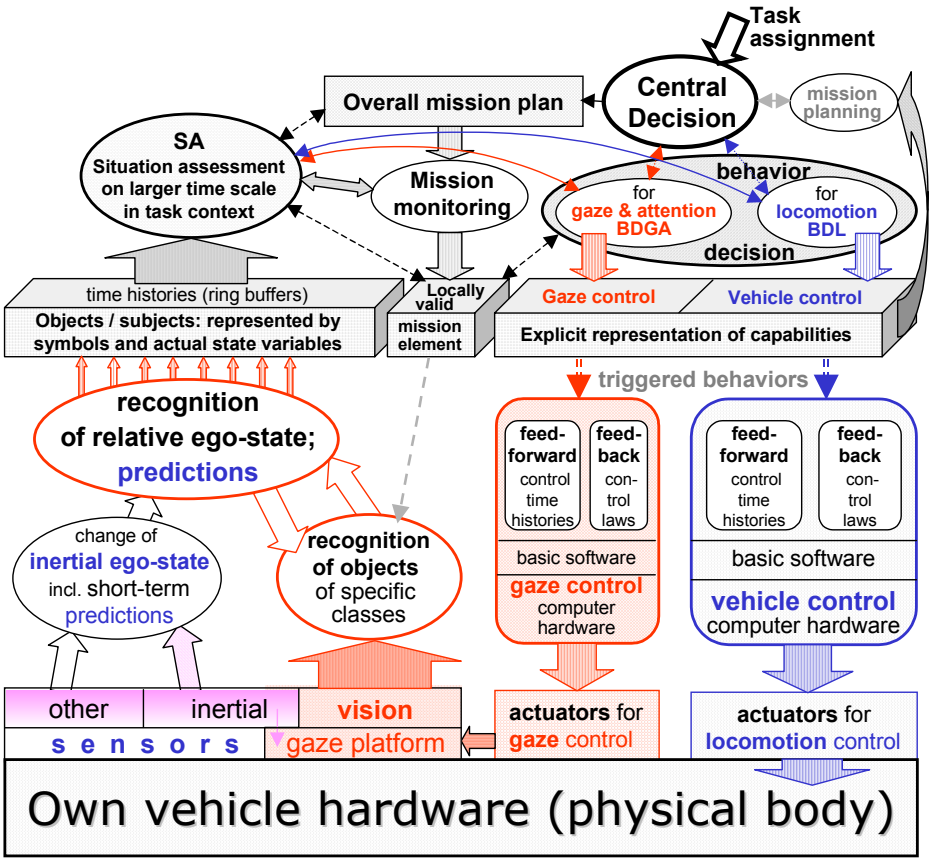


Fig. 4: Coarse resolution flowchart of an overall architecture for visual / inertial perception and intelligent control of an autonomous vehicle

3.1. Inertial sensing and data processing

This technology is well known from strap-down navigation and not detailed here. Both sets of orthonormal accelerometers (mounted close to the vehicle center of gravity) and of orthonormal rate sensors are being used. For inertial pitch stabilization of the viewing direction, a separate rate sensor is mounted on the yaw platform for gaze control. This takes care of the fact that pitching for gaze depends on the yaw angle of the platform relative to the vehicle body. [As an idealized example: for a yaw angle in gaze of -90° (to the left), vehicle roll would be a pitch angle in gaze and vehicle pitch would be a negative roll angle in gaze (which cannot be counter-acted).]

The data volume in inertial sensing is rather small. Therefore, gaze stabilization can be done at 500 Hz using a simple microprocessor, and ego-state estimation is done at 100 Hz (lower left in figure 4). Delay times are much smaller than for vision. Therefore, good estimates for own body angular pose (including the effects of perturbations) are available before higher-level interpretation of visual features starts.

For counteracting inertial drift problems, visual features from stationary objects sufficiently far away can be used. Due to fuel consumption over time, bias values for vehicle pitch cannot be assumed to be constant, in general [27]; they have to be part of the dynamical model for achieving good results. Signals from inclinometers, corrected for body accelerations have also been checked, but discarded.

3.2. Visual Perception

EMS-vision is the third-generation development step in dynamic vision [28] based on the object-oriented programming paradigm in C++. The layering according to VT1 to VT3, mentioned in the introduction, has been newly introduced as a concept. For practical reasons in a university environment and due to missing communication bandwidth for a separate 'visual feature database', VT1 and VT2 are handled in the same program module, but grouped together according to object classes. 'Roads' and 'Vehicles' are the two classes having received most of the development efforts [17, 18]. Landmarks for navigation have found some attention [29].

In the long run, it may be advantageous to have special processors for 2-D feature extraction (and thus concentrate on task VT1). The problem when looking almost parallel to a planar surface is that each image line corresponds to a different distance, and feature interpretation for physical objects has to be scaled according to distance. However, this is not uniformly true, since near objects of large vertical extension obscure the more distant ones. For such an object, all image parts covering it have approximately the same distance, irrespective of the image line. Thus, a near object may be seen directly beside a far one; their features and temporal changes have to be interpreted quite differently. This requires an object-oriented approach in a spatio-temporal framework. A zoom lens may help adjusting optimal image size to distance, so that an object further away may be seen with the same resolution as another object nearby in the wide-angle image.

3.2.1. Bottom-up object detection

Groups of horizontal and vertical edge features have proven to be good and stable indicators for objects. In most cases, the dark area underneath a vehicle also is a stable feature for object detection; exceptions are tanker trucks in combination with a low standing Sun. However, in a single image it is sometimes hard to make the decision which features belong to an object. Tracking features over time allows detecting which ones move in conjunction. Their center of gravity (c.g.) motion is the basis for solving the so-called 'Where'-problem in 2-D. The distribution of features around the c.g. allows generating hypotheses for shape, aspect conditions ('What'-problem) and rotational motion as well as the translational component in direction of the optical axis (looming). Since ambiguous object hypotheses are often possible and disambiguation can be achieved only after some period of tracking, this hypothesis generation step is often organized in conjunction with the tracking algorithm. Three to five consecutive image evaluations for consistent groups of features have proven a good basis for starting a bet on an object hypothesis with a spatio-temporal model [26].

3.2.2. Object tracking and relative state estimation

If the number of features is small in the wide-angle image and the initial range hypothesis is large (derived from the lowest part of the vehicle (feature set) under the assumption of planar ground), then a gaze shift for centering the tele-image on the object is advised. In normal road traffic, the possible aspect conditions of a vehicle are constrained due to gravity. In figure 5 left, eight classes of aspect conditions are given. For one of those (seen from rear left, as occurring when starting a passing maneuver or when the road has a curvature to the left), the distribution of characteristic features is shown in the right part. Bold letters mark the simplest set of edge features for tracking [25, 30, 31]. When area-based features (like intensity shading, color or texture) are available, the other elements imaged from the car body may alleviate recognition and tracking. The rear group of lights carries even additional information when used as

stopping light or direction indicator (dual or single intensity changes left and right). Processing power becoming available now will allow using this information in the near future.

Recursive estimation based on spatio-temporal models for shape and motion has been pioneered by our group [30] and has been generally adopted for relative state estimation in vision by the research and industrial community since. In many of the industrial application, active ranging by radar or lidar is used. With high resolution images and modeling according to figure 5, pure vision systems will probably be sufficient for normal weather conditions (as it is for humans!).

Solutions to VT2 are obtained for single objects under observation; however, by different instantiations in different image areas, results for many objects in parallel may be obtained. About half a dozen objects in three lanes in each hemisphere have been tracked in [25] already. This may be sufficient for many applications also today.

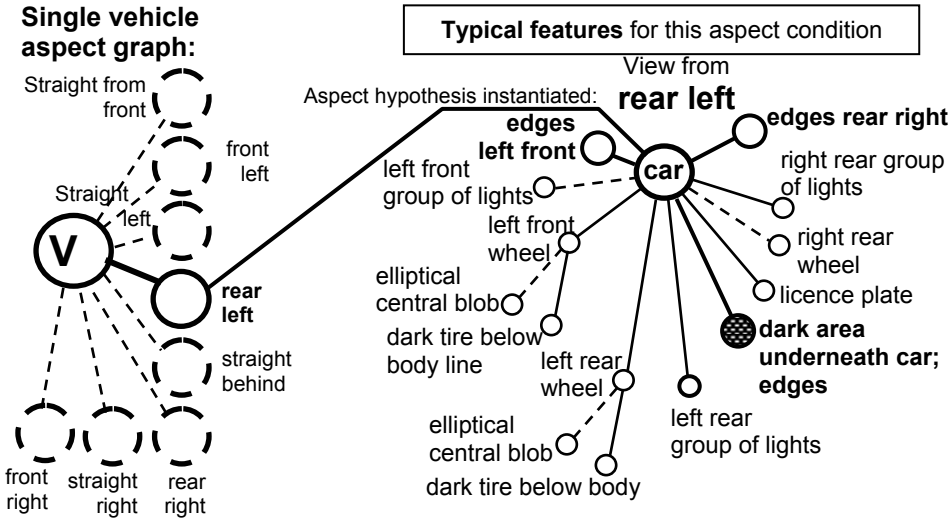


Fig. 5: Aspect graph for a vehicle V (left part), and visualization of typical (image) features for a car seen from rear left (like in passing).

*3.2.3. **D**ynamic **O**bject data**B**ase (DOB)*

In order to be able to evaluate what is happening in the environment, the different parallel results from step VT2 have to be analyzed in conjunction. As basis for this step, the best estimates for the relative states of all objects observed are collected in a database updated at video rate (see left center part of figure 4, above). For some variables, a ring buffer may be specified for tracking the last n values of state variables (their sampled time history).

As general framework for this representation, **H**omogeneous **C**oordinate **T**ransformations (HCT) in combination with a so-called 'scene tree' have been adopted [32]. Contrary to computer graphics, where all transformation variables are known beforehand, these variables are the unknowns in vision and have to be iterated in order to match model predictions and images observed. For this reason, the Jacobian matrix of first-order relations between visual features on the physical bodies (in 3-D real space) and features in the image planes have to be determined. They have to be computed for each object – sensor pair and allow spatial interpretation of objects without inverting perspective projection. A least-squares model-fit in space and time does the job efficiently. The object-oriented scene tree serves for representing the spatial distribution of objects / subjects. Subjects are objects with the capability of sensing and ego-motion control.

Figure 6 shows the scene tree for driving on a German Autobahn with three cameras mounted fix on a yaw platform (upper left part). Pitch angle control for the tele-image is done through a mirror for just the tele-camera. Each node in the figure corresponds to an object (movable object part) or to a virtual frame of reference. Each edge in the tree represents an HCT with the number of degrees of freedom (dof) indicated.

For proper road representation in generic form, the central part (vertically in figure 6 within the waved brackets) showing details of the driveway and of sub-

objects along it, may be exchanged depending on the type of road encountered. Freeways, state roads, networks of minor roads in the countryside and urban roads with many different kinds of objects require different generic representations in order to be efficient. Not all sub-objects need to be present. Excluding those that cannot be encountered in certain domains reduces the set of potential object classes for hypothesis generation and thus increases efficiency.

3.2.4. Trajectory-, maneuver- and intent recognition

For reasonable behavioral decisions in traffic it is not sufficient to know the ***actual*** state of isolated single objects, but for a defensive style of driving it is necessary to have a deeper understanding of the traffic situation in the environment. Other subjects may react depending on what *they* see and believe the situation to be. This judgement of the interdependence of motion developing can best be derived from looking at a coarser time scale, taking typical maneuvers of traffic participants into account. This can only be achieved when the corresponding maneuvers are represented in generic form together with the situation when they should be applied.

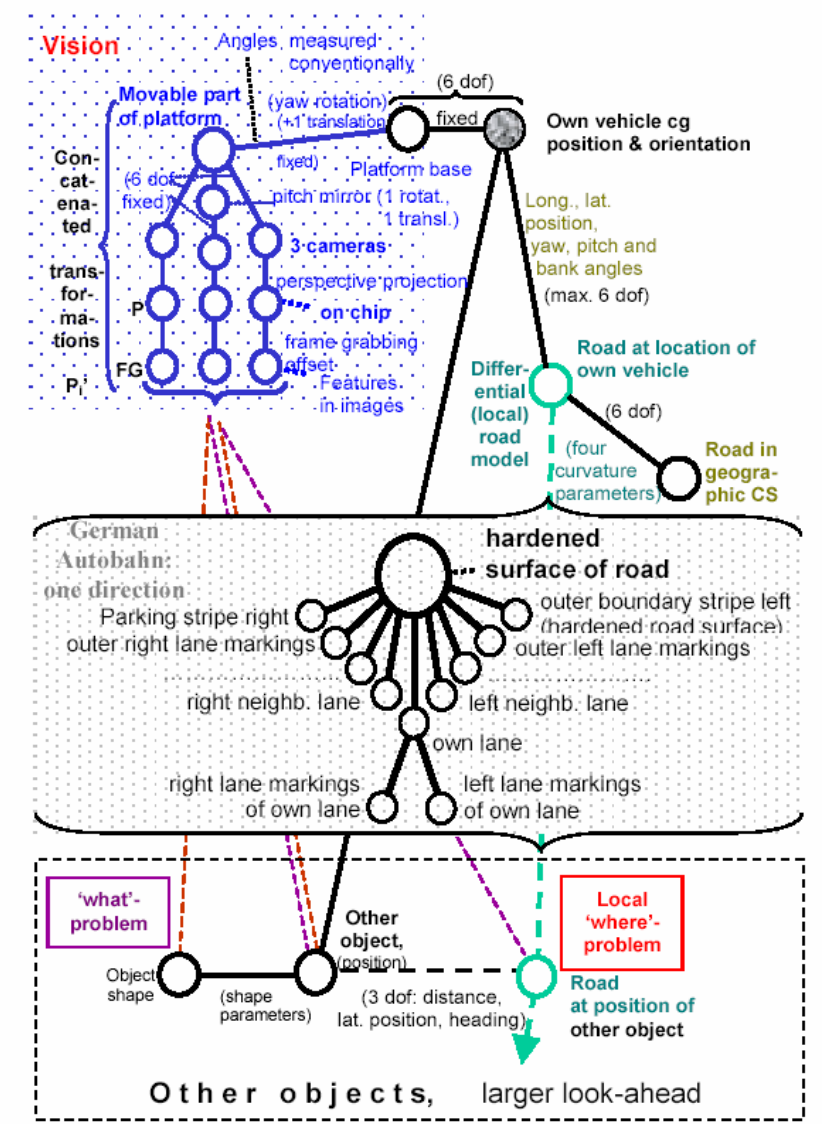


Fig. 6: Scene tree for driving on a German Autobahn with three cameras on a special platform with yaw control for all cameras and pitch control through a mirror for just the tele-camera.

This has to be available for own decision making anyway. Recognizing that these characteristic behavioral features of subjects can be used as abstract representations for characterizing subject classes, in general, has led to explicit representation of behavioral capabilities as knowledge elements for understanding subject motion.

This allows recognizing maneuvers from a short initializing typical motion component (like lane change (intended) from a systematical change of lateral position towards the lane marking). Subjects in the environment can thus be labeled as performing some action; this can be taken into account for own decision making. As a new step in the development of autonomous systems, from this, the explicit representation of all kinds of capabilities has been derived. It has not yet been realized to full extent for perception, planning, decision-making, gaze- and motion control. However, it is felt that this is an important step to really autonomous systems capable of social cooperation and of learning.

3.2.5. Integrated visual / inertial perception

The own pose is derived to a large extent from inertial measurements and signal integration with little time delay. The relative position and orientation of other objects is hypothesized and updated based on vision with smooth sequences of parallel image streams between saccades. The internal representation of the situation perceived is a mental construct relying largely on spatio-temporal models and prediction error feedback. It contains neither inertial nor visual data directly. Time delays between data exploited are bridged with these models. The latest update of the overall representation of the scene with the relevant objects is achieved after the longest delay time. Control output onto actuators takes the delay times into account via model-based prediction [33-37].

Figure 7 is meant to symbolize this 'extended presence' and the stabilized internal representation unavoidable with the mix of data and time delays in the overall system. High-frequency feedback loops with little delay time are running in parallel with slower loops based on high-level models including delay time compensation. This multiple scale solution is quite natural in distributed systems with different types of models on the different levels. Corresponding time constants for filtering allow efficient data processing. For example, the bias in pitch angle due to fuel consumption over longer time periods can be determined and compensated by a very-low-frequency model superimposed on the other ones.

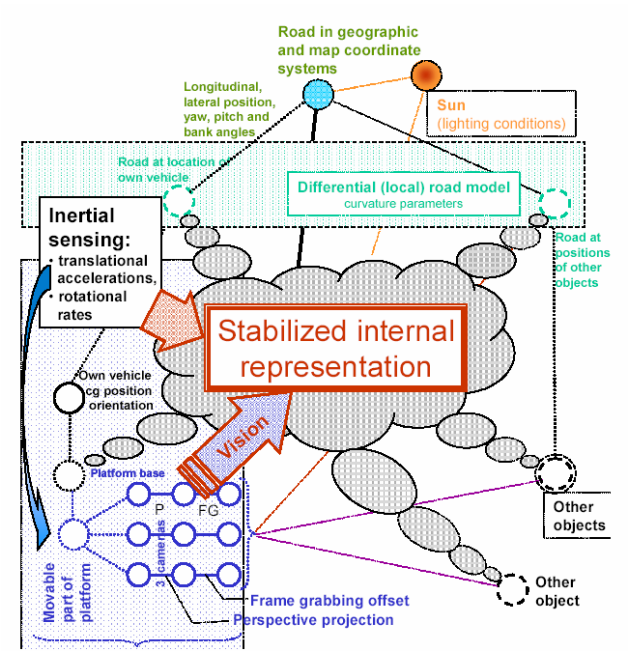


Fig. 7: Integrated visual / inertial perception by recursive estimation with spatio-temporal models and time delay compensation.

4. SITUATION ASSESSMENT

Judging the overall situation based on the best estimates of relative states of other objects/subjects is a process, which need not run at video rate, for example. When new feature sets are discovered which may be indications of a new obstacle, delay time until a proper reaction is triggered should be minimal; however, this may even be possible without understanding the new situation to the full extent. Therefore, a dual approach with a fast reflex-like reaction in the safe direction to the object in isolation, and a new evaluation of the situation with more time delay seems a good approach. For arriving at these behavioral decisions, separate evaluation processes have been realized. There are two on the process control level, one for gaze and attention, the other one for vehicle locomotion. They do have access to the description of the actual situation produced mainly (but not exclusively) by a third unit called 'Central Decision' (CD). CD has to come up with an evaluation of all objects/subjects in the context of own mission performance indicating which ones are most relevant and which ones may be discarded for decisions on the local level.

The situation is described by adding linguistic situation aspects to objects/subjects in the scene tree [22, 23]. For practical reasons, additional local storage and pointers often realize this.

5. BEHAVIOR DECISION AND IMPLEMENTATION

Figure 8 gives a symbolic description of the hierarchical behavioral decision units. The central polyhedron symbolizes the collection of situation aspects derived from observing time histories of state variables in the scene tree, referring these to known behavioral capabilities and recognizing maneuvers under performance or hypothesizing intentions for the near future.

CD takes care of setting the side constraints for achieving the mission goals. Typical tasks here are specifying the sequence of mission elements to be performed and setting reference parameters for the lower levels (like average speed for the mission element) [29]. The local decision units for gaze and attention (BDGA) and locomotion (BDL) then perform their decisions autonomously taking their more detailed knowledge about the behavioral capabilities in their category into account [35]. Especially, this level can initiate first reactions to unexpected obstacles [36]. CD may later on revise them, if necessary. In case of conflicts or if it is impossible to realize the behavior requested from CD, an information exchange is initiated for finding a solution acceptable to both sides.

6. HARDWARE REALIZATION

Figure 9 shows the hardware realization of the EMS vision system in the test vehicle **VaMoRs** (lower right corner) on a cluster of four PC plus two subsystems (19''-units in lower right). Three DualPentium PC (mounted also in a 19'' rack) perform image evaluation for the different cameras. They are configured as embedded PC by a special software package (EPC). PC1 essentially evaluates the images of the wide-angle cameras for recognizing the near road environment. PC2 analyzes the color images of the mild tele-camera, and PC3 is reserved for monochrome tele-image evaluation (e.g. for landmark recognition).

On the fourth PC ('behavior PC', to the right) all data evaluation results converge for situation assessment and decision-making (CD, BDGA and BDL in the lower left corner). The subsystems for interaction with sensors and actuators are linked to the corresponding processes (GC for **g**aze **c**ontrol, and VC for **v**ehicle **c**ontrol).

A **S**calable **C**oherent **I**nterface (SCI) for inter-processor communication with up to ~ 100 MB/s data rate also serves for synchronization between the different distributed processes. Dynamic knowledge representation (DKR) is kept identical for all processors by regular updating at video rate. To the 'Behavior PC' also the receiver for the Global Positioning System (GPS) and the **H**uman **M**achine **I**nterface (HMI) are connected, through which all commands from the operators are handled.

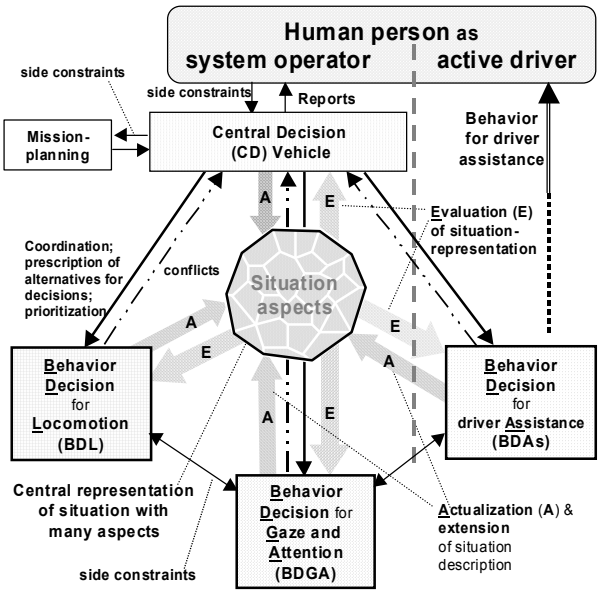


Fig. 8: Hierarchically structured, distributed decision agents based on competencies: For visual perception (lower center), locomotion (lower left), driver assistance (lower right), and for overall systems aspects (center top); a common representation of the situation is desirable [after 23; 35]

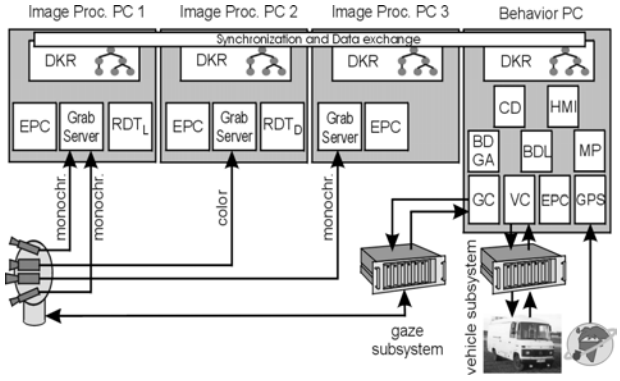


Fig. 9: Hardware realization on a cluster of PC's plus two subsystems for hardware interfacing (gaze and vehicle control).

7. EXPERIMENTAL RESULTS

EMS-vision has mainly been tested on a network of minor roads with missions as they usually occur in military

scouting. Both recognizing crossings with turnoffs onto a crossroad (left and right) and leaving roadways for driving cross country on grass surfaces as well as entering roads after these maneuvers have been demonstrated. During cross-country segments, detection and avoidance of negative obstacles like ditches have been demonstrated. For the latter task, a very powerful additional subsystem for stereo-detection and tracking of negative obstacles has been integrated in a joint project with US-partners [20]. Up to 80 billion operations per second have become available for stereo interpretation by this system. While the Pyramid Vision System PVS-200 of 2001 required a volume of about 30 liters, in 2002 the new 'Acadia'-board as a plug-in card into a PC became available and allowed full-frame, full video rate stereo evaluation in conjunction with EMS-vision. Joint intensity and stereo evaluation allowed robust ditch detection and tracking while driving on non-flat ground and bypassing the ditch with MarVeye keeping the bypassed end fixated [38].

Figure 10 shows test results from detecting, tracking and turning-off onto a crossroad. 10a shows the yaw (pan) angle time history of the platform during this maneuver. From second ~ 90 to 110 a saccading maneuver in gaze is performed in order to alternatively collect data on the crossing (distance and speed of approach) and on the geometry of the crossroad (width and angle of intersection). In 10b the saccade bit is shown telling the image evaluation processes whether it makes sense to process images or whether the images are blurred and they should stick to prediction with spatio-temporal models (when the bit is up). From sub-figures 10c - e the objects observed can be seen. The crossroad is inserted into the scene tree at around 90 seconds and becomes the new reference road (split into local (near) and distant) at around 115 seconds (s). During approach of the intersection (10f), the gaze angle in yaw increases up to ~ 60° (10g, h), telling that the vehicle turns its viewing direction 'over the shoulder' while driving still straight ahead on the old road. At ~ 116 s reorganization of the scene tree is finished with the old crossroad now being the new reference road. The gaze angle of MarVEye is constantly in direction of this new reference while the vehicle turns underneath it until at ~ 130 s gaze is almost

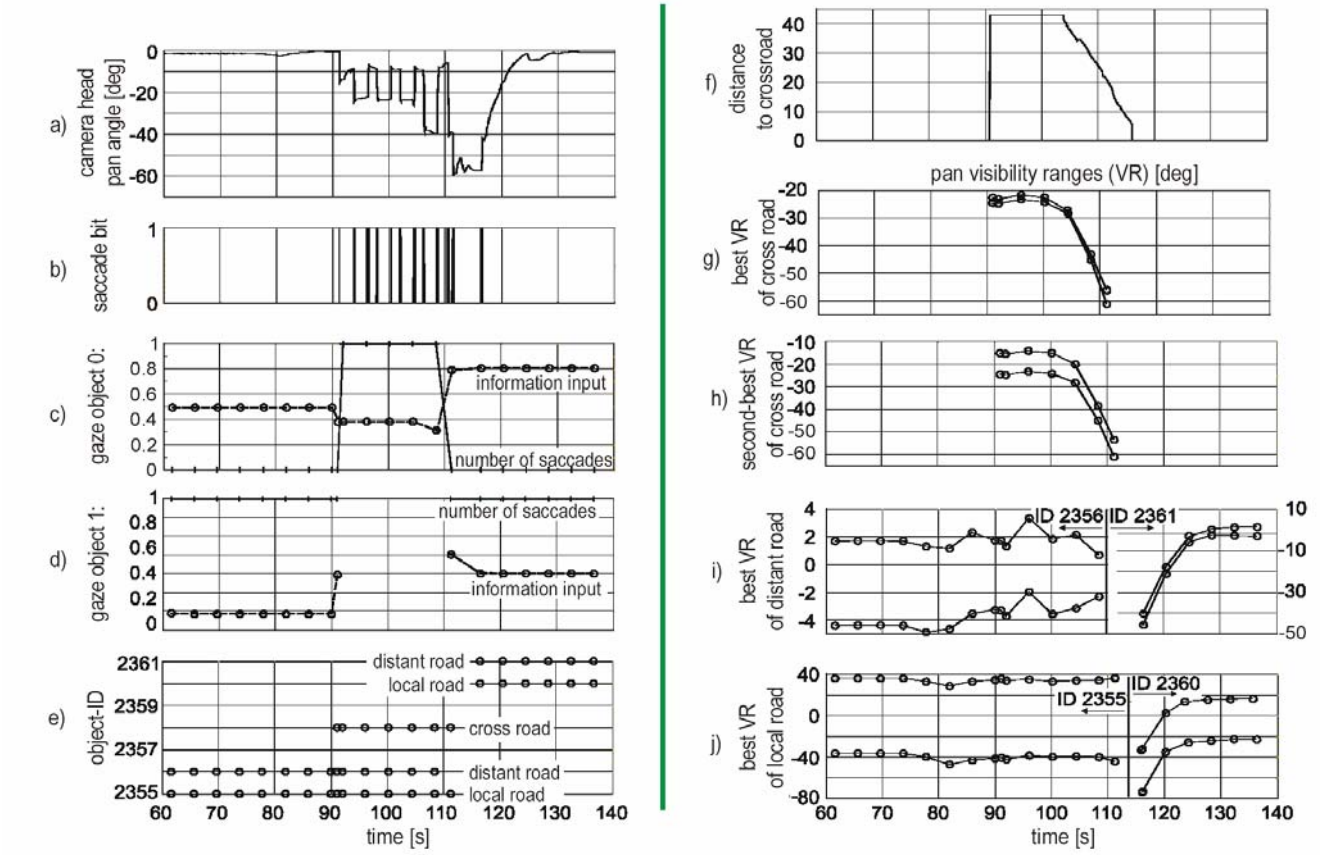


Fig. 10: Time histories of variables in EMS-vision in VaMoRs for detecting a crossroad and turning-off onto it autonomously.

in direction of the body longitudinal axis again (10i. j). (These figures show the best viewing ranges (VR) as evaluated by BDGA. Small offsets from zero may stem from the fact that the system is preparing for leaving the road and turning onto the grass surface; since only one boundary of the road can be tracked by the tele-camera at the range specified, the gaze angle selected is 2° (10i, right).

8. CONCLUSIONS

Once the assumptions of smooth surfaces being driven on and of small fields of view being sufficient for road running on high-speed roads are dropped, robust and reliable vision systems for ground vehicles become much more involved. Joint inertial and visual data evaluation in connection with active gaze control provides many advantages. The principles, which nature has found out to yield optimal solutions for the vertebrate eye, have been applied to a technical vision system looking very promising for more advanced vision systems for cars than the ones presently under development at many places. First results with an expectation-based, multi-focal, saccadic (EMS) vision system have been discussed. Further development steps desirable have been mentioned. Processing power and communication bandwidth needed for this approach will become available in the near future. The development of corresponding knowledge bases for different application areas of this general approach is a demanding step for the near future.

9. REFERENCES

[1] C. Thorpe, 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

[2] T.M. Jochem, D.A. Pomerleau, C.E. Thorpe: MANIAC. A next generation neurally based autonomous road follower. Proc. IEEE Conf. on Intell. Auton. Systems (IAS-3), Pittsburgh, USA, Febr. 1993.

[3] T.M. Jochem, D.A. Pomerleau, C.E. Thorpe: Vision-based neural network road and intersection detection and traversal. Proc. IEEE Conf. IROS, Aug. 1995, Pittsburgh, USA

[4] D.A. Pomerleau: Ralph: Rapidly adapting lateral position handler. Proc. IEEE Symp. on Intell. Vehicles'95, Detroit, Mi, USA, 1995.

[5] C. Thorpe, T. Jochem, D. Pomerleau: The 1997 Automated Highway Free Agent Demonstration. Proc. IEEE-Conf. on Intell. Transp. Systems, Bostson, USA, Nov. 1997.

[6] M. Hebert, K. Redmill, A. Stentz (eds.): Intelligent Unmanned Ground Vehicles – Autonomous Navigation at Carnegie Mellon. Kluwer Academic Publishers, 1997.

[7] B. Ulmer: VITA II – Active Collision Avoidance in Real Traffic. In [IV'94], pp. 1 – 6.

[8] U. Franke, S. Mehring, A. Suissa, S. Hahn: The Daimler-Benz Steering Assistant – a spin-off from autonomous driving. Proc. of Symp. on Intell. Vehicles'94, Paris, Oct. 1994.

[9] S. Estable, J. Schick, F. Stein, R. Janssen, R. Ott, W. Ritter, Y.-J. Zheng: A Real-Time Traffic Sign Recognition System. Proc. of Symp. on Intell. Vehicles'94, Paris, Oct. 1994, pp. 213 – 218.

[10] S. Hahn: Automation of Driving Functions – Future Development, Benefits and Pitfalls. Proc. IEEE Symp. on Intell. Vehicles'96, Tokyo, Japan, Sept. 1996, pp. 309 – 312.

[11] D.M. Reichardt: Using Automated Assistance Systems – Putting the Driver into Focus. Proc. IEEE Symp. on Intell. Vehicles'98, Stuttgart, Oct. 1998, Vol. 2, pp. 413 – 418.

[12] A. Gern, U. Franke, P. Levi: Advanced Lane Recognition Fusing Vision and Radar. Proc. Intell. Veh.'00, Dearborn, USA, Oct.2000

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

[14] Dickmanns E.D.: The development of machine vision for road vehicles in the last decade. Proc. of the Int. Symp. on ‚Intell. Veh.'02', Versailles, June 2002

[15] Proc. of the International Symposium on 'Intelligent Vehicles' starting 1992, (organized yearly by I. Masaki and various institutions)

[16] Proc. of Symp. on 'Intell. Veh.'. Dearborn, MI, USA, Oct. 2000, with the following contributions on EMS-Vision:

a) R. Gregor, M. Lützeler, M. Pellkofer, Siedersberger K.H., E.D. Dickmanns.: EMS-Vision: A Perceptual System for Autonomous Vehicles. pp. 52 – 57.
b) R. Gregor, E.D. Dickmanns.: EMS-Vision: Mission Performance on Road Networks. pp. 140 – 145.
c) U. Hofmann, A. Rieder, E.D. Dickmanns: EMS-Vision: An Application to Intelligent Cruise Control for High Speed Roads. pp. 468 – 473.
d) M. Lützeler, E.D. Dickmanns: EMS-Vision: Recognition of Intersections on Unmarked Road Networks. 302 – 307.
e) M. Maurer: Knowledge Representation for Flexible Automation of Land Vehicles. pp. 575 – 580.
f) M. Pellkofer, E.D. Dickmanns: EMS-Vision: Gaze Control in Autonomous Vehicles. pp. 296 – 301.
g) K.-H. Siedersberger, E.D.Dickmanns: EMS-Vision: Enhanced Abilities for Locomotion. pp. 146 – 151.

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

[18] M. Lützeler: Visuelle Erkennung von Verzweigungen und Knotenpunkten auf Straßen niederer Ordnung. Diss. UniBwM, LRT, 2001

[19] 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

[20.] 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

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

[22] Pellkofer M., Dickmanns E.D.: Behavior Decision in Autonomous Vehicles. Proc. of the Int. Symp. on ‚Intell. Veh.‘02‘, Versailles, June 2002

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

[24] 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

[25] F. Thomanek, E.D. Dickmanns, D. Dickmanns: Multiple Object Recognition and Scene Interpretation for Autonomous Road Vehicle Guidance. Proc. Int. Symp. on Intell. Vehicles’94, Paris, Oct. 1994, pp. 231 - 236.

[26] F. Thomanek: Visuelle Erkennung und Zustandsschätzung von mehreren Straßenfahrzeugen zur autonomen Fahrzeug-führung. Diss., UniBwM, LRT, 1996

[27] R. Behringer: Visuelle Erkennung und Interpretation des Fahrspurverlaufs durch Rechnersehen. Diss. UniBw Munich, LRT, 1996.

[28] Dickmanns E.D., Wuensche H.-J.: 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, Academic Press, 1999, pp 569-620

[29] R. Gregor: Fähigkeiten zur Missionsdurchführung und Landmarkennavigation Diss. UniBw Munich, LRT, 2002.

[30] Dickmanns E.D.; Christians T.: Relative 3-D-state Estimation for Autonomous Visual Guidance of Road Vehicles. In T. Kanade et al (eds): '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

[31] W. Efenberg, Q.H. Ta, L. Tsinas, V. Graefe: Automatic Recognition of Vehicles Approaching from behind. Proc. of Int. Symp. on Intell. Veh.’92, Detroit, 1992, pp. 57 – 62.

[32] Dirk Dickmanns: Rahmensystem für visuelle Wahrnehmung veränderlicher Szenen durch Computer. Diss., UniBw Munich, Informatik, 1997.

[33] S. Werner: Maschinelle Wahrnehmung für den bordautonomen automatischen Hubschrauberflug. Diss. UniBw Munich, LRT, 1997.

[34] Schiehlen J.: Kameraplattformen für aktiv sehende Fahrzeuge. Diss. UniBwM, LRT, Juni 1995

[35] M. Maurer: Flexible Automatisierung von Straßenfahrzeugen mit Rechnersehen. Diss. UniBwM, LRT, 2000

[36] 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

[37] K.-H. Siedersberger: Verhaltensrealisierung in EMS-Vision. Diss. UniBw Munich, LRT, 2003 (to appear).

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